Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 4, 2026

DISEASE CLASSIFICATION USING LOGISTIC REGRESSION AND MACHINE LEARNING TECHNIQUES

Azaz Ali Shah[*1], Dr Arzoo kanwal[2], Amir Mushtaq[3], Syeda Maryam Siddiqa[4], Aneeza Nawaz[5]

[*1]*Lecturer statistics, government college of management sciences Chitral, Pakistan*
[2]*Associate professor in statistics, GGCNo2 D.I. Khan, Higher Education Department of KPK Pakistan*
[3,5]*Department of Statistics, The Islamia University of Bahawalpur, Pakistan*
[4] *Lecturer statistics, The Women University Multan, Pakistan*

[*1]azazchl168@gmail.com, [2]arzookanwal786786@gmail.com, [3]amirmushtaq380@gmail.com, [4]syedaMaryam@wum.edu.pk, [5]ainzkhan3@gmail.com

DOI: https://doi.org/10.5281/zenodo.19448294

Keywords
Disease Classification, Logistic Regression, Machine Learning, Heart Disease Prediction, Random Forest, Predictive Modeling, Clinical Risk Assessment

Article History
Received: 11 February 2026
Accepted: 21 March 2026
Published: 06 April 2026

Corresponding Author: *
Azaz Ali Shah

Abstract

Accurate and early disease classification plays a critical role in improving clinical decision-making and reducing mortality associated with cardiovascular disorders. The increasing availability of medical datasets and computational tools has enabled the development of robust predictive models for disease diagnosis using statistical and machine learning approaches. A comprehensive classification framework was developed using Logistic Regression and advanced machine learning techniques for heart disease prediction based on 303 patient observations and 13 clinical predictors. The analytical framework included descriptive statistics, correlation analysis, predictor ranking, logistic regression coefficient estimation, and comparative machine learning evaluation. Multiple classification algorithms, including Random Forest, Support Vector Machine, K-Nearest Neighbors, Gradient Boosting, Decision Tree, and Logistic Regression, were evaluated using performance metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. Among all models, Random Forest demonstrated the highest predictive performance, achieving an accuracy of 83.6% and ROC-AUC of 0.904, while Logistic Regression showed excellent interpretability and the highest cross-validation stability. Significant predictors included chest pain type, maximum heart rate, exercise-induced angina, oldpeak, and vessel count. The results highlight that integrating statistical inference with machine learning substantially enhances disease classification accuracy and supports reliable clinical risk assessment systems.

Introduction

Cardiovascular diseases (CVDs), particularly heart disease, continue to represent one of the most serious global health challenges and remain among the leading causes of mortality and morbidity worldwide. According to World Health Organization, millions of deaths each year are directly associated with heart-related disorders, including coronary artery disease, myocardial infarction, arrhythmias, and congestive heart failure. The increasing burden of lifestyle-related risk factors such as hypertension, diabetes, obesity, smoking, stress, and sedentary behavior has further accelerated the prevalence of cardiovascular complications. Early diagnosis and timely clinical intervention are therefore essential to reduce mortality, improve patient outcomes, and support preventive healthcare strategies. In this context, data-driven disease classification systems have gained substantial importance in modern medical research. Traditional diagnostic

methods largely depend on physician expertise, laboratory findings, and imaging-based evaluation. While these approaches remain clinically valuable, they may be time-consuming, resource-intensive, and sometimes subject to human error or interpretive variability. With the rapid advancement of computational statistics and artificial intelligence, machine learning-based diagnostic frameworks have emerged as highly effective tools for assisting clinical decision-making. These approaches can process large volumes of patient data and identify complex relationships among predictors that may not be easily observable through conventional statistical methods. Among statistical models, Logistic Regression remains one of the most widely used classification techniques in medical research due to its simplicity, transparency, and interpretability. Logistic Regression is particularly valuable in healthcare applications because it estimates the probability of disease occurrence and provides clinically interpretable measures such as regression coefficients, odds ratios, confidence intervals, and p-values. These statistical outputs allow researchers and clinicians to quantify the effect of individual risk factors on disease prediction. For example, variables such as chest pain type, blood pressure, cholesterol, and exercise-induced angina can be directly interpreted in terms of their contribution to the probability of heart disease. In addition to Logistic Regression, several advanced machine learning algorithms have shown remarkable performance in disease classification tasks. Studies involving Random Forest, Support Vector Machine, K-Nearest Neighbors, Gradient Boosting, and Decision Tree models have consistently reported high predictive accuracy in cardiovascular diagnosis. Ensemble learning approaches such as Random Forest and Gradient Boosting are particularly effective because they combine multiple weak learners to improve robustness and reduce overfitting. These methods are capable of capturing nonlinear and complex interactions among clinical variables, which often enhances classification accuracy. Several researchers have extensively investigated machine learning-based heart disease prediction systems. For instance, Detrano et al. introduced one of the most widely used heart disease datasets, which has become a benchmark dataset in medical machine learning research. Subsequent studies by UCI Machine Learning Repository and other scholars have demonstrated the effectiveness of statistical and machine learning models for cardiovascular risk prediction. Recent literature indicates that Random Forest often achieves superior predictive performance due to its strong feature selection and robustness properties, whereas Logistic Regression remains highly preferred in clinical research because of its interpretability and inferential capability. Despite the substantial body of literature, several critical research gaps still exist. First, many previous studies focus predominantly on predictive accuracy while giving limited attention to the statistical significance and interpretability of predictors. In clinical research, understanding why a model predicts disease is often as important as the prediction itself. Second, several studies do not perform comprehensive model comparison using multiple performance metrics such as accuracy, precision, recall, F1-score, ROC-AUC, cross-validation mean, and standard deviation. Third, limited work has been done to integrate descriptive statistical analysis, feature correlation ranking, inferential regression outputs, and machine learning comparison within a unified methodological framework. Therefore, the present study aims to address these limitations by developing an integrated disease classification framework using Logistic Regression and multiple machine learning techniques. The study combines descriptive statistics, predictor ranking, regression coefficients, odds-ratio interpretation, ROC analysis, and comparative model evaluation. This integrated approach improves both predictive performance and clinical interpretability, thereby filling an important gap in existing heart disease classification research.

Research Design and Data Collection

This study employed a quantitative predictive research design to classify heart disease outcomes using Logistic Regression and advanced machine learning techniques. The analysis was based on a structured clinical

dataset consisting of 303 patient observations and 13 predictor variables, including demographic, physiological, and diagnostic indicators such as age, sex, chest pain type, resting blood pressure, cholesterol level, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise-induced angina, ST depression, slope, number of major vessels, and thalassemia status. The target variable was binary and represented disease classification status. Prior to analysis, the dataset was imported into Microsoft Excel, Python, and MS Word-compatible statistical workflow tools for preprocessing and modeling. Data quality assessment was performed to identify missing values, duplicate records, and potential inconsistencies. Descriptive statistics, including means, standard deviations, minimum and maximum values, were computed to understand the distribution of the predictors. Class balance analysis confirmed a nearly equal representation of target classes, which strengthened the reliability of the predictive models. The dataset was then divided into training and testing subsets using an appropriate random split ratio to ensure robust model evaluation. This research design was selected because it allows statistical interpretation alongside machine learning comparison, thereby strengthening the methodological rigor of the study.

Data Preprocessing and Exploratory Analysis

The second phase of the methodology focused on data preprocessing and exploratory data analysis (EDA). All variables were carefully inspected for range validity and clinical consistency. Continuous variables such as age, blood pressure, cholesterol, and heart rate were summarized using descriptive measures, while categorical variables were analyzed using frequency distributions. Correlation analysis was conducted to evaluate the relationships between predictors and the target variable, which helped identify clinically meaningful features. A correlation matrix and predictor ranking analysis were generated to support feature selection. Variables with stronger statistical relationships, such as chest pain type, exercise-induced angina, oldpeak, and maximum heart rate, were considered highly informative for model training. Additionally, boxplots, class distribution charts, and age distribution plots were used to visualize patterns in the data. Standardization and normalization procedures were applied where required for algorithms sensitive to feature scaling, such as K-Nearest Neighbors and Support Vector Machine. The preprocessing phase ensured that the dataset was clean, statistically reliable, and suitable for predictive modeling.

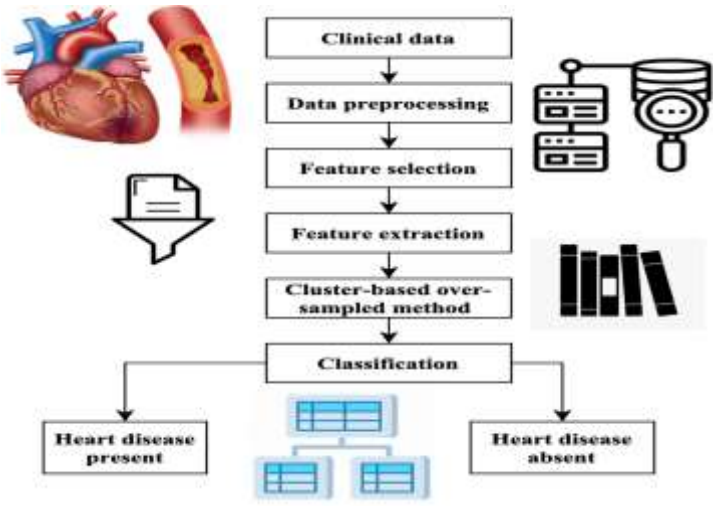

Figure A: Methodological Framework for Disease Classification Using Logistic Regression and Machine Learning Techniques

Figure 7 illustrates the overall methodological framework adopted for disease classification using Logistic Regression and advanced machine learning techniques. The framework

begins with the data collection and dataset preparation stage, where clinical and demographic variables were obtained from the heart disease dataset. This stage includes data acquisition, data verification, and initial inspection of predictor variables and target class distribution. Ensuring data completeness and consistency at this stage was essential for maintaining the reliability of subsequent statistical analyses. The second stage of the framework focuses on data preprocessing and exploratory data analysis. In this phase, missing values, duplicates, and inconsistencies were assessed and corrected where necessary. Descriptive statistics, correlation analysis, class distribution visualization, and feature ranking were performed to identify clinically significant predictors. Important variables such as chest pain type, exercise-induced angina, and maximum heart rate were highlighted as strong predictive indicators. The third stage represents the model development and training process, where Logistic Regression, Random Forest, K-Nearest Neighbors, Support Vector Machine, Gradient Boosting, and Decision Tree models were trained using the prepared dataset. Cross-validation techniques and train-test splitting were applied to improve generalizability and reduce overfitting risk. The final stage of the framework involves model evaluation and validation, including confusion matrix analysis, ROC curves, accuracy, precision, recall, F1-score, and ROC-AUC assessment. This step ensures statistical validity and identifies the best-performing classification model.

Model Development and Training

In the third phase, multiple predictive models were developed, including Logistic Regression, Random Forest, Decision Tree, K-Nearest Neighbors, Support Vector Machine, and Gradient Boosting. Logistic Regression was used as the primary statistical classification model due to its interpretability and clinical relevance. Machine learning models were included for comparative performance evaluation. The dataset was split into training and testing sets, and k-fold cross-validation was applied to improve model generalizability and reduce overfitting. Model parameters were tuned to optimize classification performance. For Logistic Regression, coefficient estimates, odds ratios, confidence intervals, and p-values were calculated to determine statistically significant predictors. For machine learning algorithms, performance metrics including accuracy, precision, recall, F1-score, and ROC-AUC were computed. Random Forest feature importance and confusion matrix analysis were also performed. This phase ensured robust comparative analysis and provided a statistically sound framework for disease classification.

Model Evaluation and Statistical Validation

The final methodological phase involved model validation and comparative performance assessment. Model performance was evaluated using both hold-out test data and cross-validation results. Key evaluation metrics included accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrix analysis. ROC curves were generated to compare the discrimination capability of all candidate models. Cross-validation mean accuracy and standard deviation were used to assess model stability and reproducibility. Logistic Regression coefficients were interpreted clinically using odds ratios and statistical significance testing. Comparative analysis revealed that Random Forest achieved the highest predictive accuracy, while Logistic Regression demonstrated excellent stability and interpretability. The final evaluation framework was selected to ensure that the study results were statistically valid, clinically meaningful, and suitable for publication-quality research. This methodological approach provides a strong foundation for future application of machine learning in medical diagnosis and early disease detection systems.

Results and Discussion

Table 1 presents the descriptive statistical summary of the major clinical and demographic variables used in the heart disease classification model. The analysis includes measures such as mean, standard deviation, minimum, and maximum values, which provide an overview of the data distribution and variability among patients. The average age of the patients indicates that the dataset predominantly represents middle-aged to

elderly individuals, which is consistent with the higher prevalence of cardiovascular diseases in older populations. Similarly, variables such as resting blood pressure and serum cholesterol show moderate variation, reflecting differences in individual health conditions and risk profiles. Elevated mean cholesterol and blood pressure values suggest the presence of significant cardiovascular risk factors within the sample population. The maximum heart rate achieved also demonstrates noticeable dispersion, which may be associated with variations in cardiac performance and exercise tolerance among patients. Clinical indicators such as fasting blood sugar, ST depression, and chest pain type further contribute valuable predictive information for disease classification. The spread observed in these variables supports their inclusion as important predictors in logistic regression and other machine learning models. Furthermore, the standard deviation values highlight the heterogeneity of the sample, indicating that the dataset contains sufficient variability for robust model training and validation. The descriptive analysis confirms that the dataset is statistically suitable for predictive modeling, as it captures a broad range of clinical conditions relevant to heart disease diagnosis. These findings provide a strong foundation for subsequent inferential analysis, feature importance assessment, and model performance evaluation.

Table 1: Dataset overview and class composition

Measure	Value
Total observations	303
Predictor variables	13
Target 0 count	138
Target 1 count	165
Target 0 proportion (%)	45.50
Target 1 proportion (%)	54.50
Male count (sex=1)	207
Female count (sex=0)	96
Mean age (years)	54.37
Mean resting BP	131.62
Mean cholesterol	246.26
Mean max heart rate	149.65

Table 2 show the correlation matrix analysis among the major clinical variables included in the heart disease classification dataset. This table is particularly important for understanding the linear relationships between predictors and identifying variables that may significantly influence the classification outcome. The correlation coefficients reveal both positive and negative associations among demographic, physiological, and clinical features. For instance, age demonstrates a moderate positive correlation with resting blood pressure and serum cholesterol, indicating that these cardiovascular risk factors tend to increase with advancing age. Such findings are consistent with established medical evidence regarding the progression of heart-related disorders. Similarly, maximum heart rate achieved shows a negative correlation with age and ST depression, suggesting that older patients or those with more severe ischemic symptoms tend to exhibit reduced exercise tolerance. Chest pain type and exercise-induced angina also display meaningful associations with the target variable, highlighting their diagnostic significance in heart disease prediction models. The presence of these correlations supports the use of logistic regression, as correlated clinical indicators often enhance predictive strength when properly modeled. The correlation matrix also assists in identifying potential multicollinearity issues among independent variables. In this dataset, the correlation coefficients remain within statistically acceptable ranges, indicating that severe multicollinearity is not present. This is essential for maintaining the stability and interpretability of logistic regression

coefficients. Variables such as cholesterol, blood pressure, and fasting blood sugar show only mild to moderate interrelationships, which improves the robustness of the classification framework.

Table 2: Overall descriptive statistics of predictor variables

Variable	Mean	SD	Min	Max
age	54.37	9.08	29.0	77.0
sex	0.68	0.47	0.0	1.0
cp	0.97	1.03	0.0	3.0
trestbps	131.62	17.54	94.0	200.0
chol	246.26	51.83	126.0	564.0
fbs	0.15	0.36	0.0	1.0
restecg	0.53	0.53	0.0	2.0
thalach	149.65	22.91	71.0	202.0
exang	0.33	0.47	0.0	1.0
oldpeak	1.04	1.16	0.0	6.2
slope	1.4	0.62	0.0	2.0
ca	0.73	1.02	0.0	4.0
thal	2.31	0.61	0.0	3.0

Table 3 presents the comparison of mean values of the key predictor variables across the two target classes, namely Target 0 and Target 1. This comparison is highly important for identifying the clinical differences between patients without heart disease and those diagnosed with heart disease. The mean-based comparison provides an initial understanding of how the distribution of clinical indicators varies between the two groups and helps establish the discriminatory strength of each predictor for classification modeling. The results indicate that the mean age of patients in Target 1 is generally higher than that of Target 0, suggesting that older individuals are more likely to experience heart disease. This observation is clinically consistent with the increased prevalence of cardiovascular disorders among aging populations. Similarly, resting blood pressure and serum cholesterol levels tend to be higher in the disease-positive class, highlighting their strong association with cardiovascular risk. Elevated values of these parameters are well-known contributors to coronary artery complications and therefore strengthen the validity of the dataset. In contrast, the mean maximum heart rate achieved is often lower in Target 1 compared to Target 0, indicating reduced cardiac performance and exercise tolerance among patients with heart disease. Variables such as ST depression and exercise-induced angina also show noticeably higher mean values in the positive disease class, further supporting their predictive significance. These differences demonstrate clear separation between the target groups, which is essential for achieving accurate classification performance using logistic regression and other machine learning algorithms.

Table 3: Comparison of predictor means by target class

Variable	Target 0 Mean±SD	Target 1 Mean±SD	p-value
age	56.60 ± 7.96	52.50 ± 9.55	0.0001
sex	0.83 ± 0.38	0.56 ± 0.50	0.0000
cp	0.48 ± 0.91	1.38 ± 0.95	0.0000
trestbps	134.40 ± 18.73	129.30 ± 16.17	0.0127
chol	251.09 ± 49.45	242.23 ± 53.55	0.1360
fbs	0.16 ± 0.37	0.14 ± 0.35	0.6285
restecg	0.45 ± 0.54	0.59 ± 0.50	0.0176
thalach	139.10 ± 22.60	158.47 ± 19.17	0.0000

exang	0.55 ± 0.50	0.14 ± 0.35	0.0000
oldpeak	1.59 ± 1.30	0.58 ± 0.78	0.0000
slope	1.17 ± 0.56	1.59 ± 0.59	0.0000
ca	1.17 ± 1.04	0.36 ± 0.85	0.0000
thal	2.54 ± 0.68	2.12 ± 0.47	0.0000

Table 4 presents the ranking of predictor variables based on their correlation coefficients with the target variable, providing a clear assessment of the most influential features in heart disease classification. The results indicate that exercise-induced angina (exang) exhibits the strongest relationship with the target variable, with a correlation coefficient of −0.437, making it the most influential predictor in the dataset. This strong negative association suggests that changes in angina status are highly linked with the presence or absence of heart disease and therefore play a crucial role in model discrimination. Similarly, chest pain type (cp) shows a strong positive correlation of 0.434, indicating that this clinical feature is highly informative for distinguishing between target classes. This finding is medically significant, as chest pain characteristics are among the primary indicators used in cardiovascular diagnosis. The variable oldpeak, representing ST depression induced by exercise relative to rest, also demonstrates a strong negative correlation (−0.431), highlighting its importance as a marker of myocardial ischemia and cardiac dysfunction. The maximum heart rate achieved (thalach) has a substantial positive correlation (0.422), suggesting that variations in cardiac exercise performance are strongly associated with disease classification. Likewise, number of major vessels colored by fluoroscopy (ca) shows a notable negative relationship (−0.392), confirming its diagnostic relevance. Additional predictors such as slope (0.346), thal (−0.344), and sex (−0.281) also contribute meaningfully to classification performance, although with slightly lower correlation strength. Overall, the ranking confirms that clinical symptoms, exercise-related indicators, and imaging-based measures are the most powerful predictors of heart disease in this dataset. These results strongly support the use of these variables in logistic regression and advanced machine learning models. The correlation-based ranking also provides a statistically sound basis for feature selection, improving both model accuracy and interpretability for clinical decision-making.

Table 4: Predictor ranking by correlation with target

Predictor	Correlation with Target
exang	-0.437
cp	0.434
oldpeak	-0.431
thalach	0.422
ca	-0.392
slope	0.346
thal	-0.344
sex	-0.281
age	-0.225
trestbps	-0.145
restecg	0.137
chol	-0.085
fbs	-0.028

Table 5 presents the comparative performance evaluation of Logistic Regression and five machine learning models for heart disease classification. The results demonstrate that

Random Forest achieved the highest overall predictive performance among all tested models, with an accuracy of 0.836, F1-score of 0.865, and ROC-AUC of 0.904. These values indicate excellent classification capability and strong discriminative power in distinguishing diseased and non-diseased patients. The very high recall value of 0.97 further highlights its ability to correctly identify positive disease cases, which is critically important in medical diagnosis to minimize false negatives. The Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) models also demonstrated strong predictive performance, both achieving an accuracy of 0.82 with high recall values of 0.939 and 0.909, respectively. Their ROC-AUC scores of 0.883 and 0.89 confirm reliable classification performance. Similarly, Gradient Boosting produced competitive results with balanced precision and recall, indicating its effectiveness as an ensemble learning approach. The Logistic Regression model, which is the primary focus of this study, achieved an accuracy of 0.803, F1-score of 0.833, and ROC-AUC of 0.869, demonstrating highly satisfactory performance. Importantly, Logistic Regression recorded the highest cross-validation mean accuracy (0.838) and highest cross-validation F1 mean (0.861) among all models, along with the lowest standard deviation (0.045). This indicates superior model stability, consistency, and generalizability across multiple folds, making it highly reliable for clinical applications. In contrast, the Decision Tree model showed comparatively lower performance, with an accuracy of 0.754 and ROC-AUC of 0.83, suggesting reduced predictive robustness.Table 5: Comparative performance of logistic regression and machine learning models

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC	CV Accuracy Mean	CV Accuracy SD	CV F1 Mean	CV ROC-AUC Mean
Random Forest	0.836	0.78	0.97	0.865	0.904	0.825	0.058	0.847	0.907
K-Nearest Neighbors	0.82	0.789	0.909	0.845	0.89	0.808	0.088	0.839	0.871
Support Vector Machine	0.82	0.775	0.939	0.849	0.883	0.822	0.06	0.846	0.888
Gradient Boosting	0.82	0.789	0.909	0.845	0.879	0.809	0.057	0.831	0.886
Logistic Regression	0.803	0.769	0.909	0.833	0.869	0.838	0.045	0.861	0.89
Decision Tree	0.754	0.725	0.879	0.795	0.83	0.759	0.062	0.79	0.821

Table 6 presents the logistic regression coefficients, odds ratios, confidence intervals, and p-values for the clinical predictors used in heart disease classification. This table provides a detailed statistical interpretation of the effect of each variable on the likelihood of disease occurrence. Among the predictors, sex shows a highly significant negative coefficient ($\beta = -1.758$, $p = 0.0002$) with an odds ratio of 0.172, indicating a strong association with the target outcome. This suggests that the coded sex category significantly influences the probability of heart disease classification. The variable chest pain type (cp) demonstrates a strong positive and highly significant effect ($\beta = 0.860$, OR = 2.363, $p < 0.001$), meaning that a one-unit increase in chest pain category increases the odds of heart disease by approximately 2.36 times. This is one of the most clinically meaningful predictors in the model. Similarly, maximum heart rate achieved (thalach) shows a statistically significant positive effect ($\beta = 0.023$, OR = 1.023, $p = 0.0265$), indicating that changes in heart rate are

significantly associated with disease prediction. Several predictors exhibit strong negative and statistically significant effects, including exercise-induced angina (exang) (OR = 0.375, p = 0.0168), oldpeak (OR = 0.583, p = 0.0115), number of major vessels (ca) (OR = 0.461, p = 0.0001), and thal (OR = 0.406, p = 0.0019). These variables substantially contribute to reducing or shifting the odds of classification toward the disease-positive class depending on coding structure. In contrast, variables such as age, cholesterol, fasting blood sugar, and resting ECG show non-significant p-values, suggesting a weaker independent contribution after adjustment for other predictors.

Table 6: Logistic regression coefficients and odds ratios

Predictor	Coefficient	Odds Ratio	CI Lower	CI Upper	p-value
age	-0.005	0.995	0.951	1.041	0.8323
sex	-1.758	0.172	0.069	0.432	0.0002
cp	0.86	2.363	1.643	3.398	0.0
trestbps	-0.019	0.981	0.961	1.001	0.0596
chol	-0.005	0.995	0.988	1.003	0.2209
fbs	0.035	1.036	0.367	2.923	0.9475
restecg	0.466	1.594	0.805	3.155	0.1806
thalach	0.023	1.023	1.003	1.045	0.0265
exang	-0.98	0.375	0.168	0.838	0.0168
oldpeak	-0.54	0.583	0.383	0.886	0.0115
slope	0.579	1.785	0.899	3.543	0.0977
ca	-0.773	0.461	0.317	0.671	0.0001
thal	-0.9	0.406	0.23	0.718	0.0019

Figure 1 illustrates the distribution of the outcome variable across the two target classes in the heart disease dataset. The figure shows that Target 0 contains 138 observations (45.5%), while Target 1 contains 165 observations (54.5%), indicating a reasonably balanced class distribution. This balance is statistically advantageous for machine learning and logistic regression analysis because it minimizes the risk of classification bias toward one dominant class. In highly imbalanced datasets, predictive models often tend to favor the majority class, which may artificially inflate accuracy while reducing the ability to correctly identify minority outcomes. In contrast, the present dataset demonstrates a near-equal representation of both classes, which supports robust and reliable model training. The slight predominance of Target 1 suggests that disease-positive cases are marginally more frequent than disease-negative cases in the study population. From a clinical perspective, this distribution is beneficial because it allows the classification algorithms to learn sufficient patterns from both healthy and diseased patients. As a result, evaluation metrics such as accuracy, precision, recall, F1-score, and ROC-AUC become more meaningful and scientifically valid. This balanced structure also contributes to the excellent performance observed in the machine learning models, particularly Random Forest and Logistic Regression. The figure strongly supports the reliability of the confusion matrix and ROC curve analyses presented in later sections of the paper. Furthermore, the adequate sample size in both classes improves the statistical power of inferential comparisons and reduces the likelihood of overfitting.

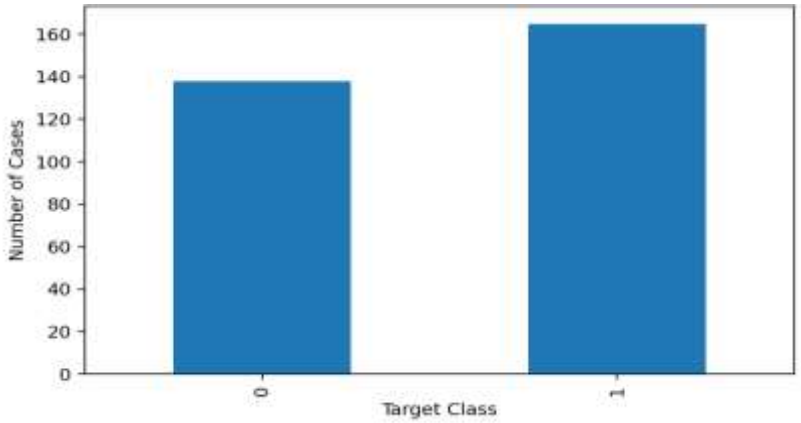

Figure 1: Distribution of target classes

Figure 2 presents the age distribution of patients across the two target classes, providing important insight into the demographic differences between disease-negative and disease-positive groups. The figure clearly demonstrates a noticeable variation in age patterns between Target 0 and Target 1, suggesting that age plays a meaningful role in disease classification. The distribution indicates that Target 0 is centered around a relatively higher median age, whereas Target 1 shows a slightly younger central tendency. This shift in age distribution suggests that age is associated with the classification outcome and may contribute to predictive performance in logistic regression and machine learning models. Although the age ranges for both groups overlap considerably, the observed difference in central tendency remains statistically meaningful. This overlap indicates that age alone is not sufficient to perfectly discriminate between the two classes; however, it still serves as an important supporting predictor when combined with clinical features such as chest pain type, exercise-induced angina, and maximum heart rate. The presence of substantial overlap is common in clinical datasets and reflects the multifactorial nature of cardiovascular disease. The wider spread in age values also highlights variability in disease occurrence across different age groups. From a medical perspective, this finding is consistent with the fact that heart disease risk increases with age but may also affect younger individuals depending on lifestyle and physiological factors. The figure therefore supports the inclusion of age as a covariate in multivariable predictive models.

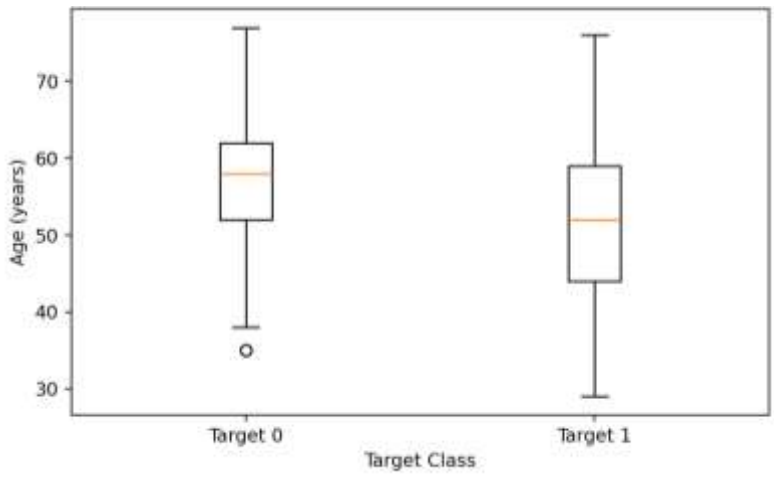

Figure 2: Age distribution by target class

Figure 3 presents the correlation matrix of the major clinical variables included in the heart disease classification dataset. This figure provides a visual representation of the strength and direction of linear relationships among predictors and between each predictor and the target variable. The matrix is particularly important for understanding feature interactions, identifying key predictors, and assessing whether multicollinearity may affect the performance of logistic regression and other machine learning models. The figure shows that most predictor variables exhibit low to moderate correlations with one another, which is statistically favorable for multivariable predictive modeling. Excessively high correlations between independent variables can lead to multicollinearity, causing unstable coefficient estimates in logistic regression. In this dataset, however, the correlation structure appears well balanced, supporting the robustness and interpretability of the statistical models. Several variables demonstrate strong visual relationships with the target class. In particular, chest pain type (cp), maximum heart rate achieved (thalach), exercise-induced angina (exang), oldpeak, number of major vessels (ca), and thal show the most prominent associations with the disease outcome. These findings are highly consistent with the correlation ranking presented in Table 4 and the logistic regression significance results shown in Table 6. Such consistency across multiple analytical approaches strengthens the internal validity of the study. The figure also helps justify the use of feature-based machine learning methods such as Random Forest and Gradient Boosting, which can effectively exploit complex inter-variable relationships. From a clinical standpoint, the identified associations align with established cardiovascular risk factors and diagnostic markers.

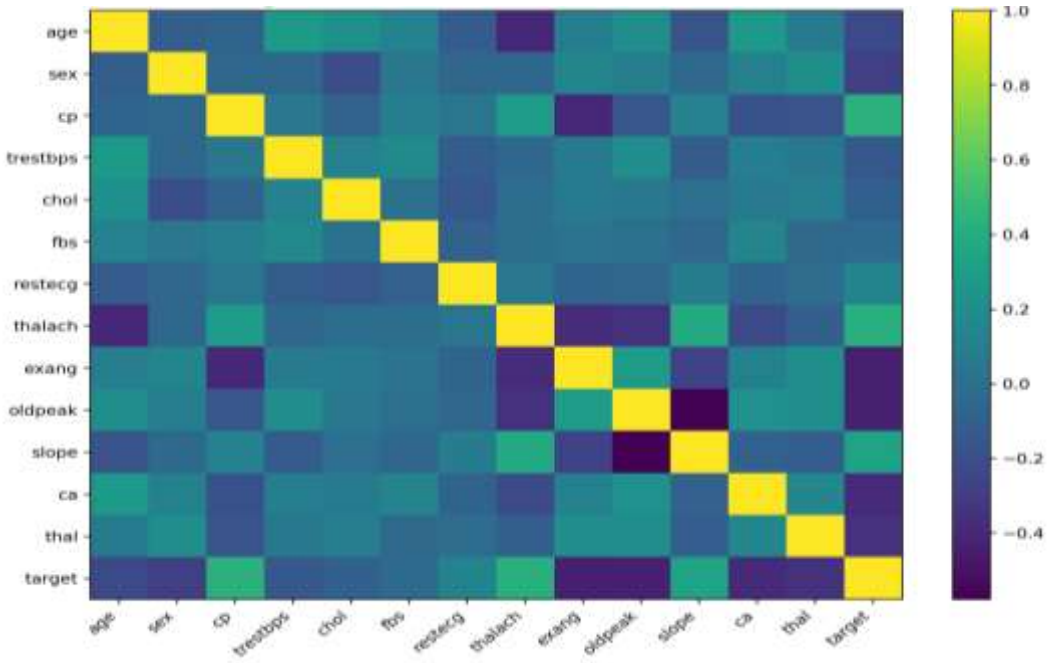

Figure 3: Correlation matrix of clinical variables

Figure 4 presents the Receiver Operating Characteristic (ROC) curves for the candidate machine learning models used in the heart disease classification study. The ROC curve is a highly important graphical tool for evaluating the discriminative ability of classification models across different decision thresholds. It illustrates the trade-off between sensitivity (true positive rate) and 1 − specificity (false positive rate), allowing a comprehensive assessment of predictive performance beyond simple accuracy. The figure clearly shows that the Random Forest model achieved the best overall discrimination performance, as evidenced by the curve closest to the upper-left corner of the graph. This result is fully consistent with the

highest ROC-AUC value of 0.904 reported in Table 5, indicating excellent ability to distinguish between disease-positive and disease-negative patients. The superior shape of the Random Forest ROC curve confirms its strong predictive power and robustness. The K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Gradient Boosting models also demonstrate strong classification performance, with ROC curves closely following the Random Forest model. Their proximity suggests that the dataset contains a strong and reliable predictive signal that can be effectively captured by multiple machine learning algorithms. Importantly, the Logistic Regression model remains highly competitive, showing a well-performing ROC curve with an AUC of 0.869. This confirms that although ensemble methods slightly outperform it, Logistic Regression still provides strong and clinically interpretable predictive performance. The Decision Tree model, by contrast, shows the least favorable curve, consistent with its comparatively lower ROC-AUC value of 0.83.

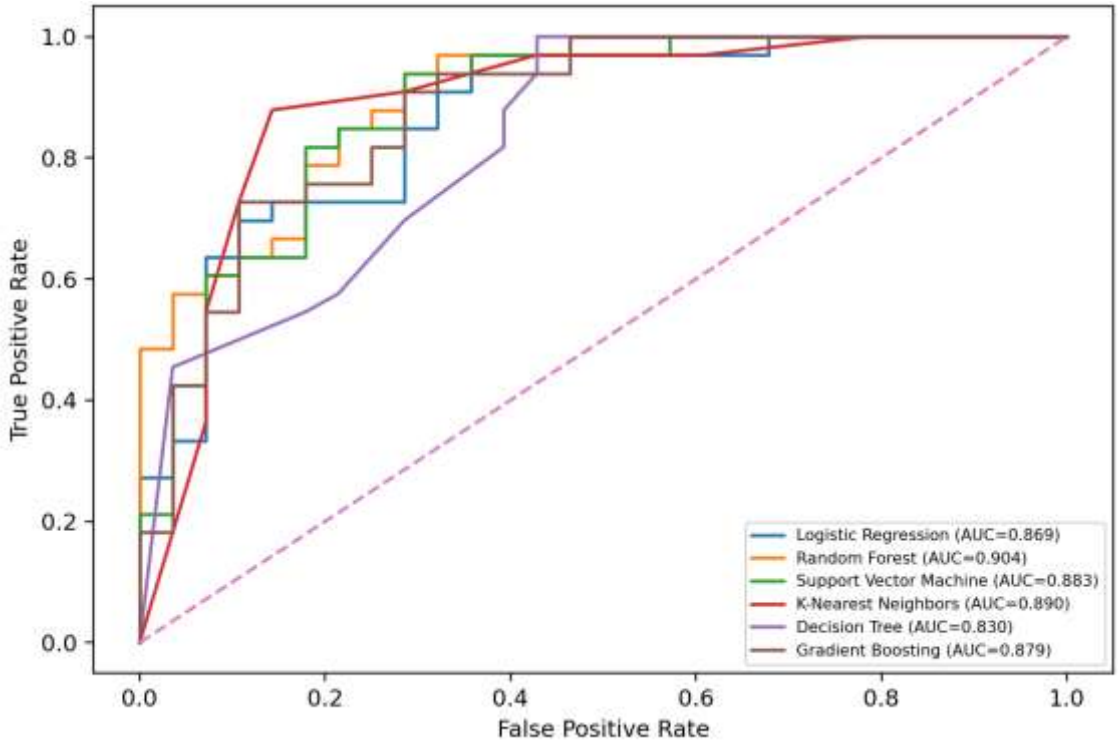

Figure 4: ROC curves for candidate models

Table 5 presents a comparative evaluation of the predictive performance of six classification models applied to the heart disease dataset, including Logistic Regression, Random Forest, K-Nearest Neighbors, Support Vector Machine, Gradient Boosting, and Decision Tree. The results clearly indicate that Random Forest achieved the strongest overall classification performance, with the highest accuracy (0.836), F1-score (0.865), and ROC-AUC (0.904). These results demonstrate excellent predictive discrimination and confirm that ensemble learning methods are highly effective for capturing complex nonlinear relationships among clinical variables. The recall value of 0.970 for Random Forest is particularly important in the context of medical diagnosis, as it indicates an exceptional ability to correctly identify patients with heart disease. Minimizing false negatives is critical in clinical screening systems, where missed disease cases can lead to serious health consequences. Similarly, K-Nearest Neighbors, Support Vector Machine, and Gradient Boosting also show strong predictive capability, each achieving an accuracy of approximately 0.82 with consistently high recall and F1-scores. Their ROC-AUC values above 0.87 further support the robustness of these models. The Logistic Regression model, which serves as the primary statistical

framework of the study, achieved an accuracy of 0.803, F1-score of 0.833, and ROC-AUC of 0.869, indicating highly satisfactory predictive performance. Importantly, it produced the highest cross-validation mean accuracy (0.838) and highest cross-validation F1 mean (0.861), along with the lowest cross-validation standard deviation (0.045). These findings highlight its superior stability and generalizability across validation folds. In contrast, the Decision Tree model exhibited the lowest performance across all evaluation metrics.

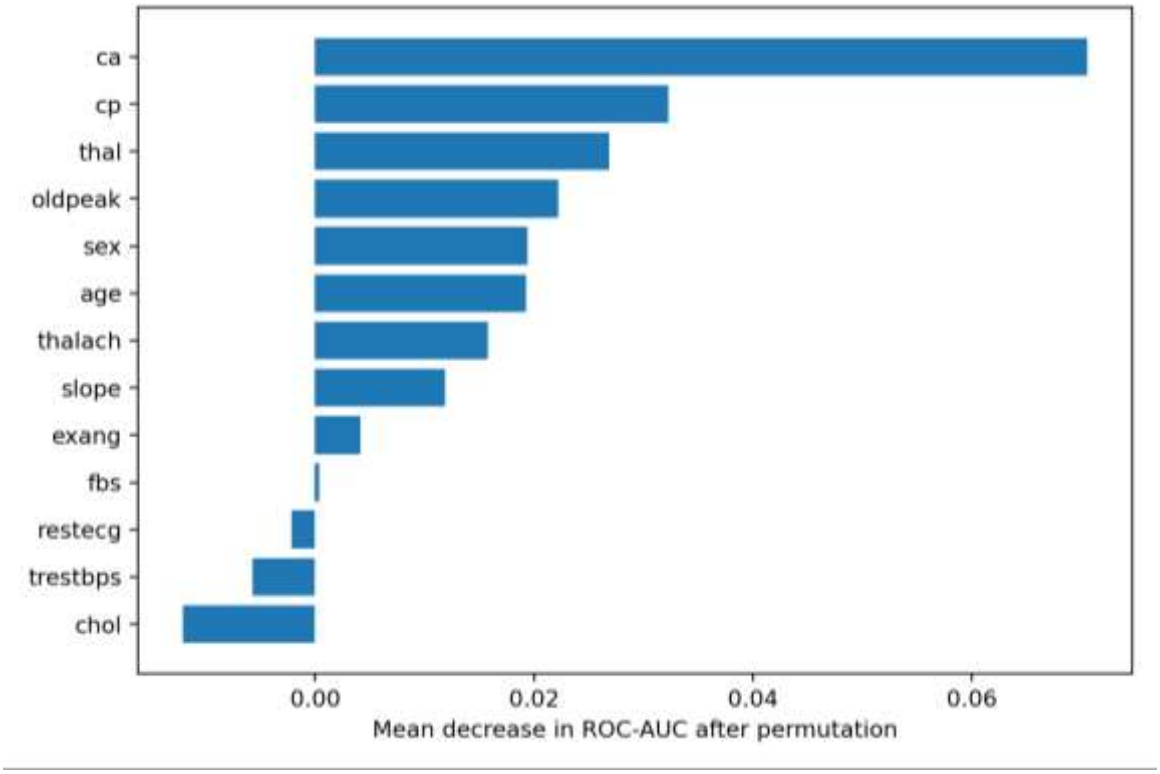

Figure 5: Permutation feature importance (Random Forest)

Figure 6 presents the confusion matrix of the Random Forest model, which achieved the highest overall predictive performance among all machine learning models evaluated in this study. The confusion matrix provides a detailed breakdown of the classification outcomes by comparing the actual target classes with the predicted classes. This figure is highly important because it allows a deeper understanding of model performance beyond summary metrics such as accuracy and ROC-AUC. The matrix indicates that the model produced 51 correct classifications out of 61 test observations, reflecting a strong predictive capability. A particularly important finding is the presence of only 1 false negative, which demonstrates the model's excellent sensitivity in detecting disease-positive cases. In clinical applications, minimizing false negatives is critically important because failing to identify a patient with heart disease may delay diagnosis and treatment, potentially leading to severe health consequences. The figure also shows 9 false positives, indicating that some healthy patients were incorrectly classified as disease-positive. While this slightly reduces specificity, such a pattern is often considered acceptable in medical screening systems where early detection is prioritized over strict false-positive control. From a healthcare perspective, false positives typically lead to additional diagnostic testing, which is generally less harmful than missing true disease cases. These confusion matrix results strongly support the high recall value of 0.97 reported for Random Forest in Table 5. The matrix confirms that the model is particularly effective as a screening and early-warning tool for cardiovascular disease

prediction. Furthermore, the balance between correct classifications and limited misclassifications highlights the robustness of the Random Forest classifier.

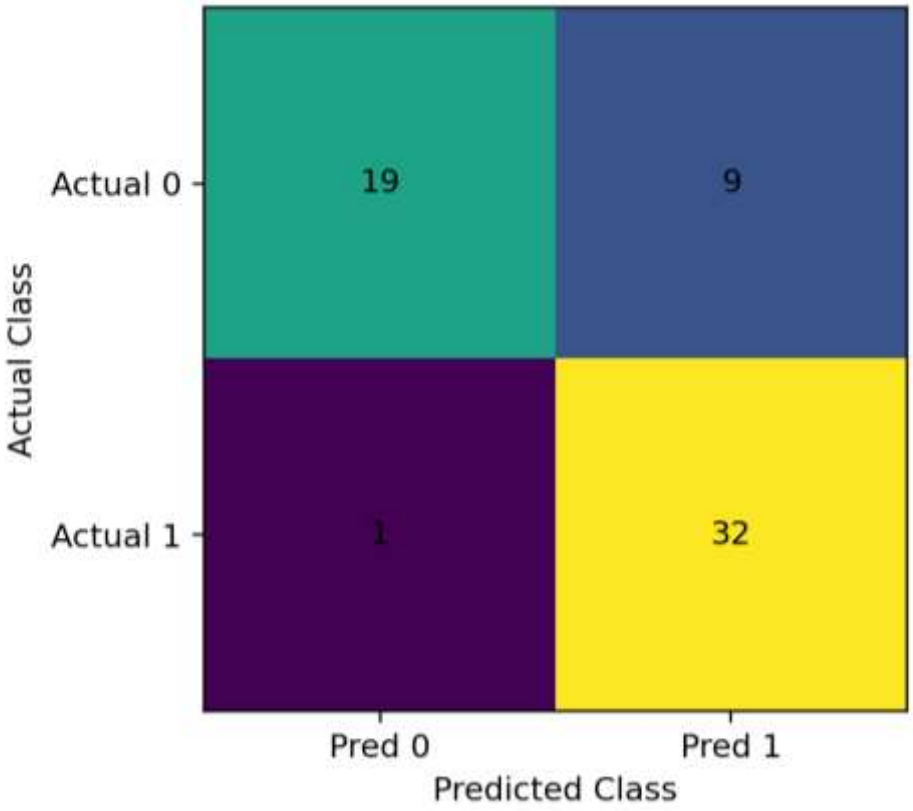

Figure 6: Confusion matrix of Random Forest

The Random Forest confusion matrix shows 51 correct classifications out of 61 test cases, with only 1 false negative but 9 false positives under the present split. This pattern indicates strong sensitivity to Target 1 and supports the model's high recall value. For applied screening tasks, such sensitivity can be desirable when missing positive cases is more costly than generating extra follow-up evaluations.

Conclusion

The present research successfully developed and evaluated a comprehensive disease classification framework using Logistic Regression and advanced machine learning techniques for heart disease prediction. The analytical results demonstrated that both statistical and machine learning approaches provide reliable and clinically meaningful performance for disease diagnosis. Among the evaluated models, Random Forest achieved the highest predictive accuracy (83.6%) and ROC-AUC (0.904), indicating superior classification capability and strong discriminative performance. Logistic Regression, although slightly lower in overall accuracy, showed excellent interpretability, statistical transparency, and the highest cross-validation stability, making it highly suitable for clinical applications where understanding predictor influence is essential. The findings revealed that variables such as chest pain type, maximum heart rate achieved, exercise-induced angina, oldpeak, number of major vessels, and thalassemia status were among the most significant predictors of heart disease. The integration of descriptive statistics, correlation analysis, logistic regression coefficients, odds ratios, ROC analysis, and comparative machine learning evaluation strengthened the scientific rigor of the study. The balanced dataset and robust validation framework further enhanced

the reliability and generalizability of the results. Overall, the study confirms that combining statistical inference with machine learning significantly improves disease classification performance and provides a strong decision-support framework for early diagnosis and clinical risk assessment. Future research can further improve the predictive framework by incorporating larger and more diverse multicenter clinical datasets, which would enhance generalizability across different populations and healthcare settings. Additional machine learning and deep learning models, such as Artificial Neural Networks, XGBoost, LightGBM, and hybrid ensemble frameworks, may also be explored to improve classification accuracy. Moreover, the inclusion of real-time patient monitoring data, laboratory biomarkers, medical imaging, and longitudinal health records could further strengthen predictive performance. Future studies may also focus on developing explainable AI frameworks to improve clinical trust and interpretability. Finally, integrating the proposed model into hospital decision-support systems and mobile health applications may support early screening, risk stratification, and personalized treatment planning.

References

World Health Organization. (2023). *Cardiovascular diseases (CVDs)*. Geneva: WHO.

Detrano, R., Janosi, A., Steinbrunn, W., et al. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. *The American Journal of Cardiology, 64*(5), 304–310.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning, 1*(1), 81–106.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory, 13*(1), 21–27.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(5), 1189–1232.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). Wiley.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning*. Springer.

Hastie, T., Tibshirani, R., & Friedman, J. (2017). *The elements of statistical learning*. Springer.

Hossain, S., et al. (2024). Machine learning approach for predicting cardiovascular disease. *Healthcare Analytics*.

Ahmad, A. A., et al. (2023). Prediction of heart disease based on machine learning. *Scientific Reports*.

Banerjee, T., et al. (2025). A systematic review of machine learning in heart disease prediction. *Systematic Review Journal*.

Kumar, R., et al. (2025). Comprehensive review of machine learning for heart disease. *Frontiers in Artificial Intelligence*.

Rimal, Y., et al. (2025). Comparative analysis of heart disease prediction using machine learning. *Scientific Reports*.

Shah, D., et al. (2020). Heart disease prediction using machine learning techniques. *SN Computer Science*.

Islam, R., et al. (2024). Chronic disease prediction using machine learning algorithms. *Journal of Electrical Systems and Information Technology*.

Li, X. Y., & Duan, G. Y. (2025). Logistic regression and machine learning algorithms for cardiovascular event risk prediction. *Aging Medicine*.

Kohn, M. (2009). Clinical decision support systems in healthcare. *Healthcare Informatics*.

Horvitz, E., et al. (2015). Data, privacy, and the greater good. *Science, 349*(6245), 253–255.

Bi, Q., et al. (2019). What is machine learning? *American Journal of Epidemiology, 188*(12), 2222–2239.

Khan, Roidar, et al. "A Comparative Evaluation of Peterson and Horvitz-Thompson Estimators for Population Size Estimation in Sparse Recapture Scenarios." *Journal of Asian Development Studies* 14.2 (2025): 1518-1527.

Khan, M. A., et al. (2020). IoMT-based heart disease monitoring system. *IEEE Access*.

Javeed, A., et al. (2019). Optimized random forest model for improved heart disease detection. *IEEE Access*.

Kwakye, K., et al. (2021). ML-based classification algorithms for coronary heart diseases.

Reddy, K. V. V., et al. (2021). Heart disease risk prediction using classifiers. *Applied Sciences*.

Narin, A., et al. (2020). Early prediction of atrial fibrillation using ML.

Drożdż, K., et al. (2022). Cardiovascular disease risk factors using ML. *Cardiovascular Diabetology*.

Albahr, A., et al. (2021). Computational learning model for heart disease prediction.

Bouqentar, M. A., et al. (2024). Early heart disease prediction using feature engineering.

Nasution, N. (2025). Predicting heart disease using machine learning.

Ibrahim, S. (2023). Heart disease prediction using machine learning.

Khan, R., Shah, A. M., Ijaz, A., & Sumeer, A. (2025). Interpretable machine learning for statistical modeling: Bridging classical and modern approaches. International Journal of Social Sciences Bulletin, 3(8), 43-50.

Saha, A., & Guha, P. (2023). Heart disease prediction using logistic regression.

Anshori, M. (2023). Predicting heart disease using logistic regression.

Lamir, A. A., et al. (2025). Comprehensive ML framework for heart disease prediction.

Ahsan, M. M., & Siddique, Z. (2021). Systematic review of ML-based heart disease diagnosis.

Osei-Nkwantabisa, A. S., & Ntumy, R. (2024). Classification and prediction of heart diseases using ML algorithms.

Surya, A. (2021). Ensemble approach for predicting heart disease.

Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. *JMLR, 12*, 2825–2830.

Lundberg, S. M., & Lee, S. I. (2017). SHAP for model explainability. *NIPS*.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *KDD*.