


# Predictive Assessment of Heart Disease Based on Multiple Machine Learning Models

Zhongyi Zhang <sup>a</sup>

Software Engineering, Tianjin University of Commerce, Tianjin, China

**Keywords:** Heart Disease Prediction, Machine Learning, Logistic Regression, Support Vector Machines.

**Abstract:** Cardiovascular disease has been one of the leading causes of many deaths, and early diagnosis can improve treatment outcomes and survival rates. In this paper, six mainstream algorithmic models applicable to dichotomization are compared to predict whether a person is suffering from heart disease based on a number of features, such as gender, age, type of chest pain, resting Electrocardiograph (ECG) results, and maximum heart rate. The study first explored the relationship between the values of these features and tried to analyze the main factors affecting heart disease among them, then the dataset was divided in a way that the test set was 30% and the training set was 70% to train the model, and finally the six algorithmic models were used to predict the dataset, and the training results showed that Support Vector Machine (SVM) algorithmic model could provide more accurate data for the prediction of heart disease. This paper provides new tools and ideas for clinical diagnosis, treatment and prevention in the field of cardiovascular medicine, and contributes to the improvement of patients' quality of life and the reduction of medical costs.


## 1 INTRODUCTION

Heart disease is a common cardiovascular disease, which is a disease that directly involves the structure or function of the heart, including coronary artery disease, myocardial infarction, heart failure, arrhythmia, etc. Stroke, pulmonary embolism, and severe hypertension are also relatively common complications, which have a great impact on the patient's quality of life, and pose a great threat to human health and life. Therefore, the ability to more accurately predict the likelihood of disease can help people develop preventive strategies to improve quality of life and longevity.

In recent years, with the progress of medicine and people's attention to health, an increasing number of people will take the means of regular testing to check whether they are suffering from heart disease and whether they may suffer from the disease in the future, and according to the results of the treatment or prevention. But if relying only on the doctor's own experience to judge, the efficiency is low, the labor cost is high, and there is a certain degree of misdiagnosis. Therefore, there is a need to rely on the

means of Artificial Intelligence (AI) to carry out the auxiliary diagnosis.

Currently, the field of artificial intelligence is developing rapidly, and has been applied in many fields including transportation, education, civil engineering, finance, biology, healthcare, etc (Qiu, 2022; Qiu, 2024; Sun, 2020; Wu, 2024; Zhou, 2023). Especially in the medical field, there are many remarkable achievements and breakthroughs in recent years. For example, Zhang, Wu, et al. focus on medical AI, and put forward the first chest X-ray diagnostic basic model for disease diagnosis based on the enhancement of the knowledge in the medical field (Zhang, 2023), Yuzhe Yang et al. develop a respiratory signal detection model for Parkinson's diagnosis through sleep breathing detection, and one important direction is to complete the diagnosis and prediction of heart disease (Yang, 2022). In addition, Ouyang et al. used the EchoNet-Dynamic model to complete the detection of echocardiograms (Ouyang, 2020). ERIC J. TOPOL et al. developed a new AI model to complete the detection of atrial fibrillation (AFib) etc (Yang, 2022). In conclusion, the rapid development of AI is bringing great changes and

<sup>a</sup> <https://orcid.org/0009-0004-3324-3938>

innovations in various fields, providing people with smarter, efficient and sustainable solutions.

Currently, numerous models exist for illness prediction; however, for diseases like heart disease, various model structures and parameter selections significantly influence the prediction outcomes. In addition, the emergence of cardiac disease is an intricate process that is typically impacted by multiple causes. For example, the environment, individuals, and lifestyle habits. For the major disease of heart disease, how to identify self-consistent prediction models and techniques, enhancing the precision of diagnosis is a matter that requires immediate attention.

In order to solve the above problems, this article will predict heart disease based on the dataset from Kaggle using various machine learning models such as Logistic Regression, Random Forest, Naive Bayes, Support Vector Machines (SVM) etc., then select the most suitable prediction model and method for heart disease by comparing the learning results.

## 2 METHOD

### 2.1 Dataset Preparation

The source of the dataset used in this study is the Kaggle platform (Kaggle, 2019). There are a total of 1025 records in this dataset, each record contains 14 attribute features, such as chest pain type, thalassemia, resting electrocardiographic results etc. In the chest pain type, '0' represents typical angina, '1' represents atypical angina, '2' represents non-anginal pain and '3' represents asymptomatic.

#### 2.1.1 Preprocessing

In order to improve the performance and accuracy of the machine learning algorithms, increase the generalization ability of the model, and reduce the overfitting and underfitting problems of the model, it is necessary to preprocess the dataset in the file. Firstly, there will modify the attribute names of the data by expanding the original abbreviated form of the names to full names to enhance our understanding of the attribute features. Secondly, it is necessary to check the tables for missing values and display them in a heat map. Then, convert variable numbers to text names while defining the feature data and target data, and divide the data proportionally into training and testing sets. Finally, standardising the data so that the models have the same scale when processing the data, which enables the algorithms to learn and generalize

more efficiently, and it also helps this projection to choose the most appropriate learning model.

## 2.2 Machine Learning Models

In this machine learning, this paper needs to use six algorithmic models Logistic Regression, Decision Tree, Random Forest, Naive Bayes, K-Nearest Neighbors (KNN) and Support Vector Machines (SVM). In order to train, evaluate and select the models, this study use various tools provided by sklearn. For instance, to divide the dataset, one can use the `train_test_split` function in the `model_selection` module. To calculate the accuracy of the six models, the `accuracy_score` function in the `sklearn.metrics` module can be used.

### 2.2.1 Logistic Regression

The main idea of the logistic regression model is to make binary predictions by building a linear model and the linear output is mapped to a probability range of  $[0, 1]$  using a logistic function. The sigmoid function (or logistic function):

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad \sigma: R \rightarrow (0,1) \quad (1)$$

The threshold is typically set to 0.5. Predictions with a probability greater than 0.5 are classified as positive, while those with a probability less than or equal to 0.5 are classified as negative.

During the training phase, the logistic regression model estimates the model parameters using methods such as maximum likelihood estimation or gradient descent. This is done so that the model's predicted probabilities for the training data are as close as possible to the actual labels. During the prediction phase, the logistic regression model utilises the learned parameters to calculate the probabilities of the input samples and make classification predictions based on a predetermined threshold. The performance of the model can be evaluated by various evaluation metrics (e.g., accuracy, precision, recall, F1 value, etc.).

### 2.2.2 Decision Tree

The main idea of decision tree modeling is to divide and predict data through a series of feature selection and node splitting. It utilizes a hierarchical tree model to depict the decision-making pathway, where each branch signifies a condition based on a feature's value or a specific threshold. Every non-terminal node symbolizes a feature or attribute, while terminal nodes, or leaves, denote a particular category or

outcome. And, according to the selected features and division criteria, it can divide the dataset into different subsets, for discrete features, each subset corresponds to one value of the feature; for continuous features, the data can be divided into two subsets based on a threshold value.

**2.2.3 Random Forest**

The main idea of random forest model is to perform classification or regression by combining multiple decision trees. This is an integrated learning method that combines the benefits of decision trees. It enhances the performance and generalisation of the model by randomly selecting features and samples for training.

Random forest will combine multiple decision trees together, each time the dataset is randomly have put back to select, at the same time randomly selected part of the features as input. In the context of classification, the combiner determines the final outcome by choosing the option that has the majority vote among various classification results. For regression problems, it computes the final result by taking the average of outcomes from multiple regressions.

**2.2.4 Naive Bayes**

The main idea of plain Bayesian model is based on Bayes' theorem and the assumption of conditional independence of features. It is a simple but effective probabilistic classification algorithm, which is called "plain" because it is simplified accordingly on the basis of Bayesian algorithm, which is the most original and simplest assumption of Bayesian classification, i.e., all the features are relatively independent of each other.

Equation (2) represents a common Bayesian formula, where P(A) represents the a priori probability, i.e. the probability of event A occurring before the occurrence of event B. P(A|B) represents the a posteriori probability, i.e. the probability of event A occurring after the occurrence of event B. The likelihood function, P(B|A)/P(B), is an adjusting factor that makes the predicted probability closer to the true probability:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{2}$$

As plain Bayes is founded on the independence of features, the above equation can be expressed as follows, given the category a:

$$P(B|A = a) = \prod_{i=1}^d P(B_i|A = a) \tag{3}$$

Finally, Equation (4) is derived.:

$$P_{post} = P(A|b) = \frac{P(A)\prod_{i=1}^d P(b_i|A)}{P(B)} \tag{4}$$

**2.5.5 K-Nearest Neighbors (KNN)**

The KNN model is an instance-based learning algorithm model that is mainly used for classification and regression problems. Its main idea is to make classification or regression predictions based on the nearest neighbor samples in the feature space, which assumes that similar samples have similar classes or objective values.

The KNN algorithm implementation process is also relatively simple. To predict an input vector *x*, it is necessary to identify the set of *k* nearest vectors to *x* in the training data set. The category of *x* can then be predicted as the one with the highest number of categories among these *k* samples. Again, this is where the *k* in the KNN algorithm comes from.

Equations 5, 6 then show the mathematical expression of the KNN algorithm. In this mathematical expression assume that the training dataset *D* now has *m* samples, *x* is the feature vector of the samples and each sample has *n* features, and *y* is the category corresponding to the sample.

$$D = \{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\} \tag{5}$$

$$x_i = (x_i^{(1)}, x_i^{(2)}, \dots, x_i^{(n)}) \tag{6}$$

The distance of the nearest neighbor samples can be measured by the following formula. In the formula:

$$L_p(x_i, y_j) = \left\{ \sum_{l=1}^n |x_i^{(l)} - x_j^{(l)}|^p \right\}^{\frac{1}{p}} \quad (p \geq 1) \tag{7}$$

**2.2.6 Support Vector Machines (SVM)**

The main concept behind SVM is to identify the most suitable hyperplane in the feature space that maximizes the interval between samples of different classes. This idea of interval maximization can enhance the model's generalization ability and mitigate the risk of overfitting.

In this algorithm, the separating hyperplane is usually denoted by  $\omega \times x + b = 0$ , for linearly divisible datasets, there are infinitely many hyperplanes that fit this description, but the one with the largest geometric interval is unique.

For nonlinear classification problems in the input space, the kernel function can be used to handle and perform linear discrimination in higher dimensional space. This is shown in the following expression (8):

$$f(x) = \text{sgn} \left( \sum_{i=1}^N \alpha_i^* y_i K(x, x_i) + b^* \right) \tag{8}$$

### 3 RESULTS AND DISCUSSION

This article is based on machine learning, in order to have a better prediction of heart disease, a total of six different algorithms are compared, this learning is divided into data sets, where the test set is 30% and the training set is 70%. From this, it can be found that the SVM algorithm model is particularly effective, as shown in Table 1, with an accuracy of 89.01%. In addition, this study evaluated the correlation of the feature values of this experiment at the beginning, and the results are shown in the Figure 1, there is a strong correlation between the heart disease patients and the features chest pain type, maximum heart rate achieved, exercise-induced angina, and ST depression with a correlation coefficient of more than 0.4; and there is a certain correlation with the features age, sex, st\_slope, num\_major\_vessels, and thalassemia. The correlation coefficient is between 0.2 and 0.4; and the correlation with chol and fasting\_blood\_sugar is weak.

Table 1: Model performance.

Model	Training Accuracy	Testing Accuracy
Logistic Regression (LF)	0.8956	0.8774
Decision Tree (DT)	0.8323	0.7282
Random Forest (RF)	0.8864	0.8721
Naive Bayes (NB)	0.8241	0.7442
K-Nearest Neighbors (KNN)	0.8147	0.7623
Support Vector Machines (SVM)	0.9128	0.8901

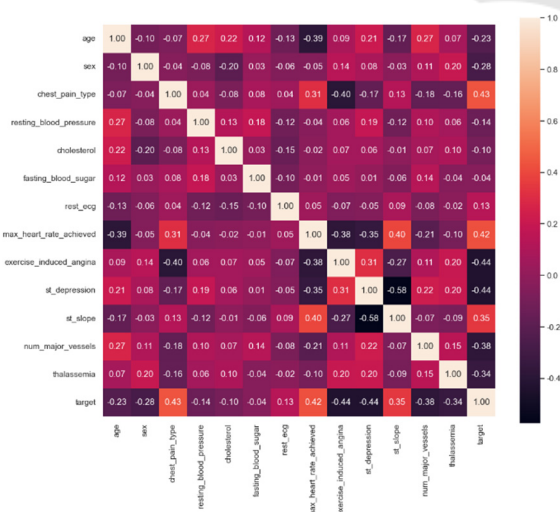


Figure 1: The correlation map of features (Picture credit: Original).

However, through further analysis, as shown in Figure 2, it can be found that angina induced by exercise is not strong evidence to confirm the diagnosis of heart disease.

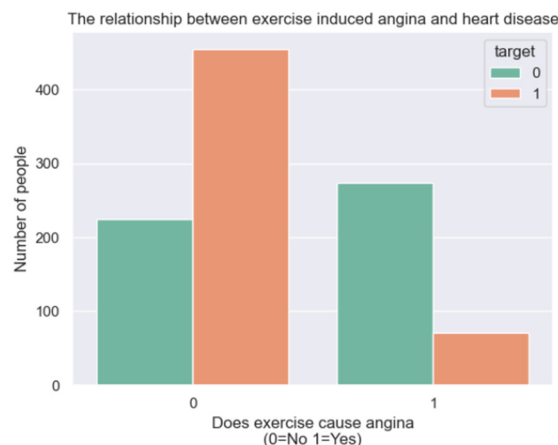


Figure 2: The relationship between exercise induced angina and heart disease (/Picture credit: Original).

Through the data in Table 1, it can be observed more intuitively that LR, RF and SVM algorithmic models have a high degree of accuracy. The reason for this is that, firstly, the problem that needs to be dealt with in this research is a binary classification problem, i.e., predicting whether or not one will suffer from heart disease. Secondly, all three algorithmic models have good robustness and can handle outliers or noise well and are not easily affected by extreme samples, and secondly they are highly interpretive, which helps to understand the contribution and impact of different features on heart disease, but the more complex feature relationships between the features in this case, and the larger number of feature values to be categorized, result in the Logistic Regression and Random forest algorithms being slightly less accurate than the SVM algorithm.

At the same time, examining the data presented in Table 1, it is not difficult to find that the two types of algorithmic models, Naive Bayes and K-Nearest Neighbors, however, perform poorly in this study. Through the study of the Naive Bayes algorithm, it is not difficult to find that the premise of this algorithm is to assume that all the features are independent of each other, but it cannot be well avoided that there may be a certain correlation between the different features of this kind of problem, for example, blood pressure and cholesterol levels, and the algorithm usually assumes that the features are discrete, and the processing of the continuous features may cause a certain degree of error, such as age, blood pressure etc.

The KNN algorithm, on the other hand, is more sensitive to the balance of the data. In this study, it can be observed that the uneven distribution of the number of samples of certain types of eigenvalues, such as gender, type of chest pain, etc., which is the main problem that leads to the unsatisfactory prediction accuracy of the model of this algorithm.

## 4 CONCLUSIONS

In this article, the study was conducted by using different machine learning algorithms to make predictions about heart disease. These six algorithms include Logistic Regression, Decision Tree, Random Forest, Naive Bayes, K-Nearest Neighbors, and Support Vector Machines, and finally the best model was found for this study, which is the Support Vector Machines classifier model, whose accuracy reaches 89.01%.

In fact, there are some shortcomings in this study. For example, the samples included in the dataset were not balanced enough, the study did not reach the expected accuracy of 95%, and only some machine learning models were used in the training of the model, and some deep learning algorithms were not considered to be used to make predictions.

In the future, in order to have a more balanced data sample, the study will continue to collect relevant data for the machine's learning, and the algorithmic model of the experiment to adjust the parameters and optimization, and will start to try to use some better algorithms to complete the optimization of the prediction results, such as XGBoost, CatBoost etc. Next, the research will expand the heart prediction problem into a multi-classification problem, classifying heart disease into different types or severity levels to better assist doctors in risk assessment and personalizing treatment for patients.

## REFERENCES

- Kaggle. 2019. Heart disease dataset. Retrieved from <https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset>
- Ouyang, D., He, B., Ghorbani, A. et al. 2020. Video-based AI for beat-to-beat assessment of cardiac function. *Nature*, 580, 252–256. <https://doi.org/10.1038/s41586-020-2145-8>
- Qiu, Y., Wang, J., Jin, Z., Chen, H., Zhang, M., & Guo, L. 2022. Pose-guided matching based on deep learning for assessing quality of action on rehabilitation training. *Biomedical Signal Processing and Control*, 72, 103323.
- Qiu, Y., Hui, Y., Zhao, P., Cai, C. H., Dai, B., Dou, J., ... & Yu, J. 2024. A novel image expression-driven modeling strategy for coke quality prediction in the smart cokemaking process. *Energy*, 130866.
- Sun, G., Zhan, T., Owusu, B.G., Daniel, A.M., Liu, G., & Jiang, W. 2020. Revised reinforcement learning based on anchor graph hashing for autonomous cell activation in cloud-RANs. *Future Generation Computer Systems*, 104, 60-73.
- Topol, E. J. 2023. As artificial intelligence goes multimodal, medical applications multiply. *Science*, 381, eadk6139. DOI:10.1126/science.adk6139
- Wu, Y., Jin, Z., Shi, C., Liang, P., & Zhan, T. 2024. Research on the Application of Deep Learning-based BERT Model in Sentiment Analysis. *arXiv preprint arXiv:2403.08217*.
- Yang, Y. et al. 2022. Artificial intelligence-enabled detection and assessment of Parkinson's disease using nocturnal breathing signals. *Nature Medicine*. doi:10.1038/s41591-022-01932-x
- Zhang, X., Wu, C., Zhang, Y. et al. 2023. Knowledge-enhanced visual-language pre-training on chest radiology images. *Nat Commun*, 14, 4542. <https://doi.org/10.1038/s41467-023-40260-7>
- Zhou, Y., Osman, A., Willms, M., Kunz, A., Philipp, S., Blatt, J., & Eul, S. 2023. Semantic Wireframe Detection. *publica.fraunhofer.de*.