Comparison of Prediction Models for Heart Disease Data: Logistic Regression, Random Forest and Extreme Gradient Boosting

Jiaming Tang*

School of Life Science and Engineering, Southwest Jiaotong University, Sichuan, 611756, China

*Corresponding author: tang113583@my.swjtu.edu.cn

Abstract. This paper aims to use predictive models such as the Logistic Regression (LR), Random Forest (RF) and Extreme Gradient Boosting (XGBoost) to predict the onset of cardiovascular disease based on relevant risk factors and make a comparison of the performance of these three models. First, this paper examined the statistical significance of the initial variables and the linear correlations between individual variables. To prepare the data for modeling, this paper first normalized all numerical and binary variables to ensure they were on a comparable scale. Dimensionality was then reduced using Principal Component Analysis (PCA), with the goal of simplifying the feature space while retaining meaningful variance. For the Random Forest (RF) model, parameter tuning prioritized minimizing RMSE, though the selected configuration also led to strong classification accuracy. XGBoost was optimized through a stepwise grid search combined with stratified k-fold validation, helping to maintain consistent performance across data splits. When comparing Logistic Regression, RF, and XGBoost, the latter consistently achieved the best classification results-likely due to its ability to capture complex and non-linear feature relationships. While these early findings are encouraging, especially for XGBoost, further validation on broader and more diverse datasets will be important before drawing firm conclusions about its clinical utility.

Keywords: Logistic regression; random forest; extreme gradient boosting.

1. Introduction

Cardiovascular diseases (CVDs), including various forms of heart disease, remain the leading cause of illness and death globally. Conditions such as coronary artery disease, heart failure, and stroke continue to impact millions each year. As the heart is central to sustaining life, early identification of cardiovascular risk is essential for both preventive care and timely medical action. More people die from CVDs than from any other disease, which highlights the urgent need for coordinated global health strategies [1]. Traditionally, clinical assessments rely on indicators like age, blood pressure, cholesterol levels, and lifestyle habits. However, recent developments in mathematical modeling and machine learning have introduced powerful tools for predicting individual risk. Timely and precise diagnosis is critical to guiding treatment and improving health outcomes. Modern diagnostic approaches-ranging from ECG and echocardiography to biomarkers and AI-assisted systems-have significantly improved early detection [2]. Beyond physical health, heart disease affects mental well-being, daily functioning, and healthcare costs, placing substantial strain on both individuals and public health systems [3, 4].

In recent years, there has been increasing interest in the use of machine learning (ML) for medical prediction, offering new possibilities for how diseases are diagnosed, monitored, and managed. These approaches draw on large and complex patient datasets to uncover patterns and associations that may be overlooked by traditional statistical methods. Techniques such as logistic regression, decision trees, and support vector machines (SVM) are among the most frequently employed in clinical predictive modeling, and have shown promise across a range of healthcare applications. Logistic regression, a fundamental statistical method, is particularly useful for binary classification problems, such as predicting the presence or absence of a disease based on patient data. Decision trees offer a straightforward and interpretable framework for medical diagnosis, translating complex clinical reasoning into clear, rule-based steps. Their simplicity makes them particularly useful in settings

where transparency is important. Support vector machines (SVMs), by contrast, are well-suited for analyzing high-dimensional data and have demonstrated strong performance in classifying various medical conditions-especially when the data involves intricate, nonlinear patterns [5, 6]. Deep learning models have additionally shown promise in evaluating intricate medical data and offering precise diagnostic insights; nevertheless, the problem of choosing a suitable model for specific populations and ensuring its clinical interpretability still arises [7-9].

About the use of mathematical statistics to model heart disease conditions, countless studies will provide answers. In order to create a training set of 220 individuals and a test set of 55 individuals, Feng randomly separated all samples into two sections: 80% for the training set and 20% for the test set. Statistical tests were first conducted to identify significant variables. To create a clinical prediction model, the data were standardized and processed, and the training set data was used to assess predictive features. The prediction models were validated using the test set, and clinical benefit was evaluated. Feng used R software to build logistic regression and random forest models based on key predictive variables [10].

This study proposes a comprehensive predictive framework for assessing heart disease risk by integrating a diverse array of clinical and diagnostic variables. Key features include demographic factors (such as age and sex), vital cardiovascular indicators (blood pressure, cholesterol levels, fasting blood glucose), and functional test results (electrocardiogram readings, maximum heart rate, and exercise-induced angina). Additionally, advanced imaging data-namely angiographic findings (including the number of major vessels affected) and Thallium stress test outcomes-are incorporated to provide a more granular characterization of cardiovascular pathology. To enhance predictive accuracy and model interpretability, this paper adopts a hybrid machine learning approach that combines Least Absolute Shrinkage and Selection Operator (LASSO) for feature selection, Random Forest for capturing complex nonlinear interactions, and Logistic Regression for probabilistic risk estimation. This integrative strategy aims to support more precise, data-driven clinical decision-making in the early identification and management of coronary artery disease.

2. Methods

2.1. Data Description

The heart disease database is available on the Kaggle platform, which provided by the University of California Irvine's Machine Learning Repository. This dataset contains 270 observations and 13 variables in the .CVS format and does not have any missing value.

2.2. Variables and Data Preprocessing

A total of 270 samples and 13 variables are in the raw dataset (Table 1).

Table 1. Variable Explanation

Name of Variables	Definition of the Variables	Range
Age	Age of the sample	29-79
Sex	Gender of the sample	0-1
Chest pain type	Type of chest pain experienced by the sample	1-4
BP	Blood Pressure of the sample	94-200
Cholesterol	Serum cholesterol level of the sample	126-564
FBS over 120	Fasting Blood Sugar level, specifically over 120 mg/dl	0-1
EKG results	Results from an Electrocardiogram test	0-2
Max HR	Maximum Heart Rate	71-202
Exercise angina	Indicates presence or absence of chest pain (angina) induced by exercise	0-1
ST depression	Depression level of the ST segment	0-6.2
Slope of ST	The slope or trend of the ST segment	1-3
Number of vessels fluro	Number of major blood vessels	0-3
Thallium	heart muscle's blood flow	3-7

For data preprocessing in Table 1, this paper replaced binary variables with "1" and "0". As all of the rest variables are numerical variables, this paper normalized them by dividing by the maximum value. And of all the data in this paper, there are no variables with missing data.

2.3. Models Introduction

This paper chooses three models which are the logistic regression model (LR), random forest model (RM), and XGBoost model to analyze and process the data. Testing and training data are chosen at random in this article at a 2:8 ratio. This paper decides to use accuracy by the confusion matrix of 2*2 for the evaluation of the models. The specific definition of the confusion matrix and the calculation formula are shown in Table 2 and Equation (1).

Table 2. Confusion Matrix

	Prediction Condition		
		Positive (P)	Negative (N)
Actual Condition	Negative (N)	False Positive (FP)	True Negative (TN)
	Positive (P)	True Positive (TP)	False Negative (FN)

$$Accuracy = \frac{TN+TP}{TP+FN+FP+TN} \quad (1)$$

The LR model creates a linear model, use maximum likelihood estimation to solve the weight parameters associated with each feature, and then assess the model's performance using metrics like the confusion matrix.

The RF model obtains the final model through cross-validation and use testing data for verification. In the XGBoost model, it uses a series of iterations to build the prediction model. The performance of XGBoost is primarily attributed to the optimization of the objective function, which utilizes both first-order gradients and second-order gradients, referred to as Hessians. And after tree constructions and regularization, it will finally get the fixed model.

3. Results and Discussion

3.1. Data Distribution

Figure 1 presents a combined histogram of all variables. It can be seen from the figure that dataset contains the samples aged from 30 to 80, which means there may be lack of samples of the young below 30 if the models are used to cover common age categories. There are also much more male samples than female ones in the dataset. Over all samples, obviously about half of the samples suffer from the worst level of chest pain, which may be associated with the incidence of the heart disease because only a small ratio of samples indicates their presence of chest pain induced by exercise. The majority of blood pressure values fall within the normal range of fluctuation. Above all the variables it turns to the heart disease showing more samples' absence than presence.

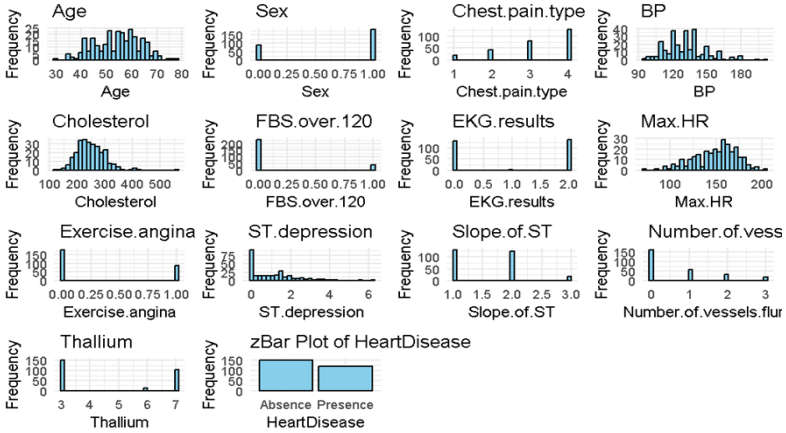

Fig. 1 Histogram of all variables

3.2. Variable Selection

Since there are totally 13 variables, it will be quite a high number for models if all the variables selected in models. A large number of variables always introduce a great deal of inconsistency for the models, so that they may influence the performance of the models. As the Table 3 shows, this paper uses the Variance Inflation Factor (VIF) value to verify if the variables exhibit multicollinearity and make Initial determination on the variable's merging or deletion. Following testing, no significant multicollinearity was found because all VIF values were less than 5. After finishing VIF value test, this paper continues to test the significance level of the p-value to determine whether the association between the variable and heart disease is statistically significant. As Table 3 shows, the P value of Cholesterol and FBS over 120 are larger than 0.05, so these two variables are excluded in the following steps (0.0000 means less than 0.0001).

Table 3. VIF value and P value of variables

Name of Variables	P Value	VIF Value
Age	0.0006	1.50
Sex	0.0000	1.32
Chest pain type	0.0000	1.29
BP	0.0119	1.21
Cholesterol	0.0563	1.17
FBS over 120	0.7886	1.08
EKG results	0.0029	1.09
Max HR	0.0000	1.65
Exercise angina	0.0000	1.37
ST depression	0.0000	1.82
Slope of ST	0.0000	1.78
Number of vessels fluro	0.0000	1.32
Thallium	0.0000	1.51

3.3. Variable Filtering

This paper also uses Principal Component Analysis (PCA) to continue to integrate and process variables again after checking VIF and P value. It can be seen according to the Table 4 and Figure 2 that the cumulative contribution arrived at 85% at pc7, so the number of variables after PCA is 7, and in the subsequent research these 7 variables are participated for training and testing the models.

Table 4. Importance of Components

Feature	Standard deviation	Proportion of Variance	Cumulative Proportion
pc1	1.7380	0.2746	0.2746
pc2	1.1907	0.1289	0.4035
pc3	1.0764	0.1053	0.5088
pc4	1.0238	0.0953	0.6041
pc5	0.9795	0.0872	0.6913
pc6	0.9209	0.0771	0.7684
pc7	0.8566	0.0667	0.8351
pc8	0.7614	0.0527	0.8878
pc9	0.6906	0.0434	0.9312
pc10	0.6384	0.0370	0.9682

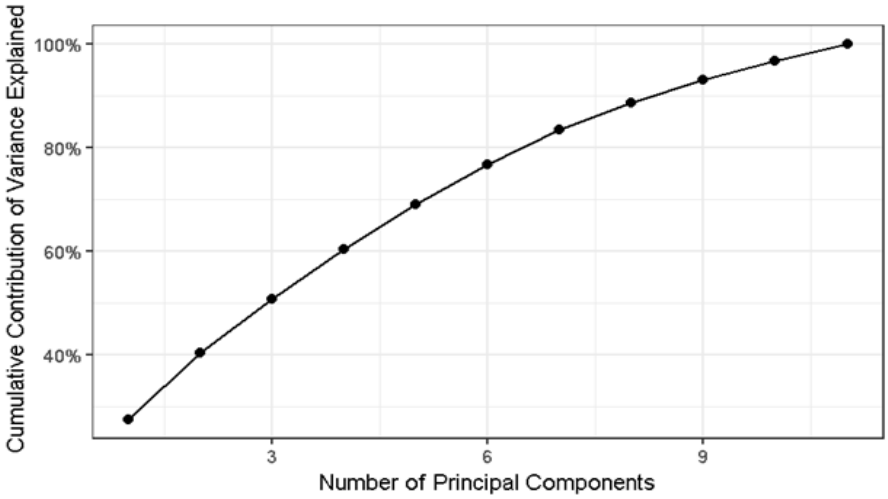

Fig. 2 Cumulative contribution variance plot

3.4. Logistic Regression Results

As confirmed by previous validation, there is no multicollinearity among the original variables; therefore, the principal components derived from PCA are also free from multicollinearity, so this paper does not check the multicollinearity again. To ensure the reproducibility of the modeling results, a random seed to "100" was employed during the model development process. The Generalized Linear Model (GLM) function was employed during the training phase of the model. Table 5 shows the result of the logistic regression model after using testing data. "Coefficient" column means the coefficient of each variable in the left column.

Table 5. Logistic Regression Coefficient

Feature	Coefficient	Std.Error
Intercept	-0.3244	0.2162
pc1	-1.4092	0.1920
Pc2	-0.3999	0.1690
Pc3	-0.2483	0.1847
Pc4	0.7586	0.2245
Pc5	-0.3206	0.2138
Pc6	0.0272	0.2135
Pc7	0.5162	0.2724

An intercept of -0.3244 indicates that, when all input features are equal to zero, the log-odds of heart disease occurrence is -0.3244, corresponding to an estimated probability of approximately 41.97%. The regression coefficients corresponding to PC1 through PC7 are -1.4092, -0.3999, -0.2483, 0.7586, -0.3206, 0.0272 and 0.5162. In logistic regression, the coefficient associated with each predictor describes how changes in that variable relate to the likelihood of the outcome. Specifically, a positive coefficient indicates that, all else being equal, an increase in the predictor is linked to higher log-odds-and thus a greater chance-of the event occurring. In contrast, a negative coefficient reflects a decrease in the predicted likelihood. The size of the coefficient, in absolute terms, gives a sense of how strongly the predictor influences the outcome.

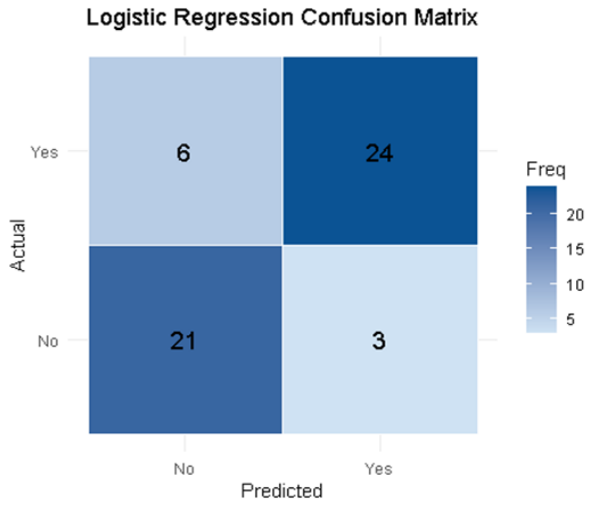

Fig. 3 Confusion matrix of LR model

As shown in Figure 3, the trained model ultimately achieved the classification accuracy of 0.8333, as indicated by the test results. According to the confusion matrix, 21 participants actually without heart disease were correctly classified, 3 were incorrectly predicted as having heart disease, 6 participants with heart disease were misclassified as not having the condition, and 24 were correctly identified as having heart disease.

3.5. Random Forest Model Results

Prior to model construction, preliminary tuning of the random forest parameters was performed to minimize the root mean square error (RMSE) and thus enhance overall model accuracy. Following the evaluation, the optimal configuration was determined to be ntree = 100 and mtry = 6, yielding an RMSE of 0.3427. In constructing the random forest model, still a random seed of"100" was consistently used to ensure the reproducibility of the results.

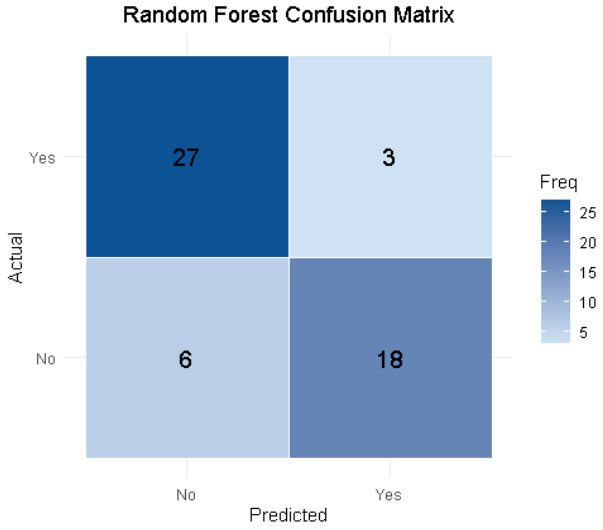

Fig. 4 Confusion matrix of RF model

As shown in Figure 4, the Random Forest model achieved a test accuracy of only 0.1667, indicating that the chosen parameter settings might have limited its predictive capability. Further parameter optimization or feature selection may be required to enhance the model's effectiveness. According to the confusion matrix, 6 participants actually without heart disease were correctly classified, 18 were incorrectly predicted as having heart disease, 27 participants actually with heart disease were misclassified as not having the condition, and 3 were correctly identified as having heart disease.

3.6. XGBoost Model Results

During the construction of the XGBoost model, hyperparameter tuning was performed to optimize predictive performance. The final configuration adopted a tree-based booster (gbtree) and included 100 boosting rounds (nrounds). To control model complexity and mitigate overfitting, the maximum tree depth (max_depth) was set to 7 and the regularization term (gamma) was set to 1, which enforces a minimum loss reduction before further splitting. A learning rate (eta) of 0.1 was used to balance learning speed and convergence.

Additional parameters included a column subsampling ratio (colsample_bytree) of 0.8, a minimum child weight (min_child_weight) of 3, and a full instance subsample rate (subsample) of 1. The model was trained using the binary:logistic objective, suitable for binary classification, and verbosity was suppressed (verbosity = 0) during training to streamline output.

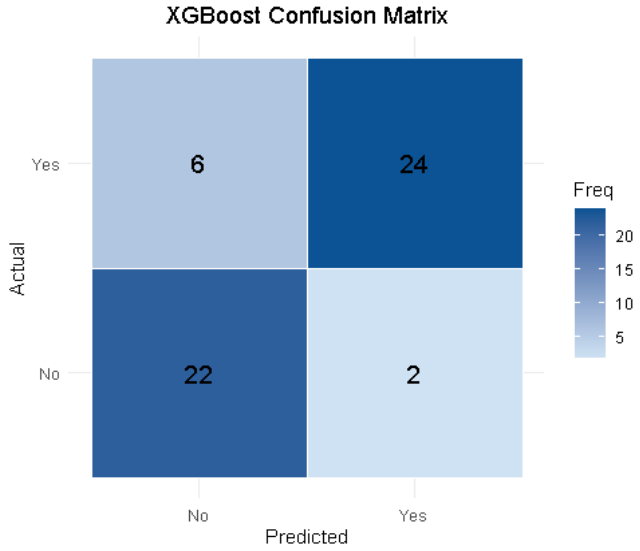

Fig. 5 Confusion matrix of XGBoost model

As shown in Figure 5, the XGBoost model achieved the highest test accuracy of 0.8519 among all models. According to the confusion matrix, 22 participants without heart disease were correctly classified, 2 were incorrectly predicted as having heart disease, 6 participants with heart disease were misclassified as not having the condition, and 24 were correctly identified as having heart disease.

4. Conclusion

This study compared the effectiveness of three widely used machine learning models-XGBoost, Random Forest, and Logistic Regression-for classifying heart disease cases based on patient data. The results indicated that XGBoost consistently achieved superior predictive performance, particularly in terms of accuracy and stability, suggesting its strong suitability for this type of classification task.

However, algorithm selection alone does not determine model success. The representation of features, the quality of input data, and especially the distribution of class labels was found to significantly affect model performance. These factors influence both learning efficiency and the model's ability to generalize to unseen data.

Future research could benefit from the use of more diverse and representative datasets that better reflect real-world clinical conditions. Additionally, exploring more complex model architectures-such as deep neural networks or hybrid ensemble frameworks-may yield further improvements in predictive accuracy and clinical applicability.

References

[1] Zhang Jingyi, Zhang Shizhong. Research progress in the epidemiology and related mechanisms of cardiovascular diseases and psychological diseases. Chinese Journal of General Medicine, 2023: 1-7.

[2] Gencer B, Evans D S, Aragam B, et al. Lipoprotein(a) and cardiovascular disease: From pathophysiology to clinical interventions. Nature Reviews Cardiology, 2021, 18(6): 339-354.

[3] Roth G A, Mensah G A, Johnson C O, et al. Global burden of cardiovascular diseases and risk factors, 1990–2019. Journal of the American College of Cardiology, 2020, 76(25): 2982-3021.

[4] Bhatnagar P, Wickramasinghe K, Williams J, et al. The epidemiology of cardiovascular disease in the UK 2014. Heart, 2016, 102(24): 1945-1952.

[5] Smith J. et al. Machine Learning for Cardiovascular Risk Prediction. Journal of Cardiology Research, 2021, 15(3): 45-58.

[6] Lee H, et al. Predictive Modeling in Cardiovascular Disease: A Comparative Analysis. International Journal of Medical Informatics, 2022, 129: 112345.

[7] Kim Y, et al. Deep Learning Approaches for Heart Disease Diagnosis. IEEE Transactions on Biomedical Engineering, 2023, 70(5): 1021-1033.

[8] Gupta R, et al. Challenges in Machine Learning-Based Clinical Risk Assessment. Computational Medicine, 2020, 18(2): 123-134.

[9] Wang T, et al. Interpretable AI Models in Cardiovascular Medicine. Nature Digital Health, 2024, 2(1): 55-69.

[10] Feng Ziqin. Analysis of Risk Factors and Development of a Clinical Prediction Model for Cardio neurosis. Beijing University of Chinese Medicine, 2023.