CARDIAC DISEASE ANALYSIS AND DISEASE DETECTION BASED ON MACHINE LEARNING ALGORITHMS

Authors: Turaeva M.[1]

[1]TATU assistant teacher.

phone: +998(93)12566521, e-mail: max1209uz@gmail.com

Abstract.

As the incidence of heart disease continues to rise globally, there is an urgent need for accurate and efficient methods to detect and prevent this debilitating condition. To address this need, this paper proposes a machine learning-based medical system for predicting the likelihood of heart disease occurrence in patients. The study utilizes the UCI dataset to analyze multiple indicators using eight different algorithms to identify the most accurate and comprehensive attributes for predicting heart disease.

Keyword: CVD, Logical Regression, KNN, SVM, Naive Bayesian, Decision Tree, Random Forest, Gradient Boosting,

Introduction. Cardiovascular disease (CVD) is a group of diseases involving the heart and brain vessels. Symptoms include angina pectoris, myocardial infarction, heart failure, etc. It is a disease that seriously affects the function of the heart. According to the World Health Organization, heart disease is currently responsible for approximately 17.5 million deaths worldwide each year on the world heart day, accounting for 30% of all deaths [1]. And in China, approximately 3 million people die from cardio cerebrovascular diseases every year [2]. The heart is difficult to effectively deliver the blood needed by the body to various parts of the body to fulfill normal physiological needs. This is a common condition in patients with heart disease, leading to a series of problems such as breathing difficulties. Symptoms related to heart disease include chest pain, difficulty breathing, and elevated blood pressure. And unhealthy lifestyles and poor diet can also increase the risk of heart disease.In the past few decades, research and analysis in the field of heart disease have accumulated numerous medical data. These datasets reflect the condition and etiology of patients with heart disease. With the development of stochastic artificial intelligence, machine learning has been widely used in health care. It has become an effective tool for the rapid mining and analysis of data and for making effective predictions and decisions in the diagnosis of cardiac diseases. Many studies have shown that machine learning has achieved very high accuracy in problems based on judgment of heart disease [3]. In recent years, the classification method based on machine learning is one of the most accurate, stable and reliable methods.

Method. This study focuses on developing a predictive diagnostic framework to aid the medical system in identifying heart disease patients at an early stage to minimize the risk associated with it. Leveraging machine learning to assess the risk of heart disease can enable early detection and corresponding response measures, thereby alleviating the hospital care burden. The heart disease prediction and diagnosis framework is shown in Figure 1. This study collected data from potential heart disease patients, put these data into an analysis system, employed machine learning algorithms, and compared the results of classifiers. The ensuing analysis is subsequently utilized to select the optimal model for ascertaining the presence of heart disease in the patient.

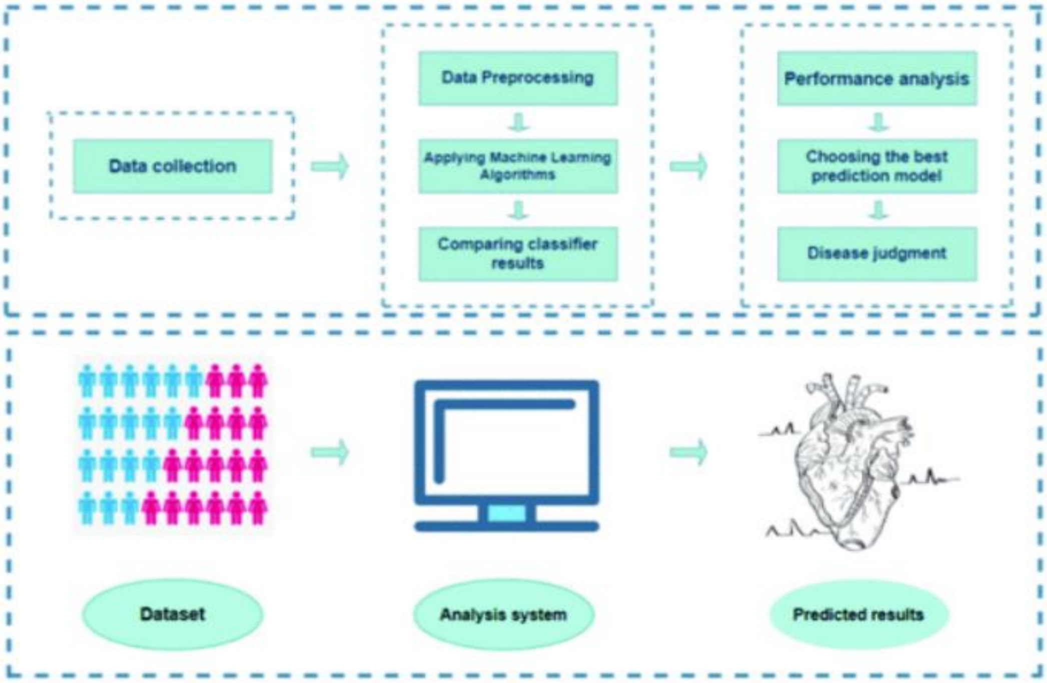
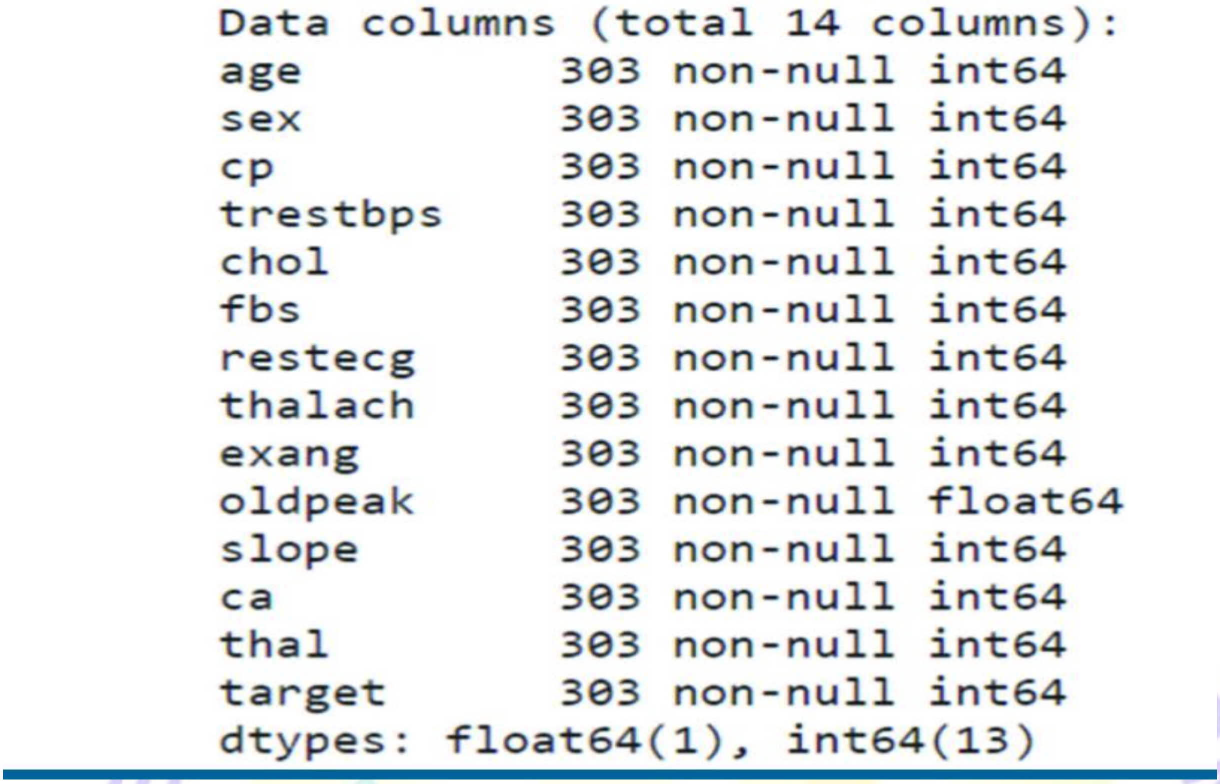


Figure 1. The whole process of heart disease analysis and judgment



```
Data columns (total 14 columns):
age         303 non-null int64
sex         303 non-null int64
cp          303 non-null int64
trestbps    303 non-null int64
chol        303 non-null int64
fbs         303 non-null int64
restecg     303 non-null int64
thalach     303 non-null int64
exang       303 non-null int64
oldpeak     303 non-null float64
slope       303 non-null int64
ca          303 non-null int64
thal        303 non-null int64
target      303 non-null int64
dtypes: float64(1), int64(13)
```

Figure 2. 14 attributes that may affect heart disease

A. Dataset description and preprocessing

The dataset used in this study is from the Cleveland Heart Disease Dataset provided by the UCI Machine Learning Repository [8]. The database contains 76 attributes, and this paper's experiment uses 14 of which are widely used in a large number of experiments, as shown in Figure 2. These 14 factors have been used in many papers, and [9]-[11] believe that these factors play an important role in their research.The data pre-processing involves transforming categorical variables using one-hot encoding technique. After creating pseudo variables, unnecessary variables appear in the data frame. These variables will be deleted. The 'target' column will be separated from independent columns. Data normalization uses a minimum-maximum normalization method to normalize data to normalize the range of independent variables or features.

The data correlation of Pearson Correlation Coefficient (PCC) was used to determine the correlation between attributes. The PCC varies from - 1 to+1, with positive and negative values indicating a high positive correlation and a high negative correlation between variables, respectively. In Figure 3, it can be observed that the correlation between the attributes chol and fbs as well as the attribute class is close to 0, indicating that the correlation between the two and the attribute class is minimal. Therefore, deleting these features can improve the performance of the model.

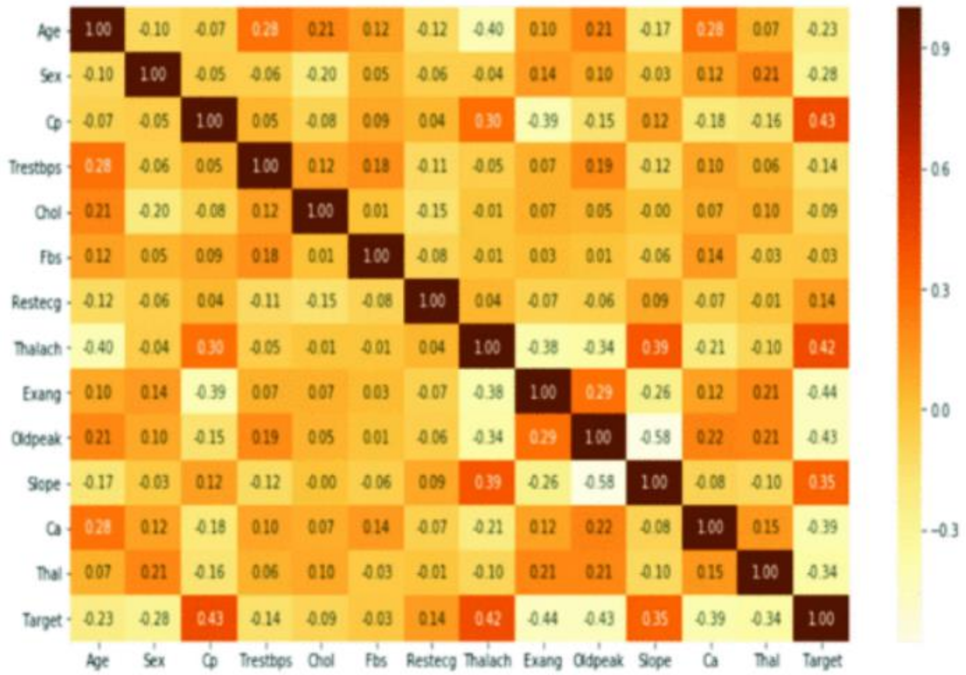


Figure 3. Heat map of each feature

B. Machine learning models. This study employed the classifiers based on the Logical Regression, KNN, SVM, Naive Bayesian, Decision Tree, Random Forest, Gradient Boosting, and AdaBoost classifiers in Sklearn's library. The introduction of these models can be found as follows:

Logical Regression: Logistic Regression is a generalized linear regression and a simple classifier for binary classification problems. It has significant effect on prediction of heart disease. The formula for logistic regression is as follows:

KNN: The KNN algorithm is a classification method that measures the distance between different feature values to perform classification. The nearest neighbor algorithm is a method for distinguishing all data samples in a dataset. In this article, the value of K is taken as 5.

SVM: The idea of the SVM algorithm is to establish the optimal decision boundary, perform binary classification of data, and place new data samples in the correct category.

Naive Bayes: Naive Bayes classifiers are based on Bayesian algorithms that assume specific values independent of any other feature. In many cases, naive Bayesian classification tools have good predictive accuracy.

Decision Tree: The decision tree algorithm is a prediction model that utilizes approximately discrete function values, representing a mapping relationship between many attributes of heart disease and prediction results. It is a tree structured classifier in which internal nodes represent the characteristics of the dataset, branches represent algorithm rules, and each leaf node represents the prediction results.

C. Performance Evaluation Metrics. Different performance evaluation indicators have been used for classifier performance evaluation [12]. This article uses four indicators: accuracy, accuracy, recall, and F1-Score for evaluation classification and regression analysis to determine the final model selection. The following is an introduction to these four indicators.Accuracy: theproportion of the quantity correctly predicted by the model to the

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

total quantity:

Precision: the proportion of samples with positive prediction results that are actually positive:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \qquad (2)$$

Recall: the proportion of samples predicted to be positive in the actual positive samples of the model:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \qquad (3)$$

F1-Score: harmonized average of accuracy rate and recall rate:

$$\text{F1 - Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \qquad (4)$$

TP: True Positive, a positive sample predicted by the model as a positive class.

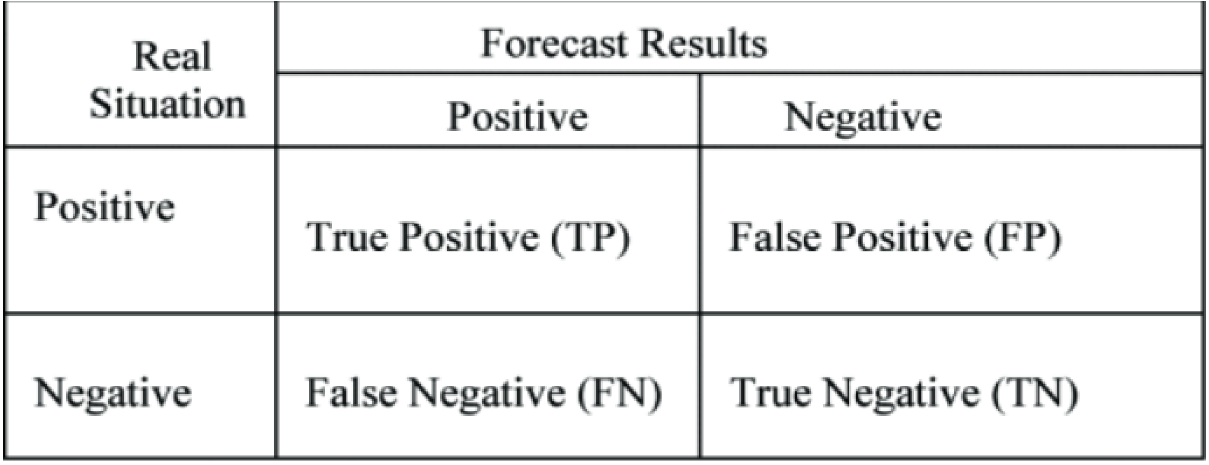
TN: True Negative, negative samples predicted as negative by the model

FP: False Positive, negative samples predicted as positive by the model.

FN: False Negative, a positive sample predicted by the model as a negative category.

Experimental results and discussion. Based on the correlation of various characteristics analyzed by heat map in Figure 3, this study selected five characteristics with high correlation, namely "age", "trestbps", "chol", "thalach", and "peak age" for analysis.

Table I. Confusion matrix



Real Situation	Forecast Results	
	Positive	Negative
Positive	True Positive (TP)	False Positive (FP)
Negative	False Negative (FN)	True Negative (TN)

B. Analysis on the correlation of three indicators: Chol, Trestbps, Thalach

Distribution diagram based on Trestbps and Thalach. As can be seen from Figure 4, target 1 indicates a heart attack, and target 0 indicates no heart disease. With the increase of trestbps, there is no corresponding significant change in the number of thalachs, indicating that there is no close relationship between the two indicators.

Distribution diagram based on chol and Thalach. As can be seen from Figure 5, when the chol level is maintained at a normal level, the thalach level is often low, indicating a linear relationship, indicating a close relationship between the two indicators.

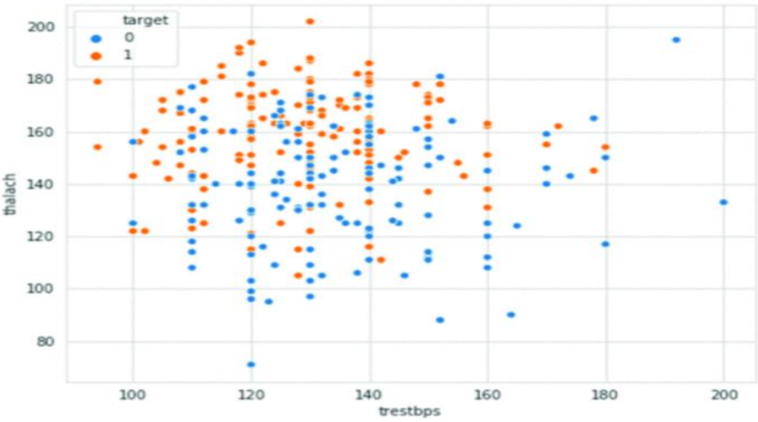


Figure 4. Distribution diagram of Trestbps and Thalach.

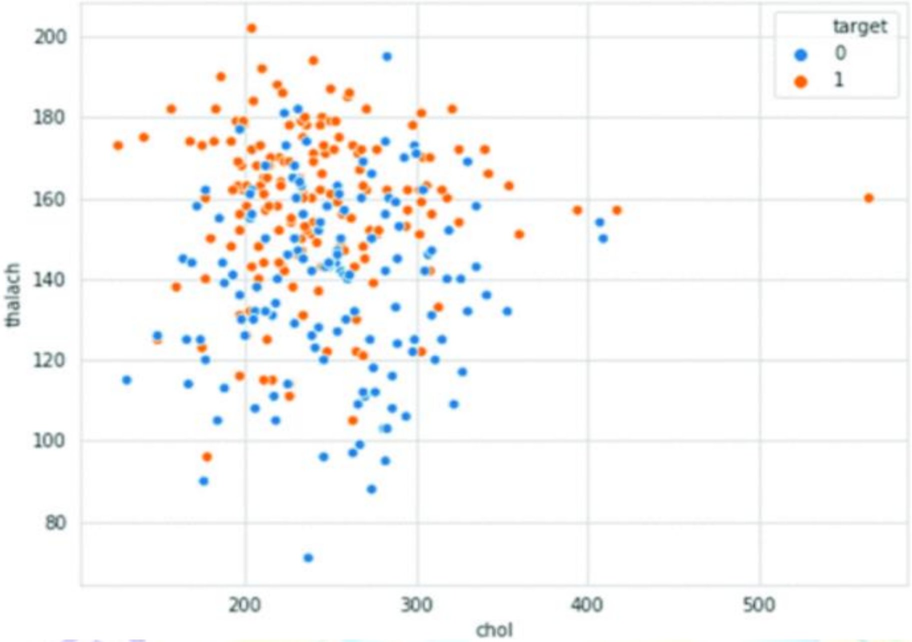
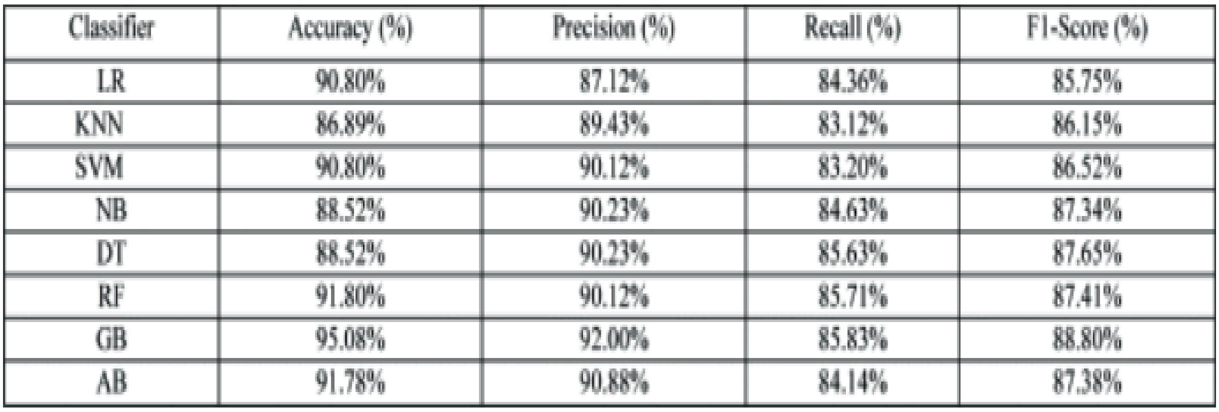


Figure 6. Distribution diagram of chol and Thalach.

C. Different machine learning algorithms predict the outcome of heart disease The following table lists the Accuracy, Precision, Recall, and F1-Score values of eight machine learning algorithms, as shown in Table II. It can be seen that the accuracy rates of the GB algorithm and the Random Forest algorithm have respectively reached 95.08%, 91.80%.

Table II. Performance of various indicators of the classifier



Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
LR	90.80%	87.12%	84.36%	85.75%
KNN	86.89%	89.43%	83.12%	86.15%
SVM	90.80%	90.12%	83.20%	86.52%
NB	88.52%	90.23%	84.63%	87.34%
DT	88.52%	90.23%	85.63%	87.65%
RF	91.80%	90.12%	85.71%	87.41%
GB	95.08%	92.00%	85.83%	88.80%
AB	91.78%	90.88%	84.14%	87.38%

D. Model analysis and evaluation.The accuracy of the eight classifiers is shown in Table II and Figure 7. It can be observed that the accuracy of GB classifier reaches the highest 95.08%. Based on the different accuracy rates obtained by the eight algorithms, the classifier with the highest accuracy rate will be selected, and the characteristics of the classifier will be analyzed from the confusion matrix shown in Table I, ROC curve and its AUC values, learning curve (including training score and cross-validation score), and binary P-R curve.

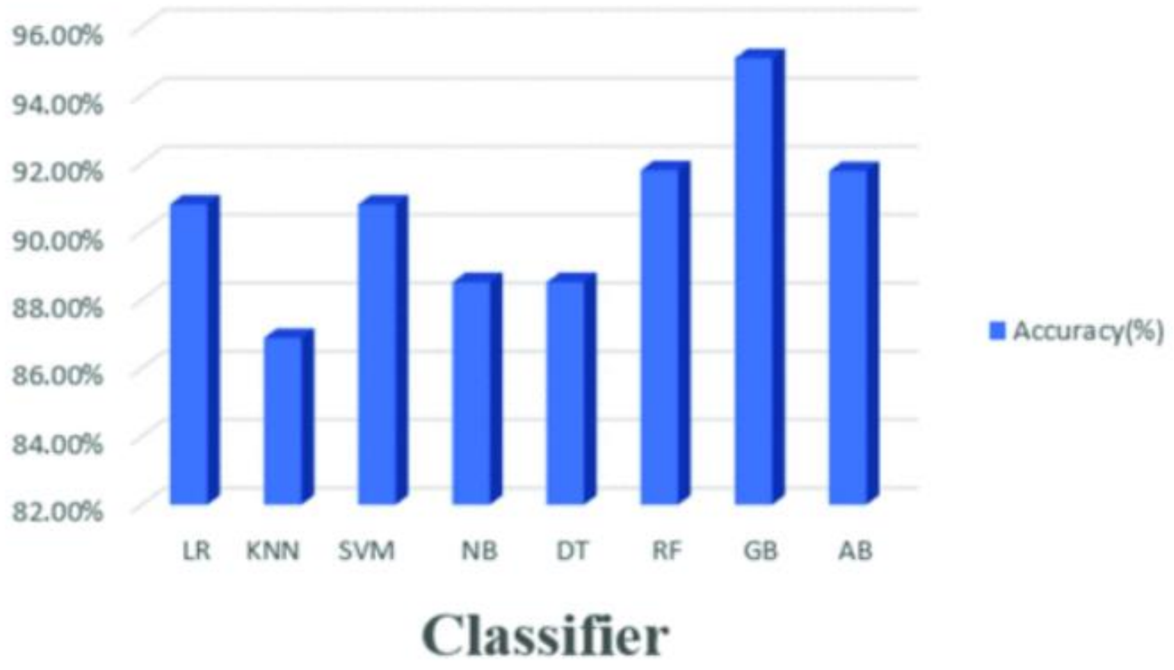


Figure 6. Accuracy visualization table

Conclusion. This paper proposed a heart disease prediction model based on machine learning to address the issue of managing the large number of heart disease patients in hospitals. The proposed model employs eight different algorithms to evaluate its feasibility, and the gradient enhancement classifier is ultimately selected with an accuracy rate of 95.08%. Further analysis of the model through the confusion matrix, ROC curve, learning curve, and precision recall curve demonstrates that the model maintains good prediction accuracy, making it suitable for real-life applications. However, the model's limitations, such as long training time and overfitting, need to be addressed to enhance its prediction efficiency in future work. Optimization of the algorithm and structure of the model can further improve the prediction accuracy and versatility for different types of heart disease to promote better application in medical systems.

References

1. Yu H. Analysis and Prediction of Heart Disease Based on Machine Learning Algorithms //2023 8th International Conference on Intelligent Computing and Signal Processing (ICSP). – IEEE, 2023. – C. 1418-1423.
2. Yu, Haonan. "Analysis and Prediction of Heart Disease Based on Machine Learning Algorithms." *2023 8th International Conference on Intelligent Computing and Signal Processing (ICSP)*. IEEE, 2023.
3. Yu, H. (2023, April). Analysis and Prediction of Heart Disease Based on Machine Learning Algorithms. In *2023 8th International Conference on Intelligent Computing and Signal Processing (ICSP)* (pp. 1418-1423). IEEE.
4. Mohan N., Jain V., Agrawal G. Heart disease prediction using supervised machine learning algorithms //2021 5th International conference on information systems and computer networks (ISCON). – IEEE, 2021. – C. 1-3.
5. Mohan, Narendra, Vinod Jain, and Gauranshi Agrawal. "Heart disease prediction using supervised machine learning algorithms." *2021 5th International conference on information systems and computer networks (ISCON)*. IEEE, 2021.
6. Mohan, N., Jain, V., & Agrawal, G. (2021, October). Heart disease prediction using supervised machine learning algorithms. In *2021 5th International conference on information systems and computer networks (ISCON)* (pp. 1-3). IEEE.
7. Ali M. M. et al. Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison //Computers in Biology and Medicine. – 2021. – T. 136. –
8. C. 104672. Ali, Md Mamun, et al. "Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison." *Computers in Biology and Medicine* 136 (2021): 104672.
9. Ali, M. M., Paul, B. K., Ahmed, K., Bui, F. M., Quinn, J. M., & Moni, M. A. (2021). Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison. *Computers in Biology and Medicine*, *136*, 104672