Proceedings of the 5th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/115/2025.18477

Prediction of Heart Disease Based on Data Classification Models

Yuanhang Yang[1,a,*]

[1]*School of Communication Engineering, Nanyang Technological University, Singapore*
a. YANG0937@e.ntu.edu.sg
**corresponding author*

Abstract: With the increasing emphasis on healthy living, the prevention of major diseases has garnered widespread attention. Heart disease, as one of the leading causes of death globally, makes early prediction and identification extremely important. Currently, the prediction of heart disease mainly relies on a multidisciplinary approach. This article focuses on machine learning and selects four supervised learning classification models, including the Back Propagation (BP) neural network classification model, Genetic Algorithm (GA) -Back Propagation neural network classification model, Support Vector Machine (SVM) classification model and Random Forests Algorithm (RF) classification model, to evaluate their performance. Simulation experiments indicate that the GA-BP neural network classification model has the best generalization ability, while the RF classification model achieves the highest classification accuracy and recall rate, with an accuracy of 82.6% and a recall rate of 88.1% on the test set. Overall, the RF classification model performs the best among the four classification models. In future research, improvements and integrations of existing multiple data classification models will play a crucial role in enhancing classification accuracy.

Keywords: Prediction of Heart Disease, Machine Learning, Classification Models.

1. Introduction

As the quality of life continues to improve, people are increasingly focusing on health and the prevention of major diseases [1]. Heart disease, being a condition with high morbidity and mortality rates, makes early prediction and identification very important. Heart disease generally refers to diseases that affect the structure or function of the heart, with common types including coronary heart disease, heart failure and arrhythmias. When heart disease develops, it can impair the heart's ability to pump blood properly, which in turn disrupts the normal functioning of different organs in the body [2]. In severe instances, heart disease can lead to death. Consequently, predicting heart disease allows for early identification of high-risk individuals, enabling doctors to implement timely preventive measures and decrease the incidence of the condition. Additionally, it facilitates the efficient allocation of medical resources, ensuring that high-risk patients receive prompt monitoring and treatment.

Currently, common methods for predicting heart disease rely on the integration of multiple disciplines, including biomarker research, big data and machine learning. As a rapidly developing technology in recent years, machine learning provides better opportunities for heart disease prediction.

DOI: 10.54254/2755-2721/115/2025.18477

Common classification methods include BP neural networks, RF, K-Nearest Neighbor (KNN), SVM, logistic regression and decision trees. Tan Pengliu extracted data using Convolutional Neural Networks (CNN) and proposed a heart disease prediction model combined with Adaboost, achieving an accuracy of 91.7% [3]. Mohan established a new heart disease prediction model using a linear mixed RF algorithm, reaching an accuracy of 88.7% [4]. Zhao Jinchao preprocessed data using KNN and then combined it with the RF algorithm, resulting in a heart disease prediction model with an accuracy of 83.2% [5].

In the aforementioned studies, researchers proposed corresponding improvements to different classification models to enhance the classification accuracy of the original models. This paper, however, uses four fundamental classification models to compare their classification accuracy and recall rates, assessing the predictive performance of different models to identify the most suitable prediction model. Based on that model, further optimizations will be performed to improve classification accuracy.

2. Research method

2.1. Data set

The dataset utilized in this study was collected by cardiologists in Cleveland, Switzerland, Hungary and Long Beach, which began in 1988. The dataset comprised a total of 1,026 samples, including 526 patients with heart disease and 500 patients without it. Each sample contains 14 features, namely age, sex, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal, cp and target. The specific meanings of each feature of the sample are shown in Table 1.

Table 1: Characteristics of heart disease prediction samples

Features	Meaning	Range
age	age	29-77 years old
sex	sex	0 means female and 1 means male
trestbps	resting blood pressure	94-200Hg
chol	cholesterol	126-564mg/dL
fbs	fasting blood glucose	0 means blood sugar is less than 120mg/dL and 1 means blood sugar is greater than 120mg/dL
restecg	resting electrocardiogram	0 indicates normal, 1 indicates ST-T abnormality and 2 indicates left ventricular hypertrophy
thalach	maximum heart rate	71-202b/min
exang	angina caused by exercise	0 means no and 1 means yes
oldpeak	exercise-induced st inhibition	0-6.2mV
slope	peak motion st segment	1 means up, 2 means flat and 3 means down
ca	number of major vessels	0-4
thal	thalassemia	0 means missing, 1 means normal, 2 means fixed defect and 3 means reversible defect
cp	chest pain category	1 is typical stranglehold disease, 2 is SARS-type stranglehold disease, 3 is non-angina pain and 4 is asymptomatic
target	status of heart disease	0 means no heart disease and 1 means having heart disease

DOI: 10.54254/2755-2721/115/2025.18477

2.2. Data model

This study uses four typical supervised learning models, including the BP neural network classification model, GA-BP neural network classification model, SVM classification model and RF classification model.

The BP neural network classification model is a multi-layer feedforward network structure that utilizes error backpropagation. It primarily optimizes network weights using the backpropagation algorithm, enabling the network to learn and predict input data effectively [6]. This classification model can establish complex nonlinear relationships and is suitable for various types of tasks. However, it may encounter challenges like getting stuck in local optima and being susceptible to overfitting.

The GA-BP neural network classification model not only retains the strong learning capability of the original BP neural network, but also incorporates the global optimization ability of genetic algorithms. The global search capability of GA allows the model to find better parameter configurations, making it adaptable to complex search spaces. But it will face challenges such as requiring larger datasets [7].

The core idea of the SVM classification model is to find a hyperplane that maximizes the margin between different classes, thereby achieving effective classification of the data. SVM can address nonlinear classification issues by employing kernel functions to transform the data into high-dimensional spaces, effectively balancing the training and test sets and minimizing the risk of overfitting [8].

The RF classification model enhances both the accuracy and robustness by building multiple decision trees and combining their outputs. This model demonstrates excellent predictive performance; however, it is inherently a black-box model, which results in lower interpretability and its performance on high-dimensional data is not ideal [9].

2.3. Experimental Method

This experiment utilized Matlab simulation software to obtain the classification accuracy of different models for heart disease prediction. When the number of training iterations is too low, the loss function of the classification model fails to converge. Conversely, when the number of training iterations is excessive, the classification results may overfit. Therefore, during the experiment, different training iterations were set and the quality of the model and the training process were assessed based on the accuracy and recall rates of the testing set, as well as the proximity of the accuracy between the training and testing sets.

The specific experimental design begins with preparing the heart disease dataset, followed by adjusting the ratio of the training set to the testing set. These sets are then input into the four classification models to obtain the classification accuracy for both the training and testing sets. A consistent and appropriate ratio is chosen to ensure that the classification accuracy of all four models is high and that they exhibit good generalization capabilities. Subsequently, the number of training iterations is adjusted to obtain varying classification accuracies and the average value of the classification accuracy that stabilizes after training is taken as the final classification accuracy and recall rate for the four models. Finally, the performance of the four models is evaluated based on their classification accuracy, recall rate and generalization ability.

3. Experimental Results

This experiment aims to evaluate the effectiveness of four typical classification models in predicting heart disease and to provide assistance for relevant medical practices. In this experiment, the ratio of

DOI: 10.54254/2755-2721/115/2025.18477

the training set to the testing set is first adjusted to obtain the line graphs of classification accuracy for the four models, as shown in Figure 1.

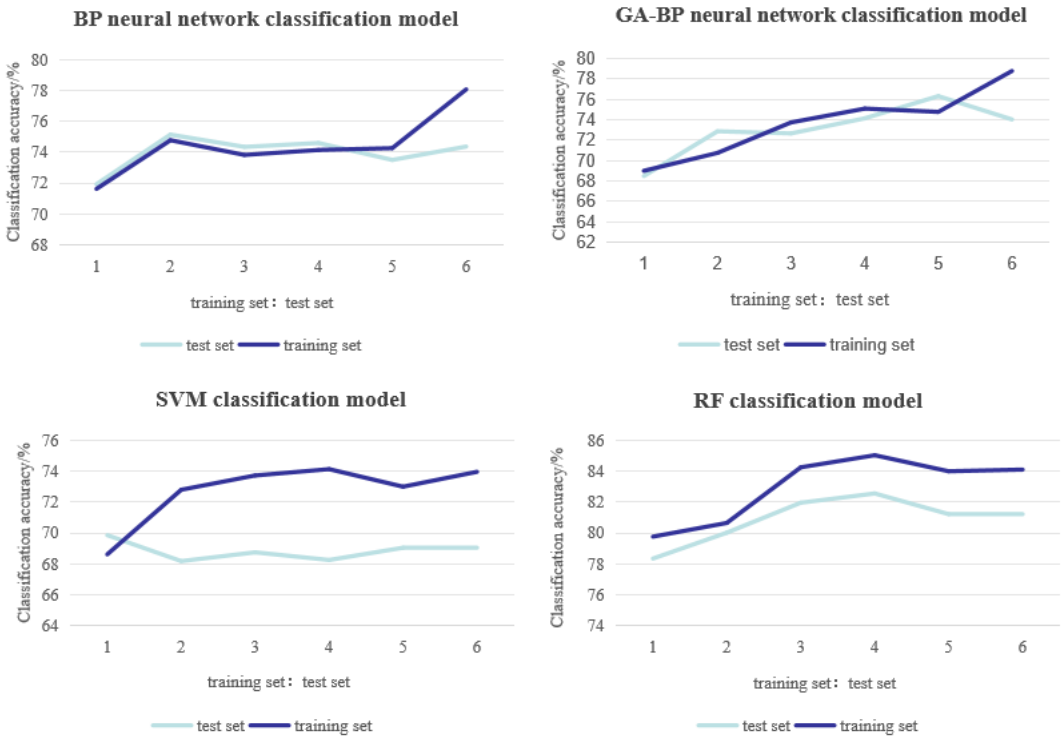

Figure 1: Classification accuracy depending on ratio of training set to test set

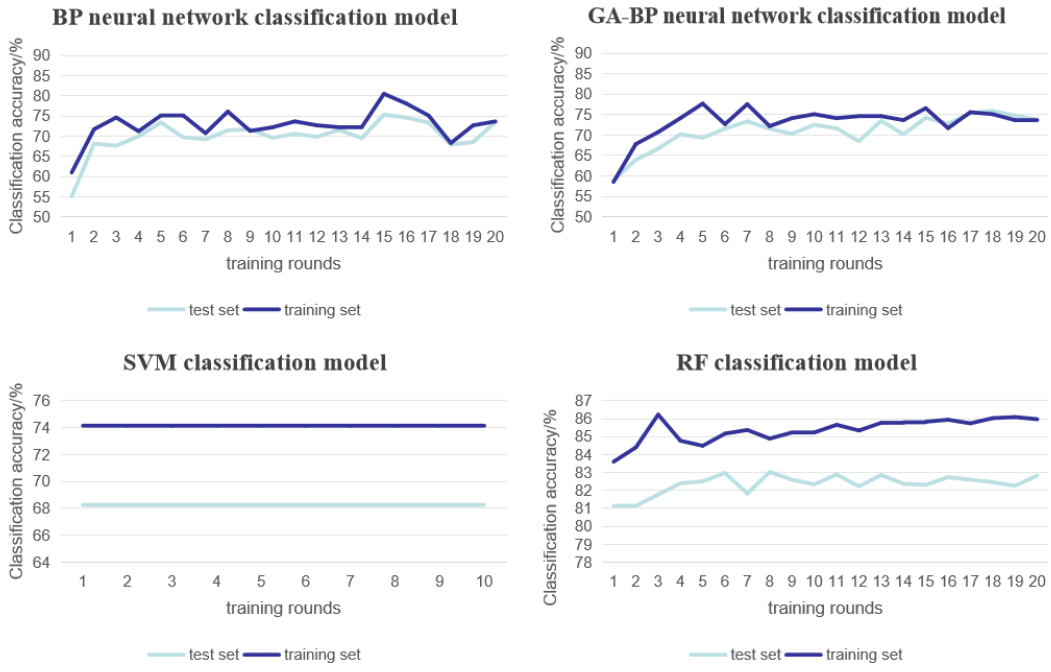

Figure 2: Classification accuracy depending on training rounds

DOI: 10.54254/2755-2721/115/2025.18477

According to the experimental data presented in Figure 1, it is evident that when the ratio of the training set to the testing set is 4:1, the classification accuracy of all four models remains relatively high and demonstrates strong generalization ability. Therefore, this ratio is selected for the next step of the experiment. The training iterations of the models are adjusted to obtain the classification accuracies of the four models, as shown in Figure 2.

From the experimental results shown in Figure 2, it can be seen that, with the exception of the SVM classification model, the classification accuracy of the other three models varies as the number of training iterations increases. Apart from the inherent randomness in the predictive capability of the models, the increase in training iterations has a positive effect on improving the classification accuracy. Upon comparison, it was found that the fluctuation of the four models becomes smaller after reaching 11 training iterations. As a result, the classification accuracy from 11 to 20 training iterations are averaged to determine the final classification accuracy for the four models. The findings are presented in Table 2.

Table 2: Accuracy and recall rates of four models

Models	BP neural network classification model	GA-BP neural network classification model	SVM classification model	RF classification model
Training set accuracy	73.91%	74.34%	74.15%	85.82%
Test set accuracy	71.47%	73.07%	68.25%	82.55%
Recall rate	74.00%	72.20%	87.35%	88.14%

Accuracy is defined as the proportion of correctly classified samples to the total number of samples, whereas recall is the proportion of true positive cases to the total number of actual positive cases. In this experiment, accuracy reflects the ratio of correctly predicted individuals to the overall number of individuals, while recall represents the ratio of accurately predicted heart disease patients to the total number of heart disease patients. The recall is important because misclassifying individuals without heart disease as having the disease can be rectified through subsequent screening, but misclassifying heart disease patients as healthy can lead to delays in treatment, resulting in serious consequences. Therefore, in heart disease prediction, the importance of recall is greater than that of accuracy. Generalization ability typically describes the degree of fitting. The smaller the difference in accuracy between the test set and the training set is, the stronger the generalization ability is.

According to the experimental data in Table 2, although the BP neural network classification model and the GA-BP neural network classification model exhibit good generalization ability, their accuracy and recall rates are relatively low. The SVM classification model has a high recall rate, but its generalization ability is poor and its accuracy is also low. The RF classification model outperforms the other three models by 10% in both accuracy and recall rate and its generalization ability is not bad. Considering the characteristics of the four classification models, the RF classification model is the most effective for predicting heart disease among the four models.

4. Conclusion

Heart disease is a significant threat to health and early prediction is crucial. This paper employs various machine learning classification algorithms, including the BP neural network classification model, GA-BP neural network classification model, SVM classification model and RF classification model, to study heart disease prediction.

DOI: 10.54254/2755-2721/115/2025.18477

In this experiment, the ratio of the training set to the testing set was first adjusted to obtain the optimal performance values for the four classification models. Subsequently, the training iterations of the models were modified to derive the classification accuracy, recall rate and generalization ability for each model. The results indicate that the GA-BP neural network classification model has the best generalization ability, while the RF classification model achieves the highest classification accuracy and recall rate, reaching 82.55% and 88.14%, respectively. Overall, the RF classification model demonstrates the best performance among the four models in predicting heart disease.

However, this experiment still has areas that need improvement and expansion, primarily focusing on the following two points. First, increasing the number of samples and features is necessary. The dataset used in this experiment mainly consists of heart disease data from four regions. This limitation may lead to poor generalization ability due to regional differences. However, due to confidentiality concerns, obtaining data from heart disease patients presents certain difficulties. Second, adding more classification models, such as KNN and logistic regression algorithms, would be beneficial. By comparing a greater variety of classification models, it may be possible to identify the model most suitable for heart disease prediction and optimize it to achieve higher prediction accuracy.

References

[1] *Nouman, A., Muneer, S. (2022). A systematic literature review on heart disease prediction using blockchain and machine learning techniques. Int J Comput Innovative Sci, 1(4): 1-6.*

[2] *Jiang, M., Zhang, H. (2024). Research on the effectiveness of machine learning algorithms for heart disease prediction. Chinese Journal of Medical Physics,41(7): 905-909.*

[3] *Tan, P., Xu, G., Zhang, L., et al. (2023). Heart disease prediction model based on convolutional neural networks and adaboost. Computer Applications, 43(z1): 19-25.*

[4] *Mohan, S., Thirumalai, C., Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. IEEE Access, 7: 81542-81554.*

[5] *Zhao, J., Li, Y., Wang, D., et al. (2021). Optimized random forest heart disease prediction algorithm. Journal of Qingdao University of Science and Technology (Natural Science Edition), 42(2): 112-118.*

[6] *Oscar, F., Deniz, R., Jose, E. (2003). Accelerating the convergence speed of neural networks learning methods using least squares. European Symposium on Artificial Neural Networks.*

[7] *Wang, W., Zhu, Q., Wang, Z., et al. (2021). Research on indoor positioning algorithm based on SAGA-BP neural network. IEEE Sensors Journal, 22(4): 3736-3744.*

[8] *Birzhandi, P., Kim, K., Youn, H. (2022). Reduction of training data for support vector machine: a survey. Soft Comput, 26(8): 3729-3742.*

[9] *Dong, Y., Zhang, S., Xu, J., et al. (2022). Random forest algorithm based on linear privacy budget allocation. Database Manage, 33(2): 1-19.*