Predicting Heart Disease Risk Using Logistic Regression and Random Forest Models: Balancing Interpretability and Accuracy

Songyu Liu *

Department of Statistics and Data Science, University of California, Santa Barbara, CA 93106, United States

* Corresponding Author Email: Songyu@ucsb.edu

Abstract. Cardiovascular disease, particularly heart disease, continues to be a major contributor to global mortality, making timely detection a critical priority. Traditional diagnostic approaches often identify the illness only after significant progression, underscoring the need for predictive tools that rely on clinical and demographic information. This study utilizes the Cleveland subset of the UCI Machine Learning Repository, which contains 270 patient records with 14 attributes, to explore predictive modeling for heart disease. We developed and assessed two methods: logistic regression and random forest. Logistic regression achieved an accuracy of 85% and an AUC of 0.918, offering transparency in interpreting risk factors such as ST depression, abnormal vessel counts, and thallium test outcomes. The random forest model delivered a comparable accuracy of 85% with an AUC of 0.906, highlighting similar predictors while capturing nonlinear patterns within the data. Our findings indicate that combining interpretable models with machine learning techniques provides a balanced and reliable strategy to support early heart disease detection and improve clinical decision-making.

Keywords: Heart disease, Machine learning, Logistic regression, Random Forest, Predictive modeling.

1. Introduction

Cardiovascular diseases, particularly heart disease, have long been recognized as leading causes of death and disability worldwide. According to the American Heart Association's 2023 Statistical Update, cardiovascular disease remains the top cause of mortality globally. The report not only highlights the persistently high rates of death and disease but also emphasizes the central role of behavioral factors such as smoking, poor diet, physical inactivity, and obesity, along with clinical indicators including blood pressure, cholesterol, and glucose levels [1]. A similar pattern has been observed in China. Liu et al., drawing on data from the Global Burden of Disease Study, reported that the number of annual cardiovascular deaths in China increased from 2.51 million in 1990 to 3.97 million in 2016, with the total number of prevalent cases approaching 94 million. Furthermore, significant variation was noted across provinces, underscoring that cardiovascular disease is not only a global health challenge but also a pressing national issue for China [2].

From a research perspective, predicting the risk of heart disease has long been a critical area at the intersection of medicine and data science. As early as 1989, Detrano et al. introduced a probability-based algorithm for diagnosing coronary artery disease using the Cleveland dataset and validated it across patient cohorts in the United States, Hungary, and Switzerland. Their findings demonstrated the dataset's reliability and a certain degree of generalizability, establishing the Cleveland dataset as a foundational resource for subsequent research on heart disease prediction [3]. Since then, this dataset has remained one of the most widely used benchmarks for developing and testing predictive models.

In recent years, advances in artificial intelligence and big data technologies have brought machine learning into the forefront of cardiovascular prediction. Karna et al. reviewed a decade of research and concluded that traditional diagnostic approaches such as electrocardiograms and angiography have limited effectiveness in early detection, whereas machine learning and deep learning models have shown far greater potential and accuracy [4]. This shift reflects a growing reliance on data-

driven methods to improve healthcare decision-making. At the same time, researchers have also recognized that different algorithms vary significantly in performance. For instance, Riyaz et al. conducted a quantitative review of various machine learning techniques and reported that artificial neural networks achieved the highest average accuracy (86.91%), while decision tree models performed the worst (74.0%) [5]. These findings suggest that careful model selection is essential to balance predictive accuracy with interpretability and practicality.

Among the many algorithms explored, logistic regression and random forest are of particular importance. Logistic regression, as a classic statistical model, offers transparency and interpretability, enabling researchers and clinicians to clearly understand how each clinical variable contributes to the risk of heart disease. In contrast, random forest is a robust ensemble learning method that can capture complex nonlinear relationships and variable interactions, often delivering superior predictive performance. Azimi Lamir et al. provided a recent example by applying logistic regression, k-nearest neighbors, and random forest to the Cleveland dataset. Their results showed that random forest outperformed logistic regression with an accuracy of 91% and an F1-score of 0.89, while logistic regression remained valuable for its interpretability [6]. This highlights the trade-off between accuracy and explainability that continues to shape the field.

Building on this background, the present study emphasizes an additional consideration: not all clinical and demographic variables contribute equally to heart disease prediction. During preliminary data analysis, we performed visualization and statistical comparison to identify the features that differed most significantly between patients with and without heart disease. Based on this analysis, we selected nine primary factors out of the fourteen available in the Cleveland dataset to serve as the foundation for model building. This feature selection approach reduces redundancy, minimizes noise, and more closely reflects clinical practice, where physicians often prioritize the most critical risk indicators. On this basis, we developed and evaluated logistic regression and random forest models, comparing their performance in terms of accuracy, sensitivity, specificity, and interpretability.

The goal of this study is not merely to build another prediction model, but rather to demonstrate that a carefully selected subset of variables can yield strong predictive performance while improving interpretability and clinical practicality. By focusing on primary factors and systematically comparing two widely used methods, we aim to provide insights that may support more efficient risk assessment, resource allocation, and early intervention strategies in cardiovascular healthcare.

2. Methods

2.1. Data Source

The dataset used in this study is the Cleveland subset of the UCI Machine Learning Repository. It contains 270 patient records with 14 attributes, including demographic features (such as age and sex) and clinical measurements (such as resting blood pressure, serum cholesterol, maximum heart rate, and ST depression). The target variable is the presence or absence of heart disease, defined as a binary outcome, where 0 indicates no heart disease and 1 indicates the presence of heart disease.

2.2. Data Processing

During preprocessing, the dataset was first checked for completeness, and missing values were handled appropriately. Continuous variables were standardized to mitigate the effect of different measurement scales. Categorical variables (including sex, chest pain type, exercise-induced angina, electrocardiographic results, and thalassemia test outcomes) were encoded into dummy variables to make them suitable for analysis.

Importantly, this study did not use all 14 attributes directly. Instead, through visualization and statistical comparison, nine primary factors that exhibited the most significant differences between patients with and without heart disease were selected as input features for modeling. This feature selection step aimed to reduce redundancy, minimize noise, and improve interpretability, while also aligning with clinical practice in which physicians typically focus on key risk indicators.

2.3. Modeling Approach

Two predictive models were developed and compared: logistic regression and random forest: (1) Logistic Regression: As a classic statistical method, logistic regression quantifies the relationship between predictor variables and the probability of heart disease, offering interpretability for clinical applications. The model is based on the logistic function and parameters were estimated using maximum likelihood estimation. (2) Random Forest: As an ensemble learning approach based on decision trees, random forest aggregates multiple trees to improve prediction stability and accuracy. It captures nonlinear relationships and interactions between variables and provides estimates of variable importance.

The dataset was divided into training and testing sets using a 75% to 25% split. Specifically, 203 observations were allocated to the training set and 67 observations to the testing set. Stratified sampling was applied to preserve the balance between positive and negative cases. For logistic regression, default parameters were used in model fitting. For random forest, hyperparameters such as the number of trees and maximum tree depth were tuned using cross-validation to optimize model performance.

2.4. Model Evaluation

To comprehensively evaluate the model's predictive performance, this study employs multiple statistical metrics and visualisation techniques. Firstly, accuracy serves as an intuitive measure of overall performance, reflecting the proportion of correctly predicted samples across the entire dataset. Secondly, sensitivity (Recall) and specificity were calculated separately to evaluate the model's recognition capability across different categories: sensitivity measures the model's detection rate for positive cases, while specificity measures the model's correct exclusion rate for negative samples. Balancing these two metrics is particularly crucial for practical applications. Concurrently, a Receiver Operating Characteristic curve (ROC curve) is plotted and the Area Under the Curve (AUC) calculated. This serves as a comprehensive metric reflecting the model's overall discrimination capability across varying decision thresholds.

Furthermore, this study generated a confusion matrix to visually illustrate the distribution of predicted outcomes across categories. By comparing the quantities of true positives, false positives, true negatives, and false negatives, one can conduct a more detailed analysis of the model's strengths and weaknesses across different categories, thereby providing a basis for subsequent model optimisation and refinement.

3. Results

3.1. Exploratory Analysis

Exploratory analysis revealed clear differences between patients with and without heart disease across both continuous and categorical variables. Age, serum cholesterol, maximum heart rate, and ST depression showed significant variation between the two groups. Similarly, categorical variables such as sex, chest pain type, exercise-induced angina, ST slope, number of major vessels, and thallium test results also displayed distinct distributions. These patterns suggested that certain demographic and clinical indicators could play an important role in predicting heart disease.

To illustrate, Figure 1 shows the distribution of age by heart disease status. Patients with heart disease were generally older, with a mean age of 56.6 years, compared to 52.7 years among those without the disease. Likewise, Figure 2 presents the distribution of maximum heart rate. Individuals without heart disease reached a higher mean maximum heart rate (158.3 bpm), whereas those with heart disease averaged only 138.9 bpm. Together, these examples highlight that both age and exercise capacity are meaningful discriminators of disease presence.

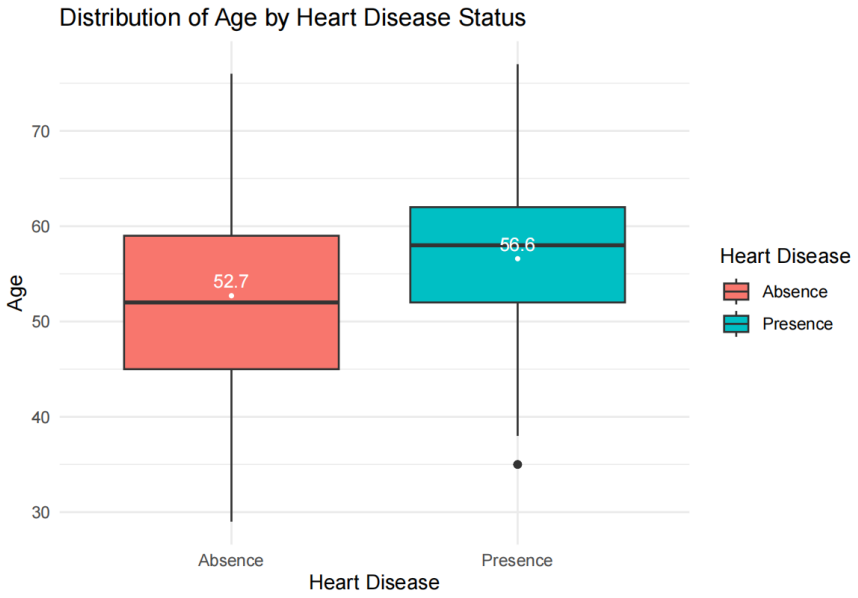

Fig 1. Distribution of age by heart disease status (Picture credit: Original)

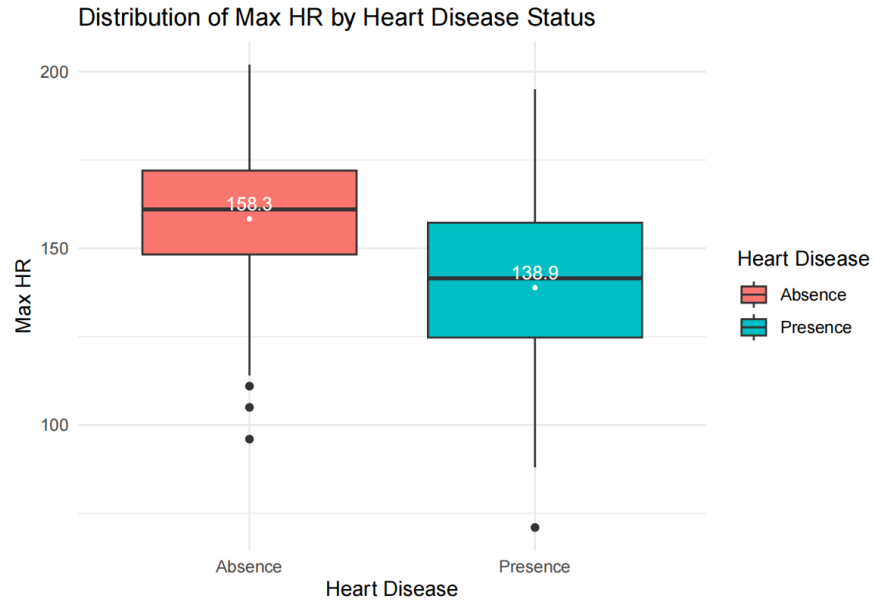

Fig 2. Distribution of maximum heart rate (MaxHR) by heart disease status (Picture credit: Original)

Overall, the exploratory findings confirmed that multiple demographic and clinical factors differed significantly between the two groups. These insights informed the subsequent feature selection step, where nine primary variables were chosen to develop predictive models.

3.2. Logistic Regression Model

The logistic regression model was trained on 203 samples and evaluated on 67 test samples. Several predictors were found to be statistically significant, including ST depression (oldpeak), sex (male), asymptomatic chest pain, number of major vessels (ca), and thallium test results (thal). The direction and magnitude of the coefficients confirmed known clinical risk patterns: for example, each one-unit increase in ST depression increased the odds of heart disease by approximately 36% ($p < 0.01$), while male patients had nearly four times the risk compared with females ($p < 0.01$). Patients with two abnormal vessels had an almost 29-fold increased risk, underscoring the strong predictive value of vessel abnormalities.

Model performance on the test set was strong. The confusion matrix (Table 1) shows that 31 individuals without disease and 26 with disease were correctly classified. Misclassifications were limited to four false positives and six false negatives. Overall, the model achieved 85% accuracy, with a sensitivity of 87% and a specificity of 84%.

In addition, the ROC curve (Figure 3) highlights the discriminative ability of the model across different thresholds. The area under the curve (AUC) was 0.918, indicating excellent performance in distinguishing between patients with and without heart disease. These results suggest that logistic regression not only provides accurate predictions but also maintains high interpretability, making it well suited for clinical applications where transparency is critical.

Table 1. Confusion Matrix (Test Set, threshold = 0.5, Positive = Presence)

	Absence	Presence
Absence	31	4
Presence	6	26

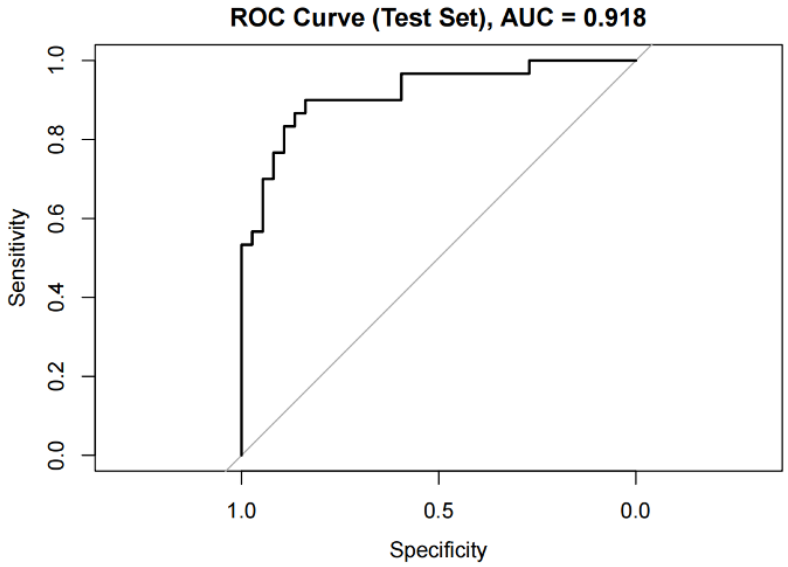

Fig 3. ROC curve of logistic regression model (Picture credit: Original)

3.3. Random Forest Model

The random forest model was trained and evaluated using the same 75/25 train-test split as logistic regression. On the test set, the model achieved an accuracy of 85% and an AUC of 0.906, which is comparable to the performance of logistic regression. The confusion matrix (Table 2) shows that 31 patients without disease and 24 with disease were correctly classified. Misclassifications included six false positives and six false negatives, indicating a balanced error distribution across classes.

The ROC curve (Figure 4) illustrates the model's discriminative capacity, with an AUC of 0.906, confirming that random forest provides reliable classification performance even with complex and nonlinear patterns in the data.

Table 2. Random Forest: Confusion Matrix (Test Set)

	Absence	Presence
Absence	31	6
Presence	6	24

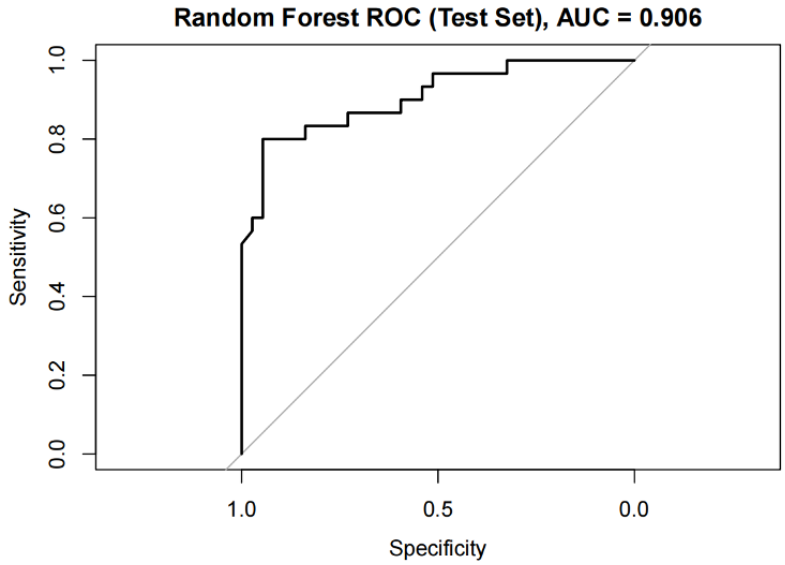

Fig 4. ROC curve of random forest model (Picture credit: Original)

Variable importance analysis (Figure 5) revealed that the number of abnormal vessels, thallium test results, and chest pain type were the most influential predictors. Maximum heart rate and ST depression also contributed meaningfully, while sex and age had relatively lower importance. These findings are largely consistent with the logistic regression model, though random forest additionally emphasizes the role of chest pain type and thallium results.

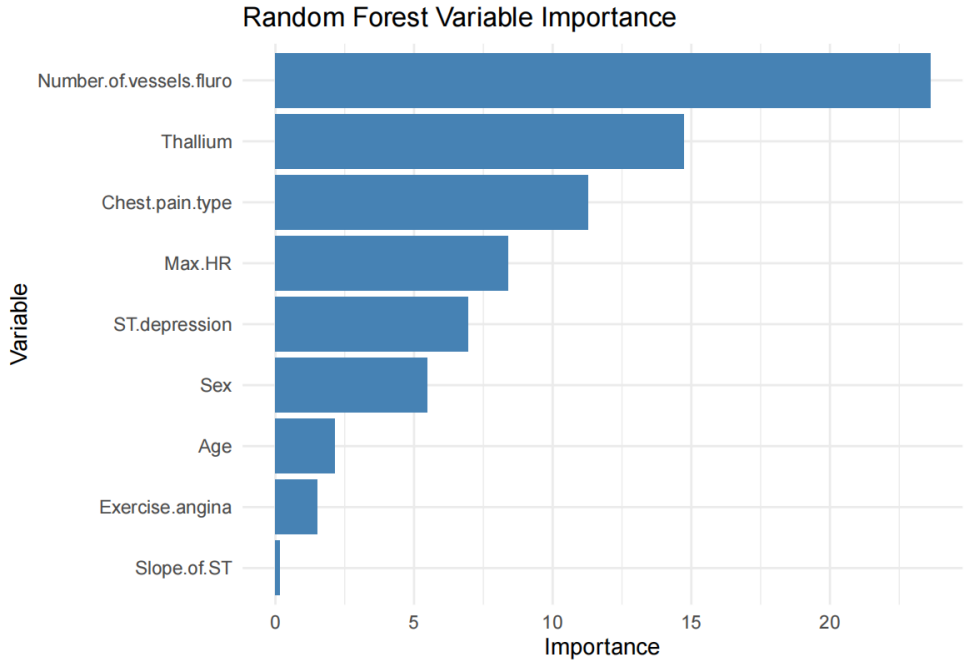

Fig 5. Variable importance plot of random forest model (Picture credit: Original)

Together, these results suggest that random forest provides robust and accurate classification performance while capturing complex variable interactions, making it well-suited for automated prediction systems.

3.4. Model Comparison

Logistic regression and random forest showed comparable overall performance. On the test set, both models achieved an accuracy of 0.85, while their AUC values were 0.918 and 0.906, respectively, confirming excellent discriminative ability (Table 3). Logistic regression demonstrated slightly higher sensitivity (0.87 vs. 0.80), indicating stronger ability in detecting patients with heart disease.

Both models maintained a specificity of 0.84, suggesting equivalent performance in identifying individuals without disease.

Table 3. Model Performance Comparison

Model	Accuracy	Sensitivity	Specificity	AUC
Logistic Regression	0.85	0.87	0.84	0.918
Random Forest	0.85	0.80	0.84	0.906

In terms of application, logistic regression offers superior interpretability, allowing clinicians to quantify the effect of each predictor and communicate risks more effectively. In contrast, random forest excels in capturing nonlinear relationships and complex interactions, making it suitable for automated prediction and large-scale deployment. Importantly, both models identified highly consistent key predictors, which reinforces the robustness and credibility of the findings.

4. Summary

This study used the Cleveland dataset from the UCI Machine Learning Repository to investigate the role of feature selection and model comparison in predicting heart disease. From 14 available attributes, we selected 9 variables that showed the most significant differences between patients with and without heart disease and applied them to logistic regression and random forest models to balance predictive accuracy with interpretability. The results showed that logistic regression identified ST depression, sex (male), asymptomatic chest pain, and number of abnormal vessels as key predictors, achieving strong discriminative ability with an AUC of 0.918. The random forest model produced comparable performance (AUC=0.906), confirming the importance of the same variables while demonstrating advantages in capturing nonlinear patterns and maintaining robustness. The complementarity of the two approaches indicates that predictive modeling can simultaneously provide clinical interpretability and reliable predictive power.

In summary, this study demonstrates that careful variable selection combined with appropriate modeling methods can deliver accurate and interpretable predictions for heart disease. These findings provide valuable insights for early risk assessment and clinical decision-making. Future work could extend this approach with larger longitudinal datasets and incorporate data from wearable devices and mobile health technologies to further enhance its practicality and scalability.

References

[1] Tsao C W, Aday A W, Almarzooq Z I, et al. heart disease and stroke statistics—2023 update: a report from the American Heart Association[J]. Circulation, 2023, 147(8): e93-e621.

[2] Liu S, Li Y, Zeng X, et al. Burden of cardiovascular diseases in China, 1990–2016: findings from the 2016 Global Burden of Disease Study[J]. JAMA Cardiology, 2019, 4(4): 342-352.

[3] Detrano R, Janosi A, Steinbrunn W, et al. international application of a new probability algorithm for the diagnosis of coronary artery disease[J]. American Journal of Cardiology, 1989, 64(5): 304-310.

[4] Karna V V R, Reddy K, Shyamala K, et al. A comprehensive review on heart disease risk prediction using machine learning and deep learning algorithms[J]. Archives of Computational Methods in Engineering, 2025, 32(3): 1763-1795.

[5] Riyaz L, Reddy P K, Bhatia M P S, et al. heart disease prediction using machine learning techniques: a quantitative review[C]//Proceedings of the International Conference on Innovative Computing and Communications (ICICC). Singapore: Springer, 2021: 627-639.

[6] Azimi Lamir A, Razzagzadeh S, Rezaei Z. A comprehensive machine learning framework for heart disease prediction: performance evaluation and future perspectives[J]. arXiv preprint arXiv:2505.09969, 2025.