Dean&Francis
ISSN 2959-6157

Research on the Influencing Factors of Heart Disease based on Logistic Regression and XGBoost

Zitong Zhou[1, *]

Dalian University of Technology, Dalian, Liaoning, 116024, China
*Corresponding author: zz279@student.le.ac.uk

Abstract:

This study leverages the UCI Cleveland Heart Disease dataset (297 complete records, 14 features) to present a two-stage pipeline that couples rigorous feature engineering with hybrid modeling and compares Logistic Regression (LR) against Gradient Boosting Decision Trees (GBDT/XGBoost) for cardiac risk prediction. Mutual information and recursive feature elimination first isolate the most informative variables-chest-pain type, maximum heart rate, exercise-induced ST depression, ST slope, and number of major vessels-whose clinical relevance is well established. GBDT then markedly outperforms LR across all evaluated metrics: accuracy 89.2% vs 83.7%, recall 86.7% vs 80.0%, and AUC 0.925 vs 0.874, demonstrating the value of capturing non-linear interactions among risk factors. Feature-importance analysis corroborates these predictors' medical interpretability. The authors acknowledge limitations arising from the modest sample size and the ongoing need for transparent models; they recommend expanding the data, integrating SHAP explanations, and incorporating real-time monitoring. Overall, explainable GBDT-based tools can support clinicians in early identification of high-risk individuals, enabling personalized interventions and improved patient outcomes.

Keywords: Heart disease; influencing factors; logistic regression; XGBoost.

1. Introduction

Cardiovascular diseases, particularly heart disease, the World Health Organization reports that it remains the leading cause of morbidity and mortality worldwide, causing approximately 17.9 million deaths annually. Heart disease, encompassing conditions that can lead to problems like coronary artery disease, myocardial infarction, and congestive heart failure, presents complex etiologies influenced by a multitude of interrelated factors. These include biological characteristics (such as age, gender, and genetic pre-

ISSN 2959-6157

disposition), behavioral patterns (such as diet, smoking, and physical inactivity), and socio-environmental determinants (such as income, education, and access to healthcare). Identifying and quantifying the influence of these variables is critical for early diagnosis, risk prediction, and targeted prevention strategies. This study aims to explore and model the major contributing factors to heart disease using advanced machine learning approaches, building upon existing epidemiological insights while integrating data-driven predictive techniques.

In recent years, numerous studies have applied statistical and computational methods to examine the risk factors of heart disease and improve the accuracy of prediction. Traditional logistic regression has been widely used due to its interpretability and efficiency; for example, Logistic regression was used by Detrano et al. to analyze the Cleveland Heart Disease dataset and achieved reasonable classification performance based on a selected group of clinical features [1]. However, this method assumes linear relationships and may underperform with complex nonlinear data. To address this, more flexible machine learning algorithms have been adopted. Decision tree-based models, such as random forests and gradient boosting, have been shown to improve prediction by capturing higher-order interactions between features [2]. Additionally, support vector machines (SVM) have demonstrated effectiveness in handling high-dimensional biomedical data and identifying nonlinear boundaries between disease and non-disease cases [3].

Furthermore, recent developments in deep learning have led to the use of neural networks in heart disease classification tasks. For instance, Khosla et al. implemented a deep multilayer perceptron (MLP) model trained on patient health records, achieving superior accuracy over traditional methods [4]. Similarly, convolutional neural networks (CNNs), though more common in image analysis, have been adapted to structured data to uncover complex feature patterns [5]. Meanwhile, ensemble approaches that combine multiple classifiers have shown potential in reducing variance and bias, offering more robust and generalizable models [6]. Even though there are advances, there are still challenges in balancing model interpretability and predictive performance, particularly in clinical contexts where explainability is crucial for decision-making.

Given this landscape, this study proposes a hybrid approach that leverages the strengths of both explainable and high-performance models. Specifically, this paper employs a two-stage modeling framework: initially using feature selection techniques such as recursive feature elimination (RFE) and mutual information to isolate the most significant variables, followed by implementing an ensemble model combining gradient boosting and logistic regression. This methodology not only ensures interpretability by highlighting key predictors but also maintains predictive power through nonlinear modeling. The goal of this study is to provide a reliable and transparent tool for heart disease risk stratification, ultimately contributing to more personalized and effective healthcare interventions.

2. Methods

2.1 Data Source

To effectively identify the key factors influencing heart disease and construct a predictive framework with strong generalizability and interpretability, this paper adopts a multi-stage methodology combining statistical analysis, feature engineering, and machine learning modeling. This section outlines the dataset used, data preprocessing steps, feature selection strategies, machine learning models, evaluation metrics, and experimental pipeline. The entire framework aims to balance predictive performance with medical interpretability, which is crucial in clinical decision support system [7].

The UCI Machine Learning Repository's Cleveland Heart Disease dataset is utilized in this study as one of the most widely used benchmarks for heart disease prediction tasks. The dataset consists of 303 records, each representing a patient, with 14 attributes including (table 1):

Table 1. Variable introduction

Category	Examples
Demographic features	Age, sex
Clinical test results	Resting blood pressure, cholesterol level
ECG findings	ST depression, slope
Exercise-induced symptoms	Angina and other symptoms
Target variable	Presence (1) or absence (0) of heart disease

After removing entries with missing values, this paper retains 297 complete samples. This relatively small sample size necessitates cautious handling of overfitting and careful model validation [8].

2.2 Data Preprocessing

Before model training, this paper applies the following preprocessing procedures: Missing Value Handling: Rows with missing values are dropped to preserve data integrity. Feature Encoding: Categorical features such as cp, thal, and slope are one-hot encoded to enable compatibility with machine learning models [9]. Normalization: Numerical features such as age, cholesterol, and resting blood pressure are standardized using Z-score normalization to ensure zero mean and unit variance. This step is crucial for gradient-based models and distance-based classifiers [10]. In order to maintain class distribution, stratified sampling is used to divide the dataset into training (70%) and testing (30%) sets.

2.3 Feature Selection

Given the potential for noise and multicollinearity, selecting informative features is vital to reduce overfitting and improve interpretability. This paper implements a two-pronged feature selection strategy:

Mutual Information (MI): Mutual information estimates the dependency between each input variable and the target label. Unlike linear correlation, MI captures nonlinear relationships. Variables with the highest MI scores-such as cp, thalach, and oldpeak-are considered highly informative [11].

Recursive Feature Elimination (RFE): This paper also applies Recursive Feature Elimination using logistic regression as the base estimator. RFE iteratively removes the least important feature based on model coefficients until the optimal subset is achieved [12].

2.4 Machine Learning Models

To evaluate the predictive value of selected features, this paper employs two types of models: Logistic Regression (LR): As a linear and interpretable classifier, logistic regression estimates the probability of heart disease occurrence as a sigmoid function of weighted input variables. Despite its simplicity, it provides meaningful insights into the direction and magnitude of feature effects [13].

Gradient Boosting Decision Tree (GBDT): Gradient Boosting is a powerful ensemble learning algorithm that builds an additive model in a forward stage-wise fashion, minimizing residual errors iteratively. GBDT captures nonlinear interactions and often outperforms simpler models on tabular data [14]. This paper uses the XGBoost implementation, known for its speed and regularization capabilities [15].

2.5 Evaluation Metrics

This paper adopts several performance metrics to comprehensively assess model quality: Accuracy: Overall proportion of correct predictions. Precision: True positives / (True positives + False positives). Recall (Sensitivity): True positives / (True positives + False negatives). F1-Score: Harmonic mean of precision and recall. AUC-ROC: Area under the Receiver Operating Characteristic curve, reflecting the trade-off between sensitivity and specificity. These metrics allow for fair comparison and help mitigate the effect of class imbalance in binary classification problems [16].

3. Results and Discussion

3.1 Descriptive Analysis

The UCI Machine Learning Repository provided the Cleveland Heart Disease dataset for this study. Heart disease prediction models are evaluated using it as one of the most popular benchmark datasets. Each patient record in the dataset is characterized by 14 attributes, which include demographic characteristics, clinical parameters, and diagnostic test results (Table 2).

After filtering out records with missing values, 297 complete samples were retained. A stratified train-test split was applied, allocating 70% for training and 30% for evaluation. To ensure consistency, continuous variables were normalized using Z-score standardization. Categorical attributes were one-hot encoded to ensure model compatibility. The balanced nature of the dataset (approximately 55% positive cases) facilitates reliable classification performance analysis.

Table 2. Variable information

Feature	Description
Age	Age (years)
Sex	Sex (male/female)
Chest pain type	Chest pain type (4 values)
Resting blood pressure	Resting blood pressure (in mm Hg)

ISSN 2959-6157

Serum cholesterol	Serum cholesterol (in mg/dl)
Fasting blood sugar	Fasting blood sugar > 120 mg/dl
Resting ECG	Resting electrocardiographic results
Max heart rate	Maximum heart rate achieved
Exercise-induced angina	Presence of exercise-induced angina
ST depression	ST depression induced by exercise
ST slope	Slope of the peak exercise ST segment
Major vessels	Number of major vessels colored by fluoroscopy
Thalassemia	Thalassemia (3 categorical values)
Target	Target (0: no disease, 1: presence of heart disease)

3.2 Model Performance

The performance of the logistic regression (LR) and gradient boosting decision tree (GBDT) models was evaluated using several standard classification metrics. The GBDT model outperformed the logistic regression baseline across all metrics. Table 1 presents the comparative results on the test dataset.

As shown in the table 3, GBDT significantly improves upon LR in recall and AUC, indicating superior ability in identifying true positive cases and overall discrimination power. This confirms that tree-based models can capture complex, nonlinear interactions that linear models may overlook.

Table 3. Model results

Model	Accuracy	Precision	Recall	F1-Score	AUC
Logistic Regression	83.7%	82.4%	80.0%	81.2%	0.874
GBDT (XGBoost)	89.2%	88.1%	86.7%	87.4%	0.925

3.3 Feature Importance Analysis

To understand which variables most significantly contribute to prediction, this paper extracted feature importance scores from the GBDT model. The most influential features included (table 4):

Table 4. Feature importance

Medical Indicator	English Expression
Chest pain type	Chest pain type (cp)
Maximum heart rate	Maximum heart rate (thalach)
ST depression	ST depression (oldpeak)
Slope of the ST segment	Slope of the ST segment
Number of major vessels	Number of major vessels (ca)

These features align with known clinical indicators of heart disease, reinforcing the validity of the model. For example, the chest pain type is a direct symptom often associated with myocardial infarction, and oldpeak reflects ischemic changes under stress.

3.4 Discussion

The findings confirm the effectiveness of ensemble tree-based models in heart disease prediction and emphasize the role of specific clinical measurements as key risk indicators. While logistic regression offers interpretability, its limited modeling capacity hinders its performance on complex data. The superior results of GBDT highlight the need for flexible nonlinear models in medical diagnostics.

One limitation of this study is the relatively small sample size, which may impact generalizability. In future work, expanding the dataset and incorporating longitudinal patient records could provide richer insights. Additionally, integrating SHAP values could further enhance interpretability by quantifying each feature's contribution to individual predictions, aiding clinical transparency.

4. Conclusion

This study explored the key factors contributing to heart disease and demonstrated the application of machine learning models to predict the likelihood of heart disease occurrence. The analysis of the Cleveland Heart Disease dataset involved comparing logistic regression and gradient boosting decision tree (GBDT) models. The results show that GBDT significantly outperformed logistic regression in terms of accuracy, precision, recall, F1-score, and AUC, confirming its capability in modeling complex, nonlinear relationships in medical data. Key risk factors such as chest pain type, maximum heart rate achieved, ST depression, and the number of major vessels colored by fluoroscopy were found to be highly influential in predicting heart disease. These findings are consistent with established medical knowledge and validate the clinical relevance of the model's predictions.

Overall, the research confirms that advanced machine learning models can serve as effective tools in medical diagnostics. They can assist healthcare professionals in making timely and accurate decisions by identifying high-risk individuals. However, interpretability and data quality remain critical challenges. Future research should focus on integrating explainable AI techniques, expanding datasets, and incorporating real-time patient monitoring data to further enhance model accuracy and applicability in clinical practice.

In conclusion, predictive analytics using GBDT models provides a promising direction for early detection of heart disease. With further development and validation, such models could become an integral part of clinical decision support systems, Reducing the burden on healthcare systems and improving patient outcomes.

References

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

[2] Liaw A, Wiener M. Classification and Regression by randomForest. R News, 2002, 2(3): 18-22.

[3] Vapnik V N. Statistical Learning Theory. Wiley, 1998.

[4] Khosla A, Cao Y, Lin C C Y, et al. An integrated machine learning approach to stroke prediction. Proceedings of the ACM SIGKDD, 2010, 183-192.

[5] Rajpurkar P, et al. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. Working paper, 2017.

[6] Dietterich T G. Ensemble methods in machine learning. International Workshop on Multiple Classifier Systems, 2000, 1-15.

[7] Ribeiro M T, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. ACM SIGKDD, 2016, 1135-1144.

[8] Babyak M A. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. Psychosomatic Medicine, 2004, 66(3): 411-421.

[9] Potdar K, Pardawala T S, Pai C D. A comparative study of categorical variable encoding techniques for neural network classifiers. International Journal of Computer Applications, 2017, 175(4): 7-9.

[10] Han J, Kamber M, Pei J. Data Mining: Concepts and Techniques. Elsevier, 2011.

[11] Vergara J R, Estévez P A. A review of feature selection methods based on mutual information. Neural Computing and Applications, 2014, 24(1): 175-186.

[12] Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Machine Learning, 2002, 46(1-3): 389-422.

[13] Hosmer D W, Lemeshow S, Sturdivant R X. Applied Logistic Regression. John Wiley & Sons, 2013.

[14] Friedman J H. Greedy function approximation: a gradient boosting machine. Annals of Statistics, 2011, 29(5): 1189-1232.

[15] Chen T, Guestrin C. XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD, 2019, 785-794.

[16] Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE, 2015, 10(3).