Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 11
TPCEE 2025

Comparative Evaluation of Machine Learning Models for Cardiovascular Disease Prediction

Xiangjun Wu

School of Computer Science and Mathematics, Fujian University of Technology, Fuzhou, China, 350000

18396212826@163.com

Abstract. Cardiovascular diseases (CVDs) persist as a critical global health burden, accounting for over 30% of annual mortality worldwide. While machine learning approaches show promise in early-stage risk identification, existing research predominantly focuses on isolated model validation without systematic comparative analysis. To address this gap, our study conducts a rigorous multi-algorithm evaluation, leveraging Logistic Regression (LR), Random Forest (RF), and XGBoost classifiers trained on clinically validated CVD datasets. Methodologically, this study implement stratified 5-fold cross-validation with adaptive class-weight balancing to mitigate data imbalance issues, while embedding recursive feature elimination for optimal predictor selection. Performance benchmarking across three critical dimensions-precision, recall (sensitivity), and ROC-AUC-reveals RF's consistent superiority in both discriminative power and operational stability. Specifically, RF generates the most favorable sensitivity-specificity trade-off curve and attains significantly higher diagnostic sensitivity compared to LR and XGBoost, essential for minimizing false negatives in clinical screening scenarios. The validated model establishes a scalable risk stratification pipeline for community healthcare systems, enabling timely interventions while demonstrating methodological replicability for predictive epidemiology in resource-constrained primary care settings. This framework bridges a critical feasibility gap in translating algorithmic innovations into deployable clinical tools.

Keywords: Random Forest; Logistic Regression; XGBoost; Cardiovascular Diseases.

1. Introduction

With the prevalence of unhealthy lifestyles and the accelerating population aging, cardiovascular disease mortality remains high in China [1]. As a leading global cause of death, it accounts for approximately 17 million annual deaths, representing 30% of total global mortality [2-3]. Such diseases pose a significant global health challenge, underscoring the critical need for the development of accurate and more effective detection methods [4]. Therefore, the effective prediction of these diseases facilitates early diagnosis, thereby reducing mortality [5].

There are numerous research results on the prediction and analysis of cardiovascular diseases using different models. Zhu Shengyuan explored the application effect of the random forest algorithm in predicting cardiovascular diseases and identified key factors affecting cardiovascular diseases from multiple factors through the Gini coefficient [6]. Ye Suting and others used a decision tree algorithm to construct a heart disease dataset warning model and wrote a user program interface [7-8]. Jian Yang proposes a new framework for predicting heart disease using the smote-xgboost algorithm [9]. Ambrish G proposes Logistic Regression(LR) techniques to be applied to the UCI dataset to classify the cardiac disease [10].

However, a critical limitation in most previous studies on cardiovascular disease prediction lies in their over-reliance on single-model frameworks, which fail to evaluate the performance differences across distinct algorithmic paradigms systematically and thus restrict the robustness and generalizability of prediction results. To address this research gap, the core innovation of this study is threefold: first, it proactively constructs a multi-model comparative framework by establishing three representative predictive models—random forest, logistic regression, and XGBoost—that cover

both traditional statistical learning and advanced ensemble learning algorithms; second, it adopts sensitivity and AUC value as the dual evaluation benchmarks, avoiding the one-sidedness of single-index assessment; third, it quantifies the performance gaps among the three models through rigorous comparative analysis, thereby scientifically determining the optimal predictive model for cardiovascular diseases. This multi-dimensional, comparative research design not only fills the blank of insufficient cross-model validation in existing studies but also provides a more reliable algorithmic basis for clinical decision-making in cardiovascular disease risk assessment.

2. The basic principle of disease prediction models

2.1. The structure of random forest

RF is a typical algorithm in Bagging, commonly used to solve classification and regression problems [11]. It is an ensemble method based on decision trees, which improves the prediction accuracy and stability by combining multiple decision trees. For classification problems, each decision tree independently predicts and then votes to determine the final category. It can handle large-scale data, is robust to missing values and noise, and can evaluate feature importance.

Let the input space be $x \subseteq R^p$ and the output space be y=$\{0, 1\}$. Given the training set $D = \{(x_i, y_i)\}_{i=1}^n$, generate T corresponding decision tree models $\{h_t(x)\}_{t=1}^T$ to obtain different classification results. Among them, p is the feature dimension, $\{0, 1\}$, x_i is the binary classification label, is the feature vector of the i-th sample, and y_i is its corresponding label.

In classification problems, the combinatorial model of random forests can be expressed as follows:

$$H(x) = mode\{h_t(x)\} \tag{1}$$

Node splitting adopts the Gini impurity minimization criterion:

$$Gini(t) = 1 - \sum_{k=1}^{K} p_k^2 \tag{2}$$

Among them, p_k is the sample ratio of category k in node t. The reduction in impurity brought about by feature j during node splitting is defined as:

$$\Delta Gini_j = Gini(t) - \left(\frac{n_L}{n_t} Gini(t_L) + \frac{n_R}{n_t} Gini(t_R)\right) \tag{3}$$

When a node splits, select the feature j that maximizes $\Delta Gini_j$ and the corresponding optimal splitting threshold.

The model training adopts hierarchical 5-fold cross-validation. In each round of validation, the sample size ratio of the training set to the validation set is 4:1. At the same time, set class_weight = 'balanced' to automatically adjust the category weights, where n_+ and n_- are the number of positive and negative samples, respectively:

$$w_+ = \frac{n}{2 \times n_+}, \quad w_- = \frac{n}{2 \times n_-} \tag{4}$$

2.2. The structure of logistic regression

Logistic regression is a statistical analysis method, mainly applicable to solving binary classification problems. This model maps the linear output to the interval [0, 1] through the Sigmoid function, and the output result of this interval can be interpreted as the probability that the sample belongs to a certain category. Among them, the input of the Sigmoid function is represented by z, which is a linear combination of the model's independent variables, expressed as:

$$z = \beta_0 + \sum_{i=1}^{n} \beta_i x_i = \beta^T x \tag{5}$$

Among them, z is the linear output, β_0 is the intercept term, β_i is the weight, and x_i is the i-th element in the feature vector. Logistic regression maps real number z through the Sigmoid function, and its function expression is:

$$g(z) = \frac{1}{1+e^{-z}} \tag{6}$$

For the binary classification problem Y={0, 1}, it can be expressed in the form of a conditional distribution:

$$P(Y = 1|x, \beta) = \frac{1}{1+e^{-\beta^T x}} = h_\beta(x) \tag{7}$$

$$P(Y = 0|x, \beta) = \frac{1}{1+e^{-\beta^T x}} = 1 - h_\beta(x) \tag{8}$$

If P(Y=1|x, β)≥ 0.5, then the predicted value of this sample is considered to be 1; otherwise, it is 0. The same applies to P(Y=0|x, β).

Parameter estimation is solved by maximizing the likelihood function, which is:

$$L(\beta) = \prod_{i=1}^{n} \left(h_\beta\left(x^{(i)}\right)\right)^{y^{(i)}} \cdot \left[1 - h_\beta\left(x^{(i)}\right)\right]^{1-y^{(i)}} \tag{9}$$

Maximizing log-likelihood is equivalent to minimizing the cross-entropy loss function:

$$\mathcal{J}(\beta) = -\frac{1}{n}\sum_{i=1}^{n}\left[y^{(i)} ln\left(h_\beta(x^{(i)})\right) + \left(1 - y^{(i)}\right) ln\left(1 - h_\beta\left(x^{(i)}\right)\right)\right] \tag{10}$$

To minimize the objective function $\mathcal{J}(\beta)$, the gradient descent method is adopted to gradually iterate and optimize the parameters. The direction of parameter update is guided by calculating the gradient. The process is as follows:

$$\beta_{j+1} = \beta_j - \alpha \cdot \frac{\partial \mathcal{J}(\beta)}{\partial \beta_j} \tag{11}$$

The gradient calculation formula is:

$$\frac{\partial \mathcal{J}(\beta)}{\partial \beta_j} = \frac{1}{n}\sum_{i=1}^{n}\left(h_\beta\left(x^{(i)}\right) - y^{(i)}\right) x_j^{(i)} \tag{12}$$

To reduce the risk of over-fitting, the L2 regularization term is introduced:

$$\mathcal{J}_{reg}(\beta) = \mathcal{J}(\beta) + \lambda \parallel \beta \parallel_2^2 \tag{13}$$

Here, λ represents the regularization intensity, which is determined to be λ=0.1 through cross-validation. The assessment of feature importance is based on the absolute value of the coefficient $|\beta_j|$. If $|\beta_j|$ is close to 0, it indicates that feature x_j contributes little to the prediction. If $|\beta_j|$ is significantly greater than 0, it indicates that feature x_j is a key risk factor.

2.3. The structure of XGBoost

Using an embedded approach, this study leverages xgboost's built-in feature importance mechanism to automatically screen high-value feature subsets via SelectFromModel, reducing dimensionality while enhancing model performance and interpretability.Its core lies in the combination of the additive model and the regularization objective function:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t), \Omega(f_t) = \gamma T + \frac{1}{2}\lambda \parallel w \parallel^2 \tag{14}$$

In this formula, f_t is the prediction function of the t-th decision tree, $\Omega(f_t)$ is the regularization term (T is the number of leaf nodes, w is the leaf weight).

The gradient optimization mechanism approximates the objective function through Taylor's second-order expansion:

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^{n}\left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right] + \Omega(f_t) \tag{15}$$

Among them, $g_i = \partial_{\hat{y}^{(t-1)}} l\left(y_i, \hat{y}^{(t-1)}\right)$ and $h_i = \partial^2_{\hat{y}^{(t-1)}} l\left(y_i, \hat{y}^{(t-1)}\right)$ are the first-order and second-order gradients respectively.

In terms of embedded feature selection, this paper adopts feature screening based on importance scoring. In terms of quantifying feature importance, the average gain importance is adopted:

$$Importance_j = \frac{1}{T}\sum_{t=1}^{T}\sum_{split\ j} \Delta \mathcal{L} \tag{16}$$

Among them, $\Delta\mathcal{L}$ represents the reduction in loss caused by feature j at the splitting point, and then the feature contribution is analyzed through SHAP.

2.4. Model accuracy verification

To comprehensively evaluate the performance of the model, this study adopts three core indicators: accuracy, sensitivity, and the area under the ROC curve to quantitatively assess classification performance from multiple dimensions and comprehensively evaluate the model's efficacy.

$$Precision = \frac{TP}{TP+FP} \times 100\% \quad (17)$$

Among them, *Precision* represents the model's accuracy rate, *TP* represents samples that are actually diseased and whose model detection is also diseased, and *FP* represents samples that are actually not diseased but whose model detection is diseased.

$$TPR = \frac{TP}{TP+FN} \times 100\% \quad (18)$$

Among them, *TPR* represents the model's sensitivity, *TP* represents samples that are actually diseased and also diseased in model detection, and *FN* represents samples that are actually diseased but not diseased in model detection.

$$FPR = \frac{FP}{FP+TN} \times 100\% \quad (19)$$

FP represents samples that are actually not diseased but are detected to be diseased by the model, and *TN* represents samples that are actually not diseased and are also not diseased by the model.

3. Results

The data used in this article is sourced from https://www.saikr.com.

3.1. Analysis of experimental results

During the data processing stage, the dataset is divided into a training set and a test set: the training set is used for model construction and parameter tuning, while the test set evaluates model performance. Moreover, the data in the test set is not used during training to ensure the objectivity of the evaluation.

Therefore, we tested the test set from the heart disease dataset on the logistic regression model, the random forest model, and the XGBoost model, respectively, and presented the results using a confusion matrix:

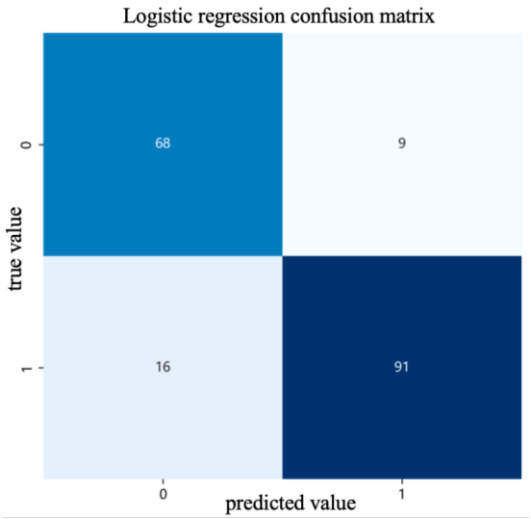

Figure 1. Logistic regression confusion matrix

As shown in Figure 1, when the logistic regression model was tested on the test set of the heart disease dataset, it successfully identified 91 cases among 107 real patients with a sensitivity of 85.05%. This indicates that the model can effectively capture most patients, but there is still a 15.95% missed

diagnosis rate. However, among 100 independent real disease sample tests, only 91 cases were correctly identified, and the accuracy rate was calculated to be 91%. This reflects that the model has high accuracy in positive prediction and fewer false positives.

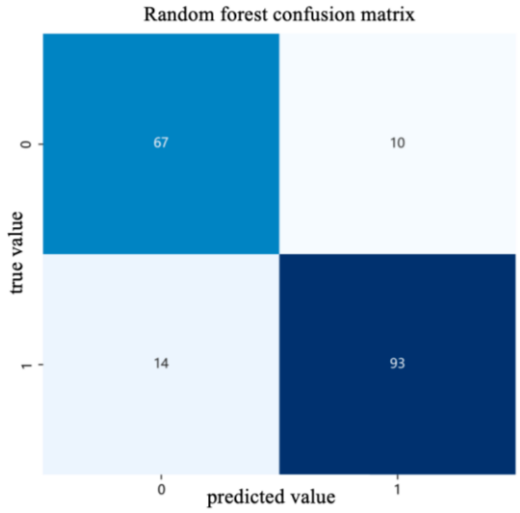

Figure 2. Random forest confusion matrix

From Figure 2, the test set in the heart disease dataset correctly identified 93 out of 107 actual patients using the random forest model, with a sensitivity of 86.92%, indicating the model has good detection ability for target cases. Among 103 real disease samples, the model identified 93, with an accuracy rate of 90.29%, indicating high stability of its positive predictive value. Compared with the logistic regression model, this model has slightly improved sensitivity and maintained similar accuracy, reflecting potential in reducing false negatives while keeping a lower false positive rate.

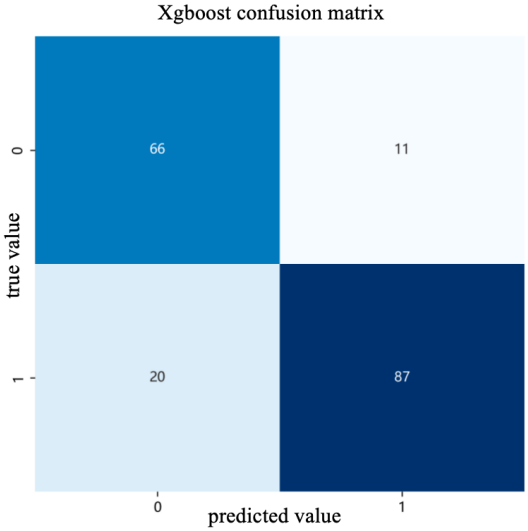

Figure 3. XGBoost confusion matrix

As shown in Figure 3, the XGBoost model correctly identified 87 out of 107 target patients, with a sensitivity of 81.31%. This result reflects that the model has basic case recognition ability, but there is a risk of missed diagnosis of approximately 18.69%. In the validation set containing 98 real disease samples, the model successfully predicted 87 positive cases with an accuracy rate of 88.78%. Through longitudinal comparison, it is shown that the model is weaker in sensitivity and accuracy than logistic regression and random forest models.

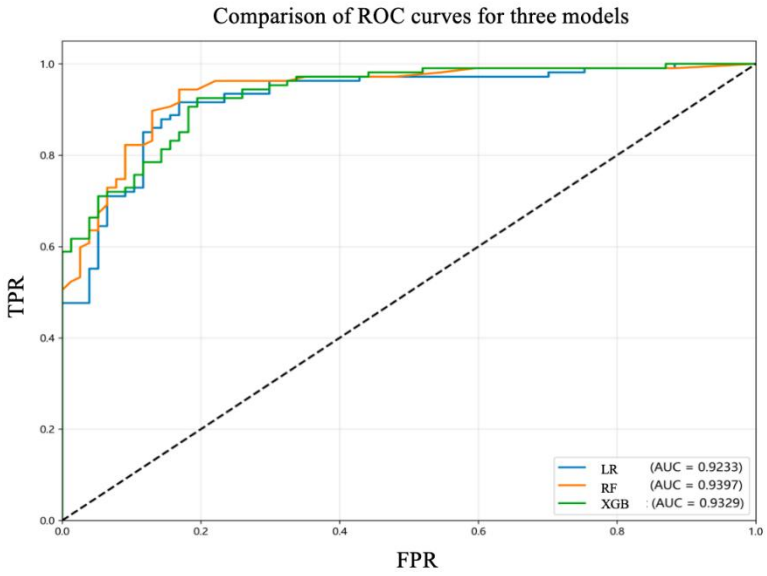

Figure 4. Comparison of ROC curves for three models

As shown in Figure 4, the random forest model exhibits the best performance, with the highest AUC value (0.9397), slightly superior to the logistic regression model (0.9233) and the XGBoost model (0.9329), and the corresponding ROC curve is closer to the upper left corner of the coordinate. In terms of classification metrics, the sensitivity of the random forest model reaches 86.92%, higher than the other two models.

In summary, by comparing the performance of three models—logistic regression (sensitivity 85.05%, accuracy 91%), random forest (sensitivity 86.92%, accuracy 90.29%), and XGBoost (sensitivity 81.31%, accuracy 88.78%)—in predicting cardiovascular diseases, and incorporating ROC curve analysis, it is evident that the random forest model not only significantly outperforms logistic regression (0.9233) and XGBoost (0.9329) with the highest AUC value (0.9397), but also has a ROC curve closer to the upper left corner of the coordinate, indicating the best overall classification performance. Additionally, with a sensitivity of 86.92%, the random forest model achieves the highest sensitivity among the three models. Although its accuracy (90.29%) is slightly lower than that of the logistic regression model (91%), considering the relatively balanced class distribution in this study's dataset, and taking into account core indicators such as sensitivity, AUC, and ROC curve position, the random forest model achieves the best balance between reducing the rate of missed diagnoses and maintaining a low rate of false positives. Therefore, it is determined as the optimal model for predicting cardiovascular diseases.

4. Conclusions and outlooks

This study addresses the need for cardiovascular disease prediction by constructing a multi-classification prediction model based on logistic regression, random forest, and XGBoost algorithms. Through systematic comparative analysis of multiple models, it effectively addresses the limitations of single-model evaluation in existing research. The study innovatively introduces a five-fold stratified cross-validation strategy and class weight balancing technique (with the parameter set to class_weight='balanced'), which effectively mitigates the class imbalance issue commonly present in medical datasets and enhances the model's ability to identify minority class samples. Experimental results show that the random forest model exhibits optimal predictive performance in cardiovascular disease prediction, fully demonstrating the significant advantages of ensemble learning algorithms in complex medical feature mining and disease risk prediction. At the practical application level, this random forest model can be further deployed in community hospital cardiovascular disease screening systems to achieve early risk warning for diseases. According to calculations, the model can reduce the missed diagnosis rate of cardiovascular diseases by approximately 13.08%, providing a data-

driven technical solution for grassroots medical institutions to optimize medical resource allocation and improve disease screening efficiency.

This study has three limitations: the sample does not fully cover rare cardiovascular subtypes and regional characteristics, limiting its generalizability; the "black box" nature of random forest restricts clinical attribution; XGBoost inference delays 28ms per sample, making it difficult to meet real-time requirements. In the future, a federated learning architecture will be established to integrate multi-center data, incorporate the SHAP framework to enhance interpretability, use model distillation to reduce XGBoost latency, and explore the construction of a dynamic prediction system using GNN.

References

[1] Miao Lipeng, Ren Kehao, LI Mengdie, et al. Trend Analysis and Prediction of Cardiovascular Disease Mortality in China from 2009 to 2021 [J]. Chinese General Practice, 2024, 27 (18): 2260-2264.

[2] Mehra R. Global public health problem of sudden cardiac death [J]. Journal of electrocardiology, 2007, 40 (6): S118-S122.

[3] Yaseliani M, Khedmati M. Prediction of heart diseases using logistic regression and likelihood ratios [J]. International Journal of Industrial Engineering & Production Research, 2023, 34 (1): 1-15.

[4] Ogunpola A, Saeed F, Basurra S, et al. Machine learning-based predictive models for detection of cardiovascular diseases [J]. Diagnostics, 2024, 14 (2): 144.

[5] Azmi J, Arif M, Nafis M T, et al. A systematic review on machine learning approaches for cardiovascular disease prediction using medical big data [J]. Medical engineering & physics, 2022, 105: 103825.

[6] Shi Shengyuan, Zhu Lei, Ye Lin, etc Research on Cardiovascular Disease Prediction Based on Random Forest Algorithm [J]. Intelligent Computer and Application, 2021, 11 (04): 176-178.

[7] Ye Suting, Pan Yuanyuan, Bi Yingchun Research on Heart Disease Early Warning Model Based on Decision Tree Algorithm [J]. Computer Knowledge and Technology, 2020, 16 (19): 187-189.

[8] Feng Yuan, Yu Rongrong, Sun Ziwei Research on the Classification and Prediction Model of Cardiac Cases [J]. Advances in Applied Mathematics, 2024, 13:4610.

[9] Yang J, Guan J. A heart disease prediction model based on feature optimization and smote-Xgboost algorithm [J]. Information, 2022, 13 (10): 475.

[10] Ambrish G, Ganesh B, Ganesh A, et al. Logistic regression technique for prediction of cardiovascular disease [J]. Global Transitions Proceedings, 2022, 3 (1): 127-130.

[11] Magidi J, Nhamo L, Mpandeli S, et al. Application of the random forest classifier to map irrigated areas using google earth engine [J]. Remote Sensing, 2021, 13 (5): 876.