Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/46/20241124

Risk prediction of acute myocardial infarction based on deep neural network

Guanyu Chen

Computer Science and Technology, Wuhan University of Technology, Wuhan, 430070, China

guanyuchen@whut.edu.cn

Abstract. The cardiovascular disease situation in China is becoming more and more serious, among which the prevalence and mortality rate of acute myocardial infarction (AMI) have been increasing in recent years, so it is urgent to prevent cardiovascular disease. Based on the personal behavior data and basic physical data covering 30,000 respondents within 30,000 days, a deep learning algorithm was used to predict whether the respondents would have acute myocardial infarction in the near future. The prediction accuracy and generalization ability of different algorithms, including logistic regression, decision tree, XGBoost, and GNB algorithms, are compared. The results show that it is feasible to predict the short-term AMI disease split line based on the patient's personal basic information, and the DNN algorithm can achieve the highest prediction accuracy. The research results can assist clinical medical staff in the prevention and treatment of acute myocardial infarction to a certain extent, and also have good civilian characteristics, which can serve medical insurance and other industries.

Keywords: Myocardial, Machine Learning, Multilayer Perceptron, Feature Selection.

1. Introduction

In recent years, in the medical field, using machine learning or deep learning algorithms to assist medical staff to complete disease diagnosis has always been a research hotspot and difficulty. At present, it has been widely used in the prediction of cardiovascular diseases, diabetes, tumors, kidney diseases and other diseases, and has achieved better application effect. In relevant application research, the researchers used the patient's personal information data as the input of the deep learning model, and the risk of disease occurrence as the predicted output, effectively assisting medical staff to take intervention measures in advance to reduce the risk of disease, and achieve early detection and early prevention. The purpose of diagnosis and treatment of early treatment.

Among the total causes of death between urban and rural residents in our country, cardiovascular disease death ranks first, higher than tumors and other diseases, according to the "China Cardiovascular Health and Disease Report 2020" released in July 2021. Myocardial infarction (MI), as a common cardiovascular disease, seriously threatens people's life and health. Since 2005, its mortality rate has been rising rapidly. AMI is a complication of multiple diseases, affected by multiple risk factors, and has an acute onset. Therefore, prediction based only on traditional risk factors of cardiovascular disease has low accuracy and poor explanatory power. Therefore, the prediction of AMI needs to be based on more variables and try different machine learning algorithms.

DOI: 10.54254/2755-2721/46/20241124

In this paper, the variables of various personal physiological indicators of the patient and the neural network algorithm are used to construct a prediction model for whether the patient will suffer from AMI [1-2]. And compared the prediction accuracy and generalization ability of different algorithms, including decision tree, logistic regression, Support Vector Machines and other algorithms. The research results show that it is feasible to predict the risk of AMI by combining the personal data of patients, and the Deep-Learning Neural Network (DNN) can achieve higher prediction accuracy. The research results have certain reference significance for improving the ability to predict the risk of AMI and better serving the disease prevention management of AMI, and can be applied to the insurance industry.

2. Related research

The characteristics proposed by Liang et al. and the constructed age stratified prediction model, Historical records can predict whether a patient diagnosed with MI will develop heart failure (HF) in a short period of time. The researchers used different algorithms of machine learning to build risk predictive models and verified the importance of age stratified characteristics. Among all models, random forests have the best prediction performance, but XGBoost predictions are not much different from random forests, but the calculation time is greatly reduced. Ibrahim et al. used three machine learning models of convolutional neural network (CNN), recurrent neural network (RNN) and XGBoost to predict the risk status of AMI, and found that the effect of XGBoost model was the best [3]. Azwari constructed a random forest model to predict myocardial rupture after acute myocardial infarction in hospitalized patients. Rossiev et al. applied neural networks to predict myocardial infarction complications.

It can be seen that the data used by researchers at home and abroad are very professional medical data, which requires high costs. None of these models can have a good civilian type and cannot play a preventive role. 1) In this paper, more recent behaviors of patients are included in the prediction of AMI, fully considering that recent behaviors have a greater impact on the disease, and innovations in feature selection are conducted. 2) Since the influencing factors of heart disease are too complex, the accuracy of all training models on the data set is poor, so this study uses relatively common personal daily information to construct a DNN to forecast the risk of myocardial infarction. 3) The accuracy and generalization of other different algorithms for AMI risk prediction were compared.

3. Methods

3.1. DNN neural network algorithm

A multi-layer perceptron is based on a single-layer neural network, adding one to more hidden layers on the basis of a single-layer neural network [4], and in order to overcome the limitations of the linear model to use a nonlinear activation function to improve nonlinear data processing capabilities, and its basic structure is shown in Figure 1.

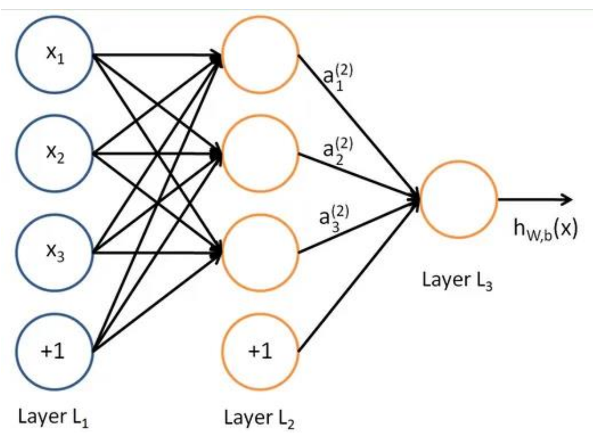

Figure 1. Schematic diagram of DNN model.

From figure 1, DNN is composed of an input layer, an output layer and several hidden layers. The input layer is responsible for taking the j features of myocardial infarction patient's i as input, and the neuron nodes in the hidden layer take the weighted summation of the input features of the input layer, and generate an activation response through a nonlinear activation function, so that this allows the model to approach arbitrary nonlinear functions. During the training process, the forward propagation and back propagation algorithms are adopted to continuously adjust the model weights and bias parameters, so as to achieve better classification results.

3.2. Logistic linear regression algorithm

Logist Regression [5] is a widely used statistical learning method. Its essence is a classification model. Its basic idea is to introduce nonlinearity into the model through the Sigmoid function on the basis of linear regression, and map the independent variable to 0 ~ 1 And as the probability of event occurrence, 0.5 is taken as the threshold value, above 0.5, myocardial infarction disease occurs, and below 0.5, myocardial infarction disease does not occur.

3.3. Decision tree algorithm

The decision tree takes a single node as the root node [6], which stores all the data, and classifies and recurses the existing nodes based on the object characteristics, so as to divide them into more nodes. When the divided node satisfies the termination condition, the node is set as a leaf node, and the classification is completed.

3.4. XGBoost algorithm

XGBoost is an efficient, flexible and scalable gradient boosting decision tree ensemble algorithm, proposed by Chen Tianqi in 2014 [7][8]. The model adopts the Boosting modeling idea when performing decision tree integration, and establishes multiple regression decision trees, and the outputs of each decision tree are interrelated. Every time a new decision tree is added, the existing decision tree needs to be considered. Error, so that the accuracy of the model can be continuously iteratively improved.

3.5. Gaussian NB Algorithm

GNB is a classification technique based on probabilistic methods and Gaussian distribution for machine learning [9-10]. In this algorithm, the conditional probability of each feature variable follows a Gaussian distribution, and then the posterior probability of a new sample for each class is calculated according to Bayes' formula. Finally, the sample type is determined by maximizing the posterior probability.

4. Analysis of experimental results

4.1. Experimental details

1) Training platform: This study mainly uses pycharm software for data cleaning, data visualization, model training, and model comparison.

2) Training parameters: a) There are two hidden layers in DNN, the first layer contains 64 neurons. The second layer contains 20 neurons. The activation function of them is relu function. b) The reciprocal of the regularization coefficient in Logist icRegression is 100, and the random number seed is 1. c) The number of trees in XGBoost is 10, the depth is 4, the minimum weight of leaf nodes is 1, and the random number is 27.

4.2. Evaluation index

In order to verify the validity of the prediction model for complications in patients with myocardial infarction, it is necessary to evaluate its prediction effect. In this paper, the model is evaluated from three dimensions: precision, recall, and AUC.

(1) Accuracy: the degree of accuracy in predicting the occurrence and non-occurrence of complications in patients with myocardial infarction, as shown in the expression (1).

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + FP + TN} \tag{1}$$

In the formula, TP represents the number of correctly predicted myocardial infarction patients with complications, TN represents the number of correctly predicted myocardial infarction patients without complications, FN represents the number of wrongly predicted myocardial infarction patients without complications, and FP represents the number of patients with complications who wrongly predicted myocardial infarction.

(2) Recall rate recall: the proportion of correctly predicting the number of complications in patients with myocardial infarction to the number of complications in the sample set, as shown in the expression (2).

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2}$$

(3) AUC [11]: It represents the area enclosed by the abscissa axis under the ROC curve, which is used to judge the advantages and disadvantages of the model effect, and its value is generally between 0.5 and 1.

4.3. Analysis of the experimental data set

The data set used in this article comes from the personal key indicator data set of myocardial infarctions in the UCI machine learning database of the University of California, Irvine, with a total of 319,774 cases. Depression, etc.), a total of 18 input feature vectors, whether suffering from myocardial infarction is used as the sample label result.

The dataset contains sample data of 319,774 patients, and each patient sample contains 18 personal key indicator data (Table 1).

Table 1. Dataset Introduction.

variable name abbreviation	Specific measurement content	type of data	value
Heart Disease	Whether you have heart disease or myocardial infarction	binary	0,1
BMI	Body mass Index	Discrete	12, 24, 27
Smoking	do you smoke	binary	0,1
AlcoholDrinking	Alcoholism	binary	0,1
Stroke	Have you ever had a stroke	binary	0,1
Physical Health	Whether the individual is sick or injured within 30 days	binary	0,1
Mental Health	Whether the individual feels mentally unhealthy within 30 days	binary	0,1
DiffWalking _	Whether the individual has difficulty walking	binary	0,1
Sex	gender	binary	0,1
AgeCategory	Age range distribution	Discrete	12, 35, 77
race	Race	character type	White
Diabetic	Whether you have diabetes	binary	0,1
PhysicalActivity	personal judgment of health	Discrete	12, 24, 27
GenHealth _	Personal assessment of mental condition	Discrete	0, 10, 15
Sleep Time	Individual's average sleep time per day	Discrete	6, 8, 10
Asthma	Whether the individual has a history of asthma	binary	0,1

DOI: 10.54254/2755-2721/46/20241124

Table 1. (continued).

Kidney Disease	Whether the individual has kidney disease	binary	0,1
Skin Cancer	Whether the individual suffers from a skin disease	binary	0,1

Above eigenvectors can get the correlation matrix shown in Figure 2. That is to say, there is only a weak correlation between walking difficulty and physical health indicators, while there is almost no correlation between other feature vectors.

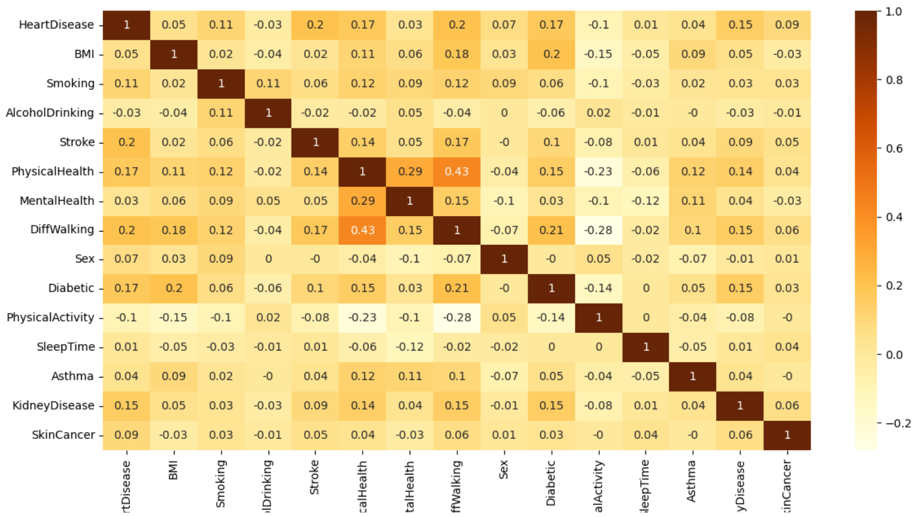

Figure 2. Myocardial infarction correlation matrix.

Myocardial infarction is common in middle-aged and elderly people. The age distribution of patients with myocardial infarction in this data set is shown in Figure 3. It can be seen from the figure that the age of patients with myocardial infarction is mostly concentrated between [40,80].

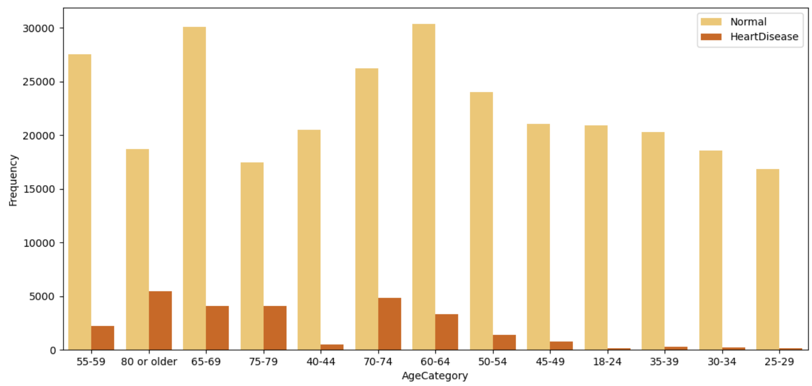

Figure 3. Age Distribution of Myocardial Infarction.

In many medical models, BMI, gender, and smoking are all very important indicators, so this paper also observes data such as BMI, gender, and smoking. From Figure 4, it can be seen that in patients with myocardial infarction Among the respondents and those without myocardial infarction, most of the BMIs were concentrated in [20,40], and there was no obvious difference. In gender and smoking data, the sample is relatively balanced and the correlation with the prevalence of myocardial infarction is weak.

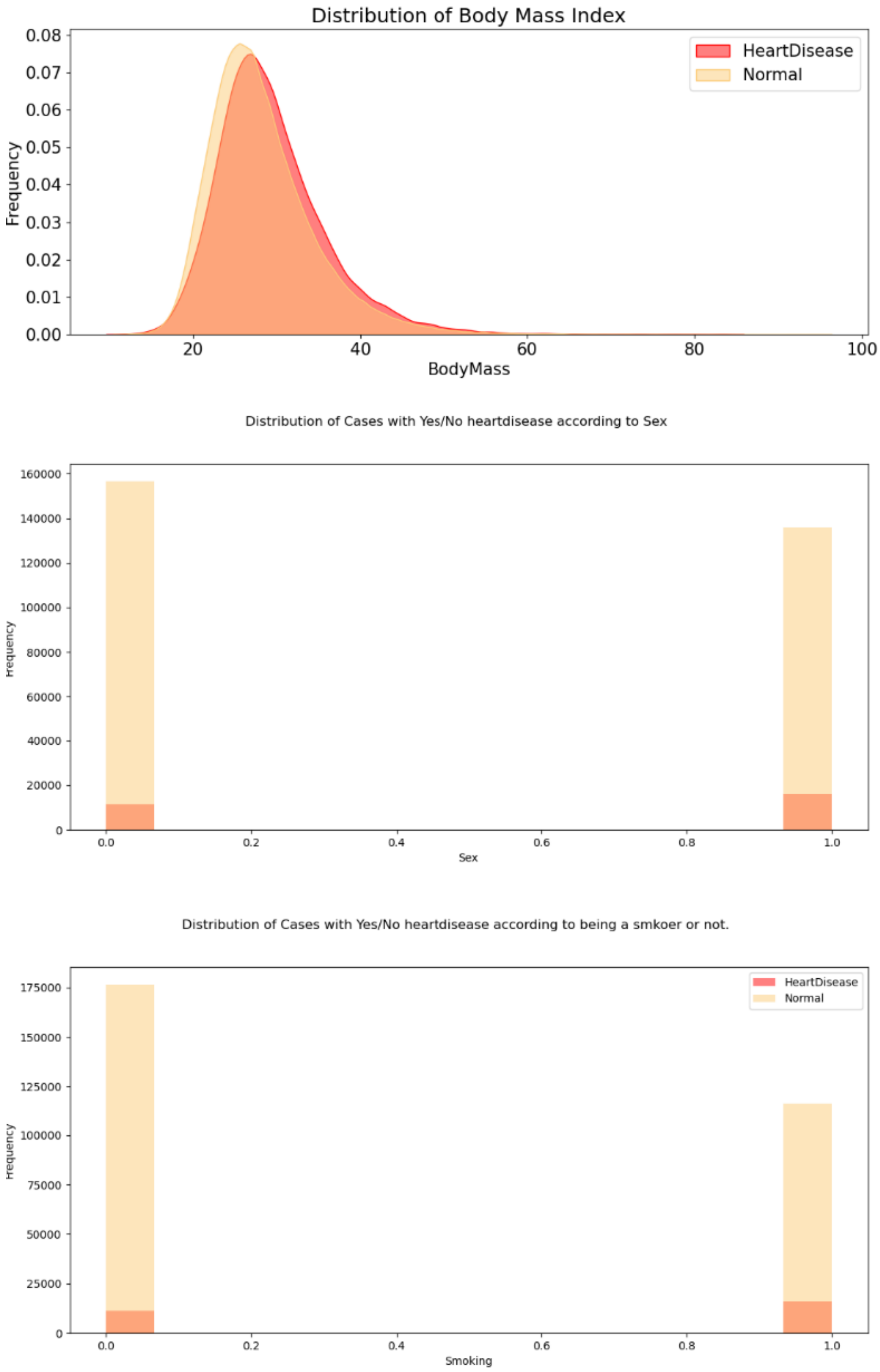

Figure 4. BMI/SEX/S moking data visualization.

According to Figure 2, heart disease is strongly correlated with stroke, diabetes, and kidney disease. Data visualization is performed for these four factors. As shown in Figure 5, according to the data analysis, the probability of myocardial infarction in patients with stroke, diabetes and kidney disease is much higher than that of the group without the three diseases.

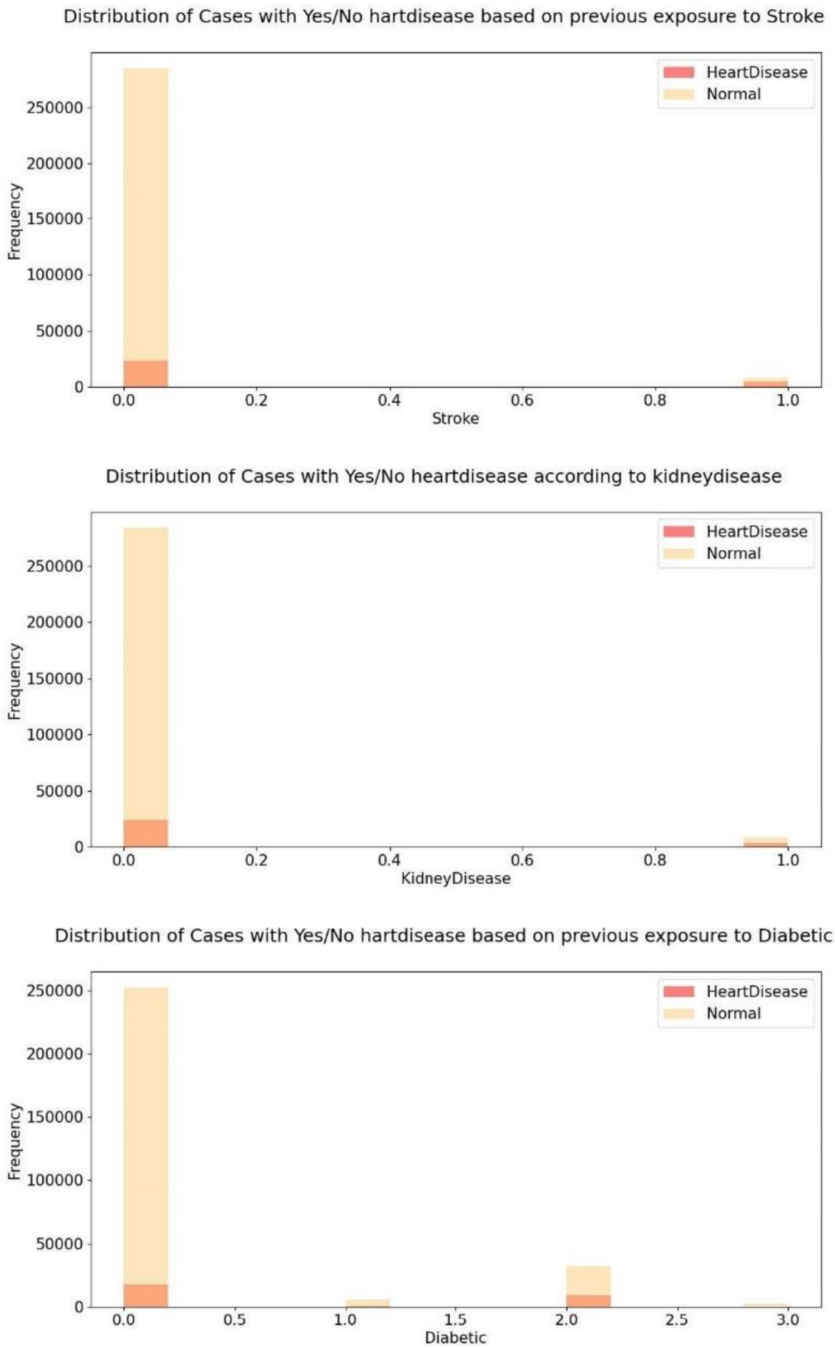

Figure 5. Stroke/kidney disease/diabetes data visualization.

4.4. *Analysis of model results*

DNN, GaussianNB, XGBoost, Logistic Regression and Decision Tree predict each complication in the data set, and their performance is shown in Figure 6. The DNN model has a very good performance in terms of precision and recall, which shows that the accuracy rate of the correct results in the classification results of the DNN model is higher. On Accuracy: Except for Gaussian Naive Yebes, other models are above 80%, and some models have reached above 90%, while the DNN model has a better fit for the data, and its accuracy rate reaches 94.57%. DNN has a better sorting ability for the overall sample and performs better on auc (Table 2).

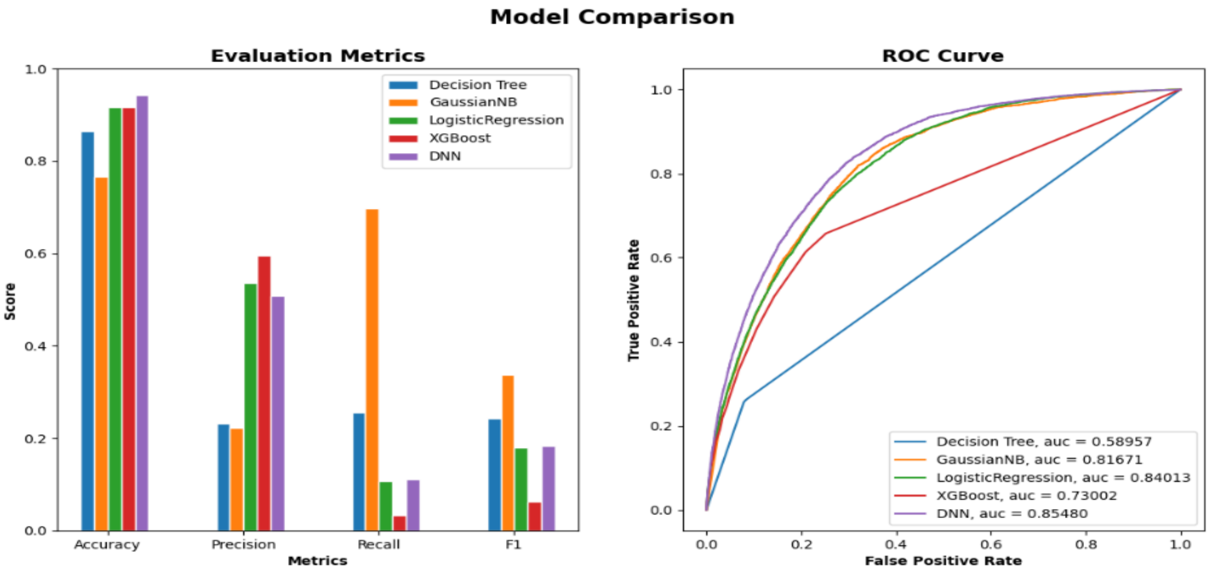

Figure 6. Schematic diagram of the results.

Table 2. Algorithm Index Comparison.

Method	Acc	Pre	Recall	F1
Decision Tree	0.86469	0.23222	0.25545	0.24326
Gaussian NB	0.76638	0.22224	0.69757	0.33708
LogisticRegression	0.91611	0.53678	0.10852	0.18054
XGBoost	0.91575	0.59539	0.03323	0.06295
DNN	0.94192	0.50753	0.11146	0.18278

5. Conclusion

At present, researchers focus on mining the association between some characteristic data of the diagnosis and treatment process and the incidence of myocardial infarction, ignoring the medical history data of patients before diagnosis and treatment. This paper uses the public real data set, according to the personal information data of the respondents, through the DNN model in the deep learning algorithm, to establish a risk prediction model for acute myocardial infarction, and compares the accuracy of different models. Experiments show that the DNN model has achieved good prediction results under this data set, and it has a certain reference value for clinical medical staff to detect myocardial infarction early.

This article has certain reference significance for predicting the possibility of AMI in the future by combining the respondents' physical indicators and short-term behavior. Most of the related research is based on the patient's physiological indicators to predict the risk of disease, and most of the prediction models' prediction cycle too long. In addition, the research ideas in this paper have certain reference value for the insurance industry, and can better set the insurance compensation amount, premium and other terms.

In follow-up research, on the one hand, it is necessary to continuously optimize model parameters, try to use more classification measurement models, and further improve the prediction accuracy, recall rate, and AUC value of the model; on the other hand, deep learning and other methods can be used to continue in-depth analysis of existing Input the internal relationship between the characteristics and the prevalence of myocardial infarction, and rank the importance of the influencing factors of the prevalence.

References

[1] Yang Chushi. Risk Prediction of Acute Myocardial Infarction Based on Behavioral Data. *Shanghai Jiaotong University*, 2020.

[2] Tan Pengliu, Xu Guangyong, Zhang Luyu, etc. Heart Disease Prediction Model Based on Convolutional Neural Network and Adaboost. *Computer Applications*, 2023, 43(S1): 19-25.

[3] Liu Hailiang, He Yalin, Chen Jianhao, etc. Design and Implementation of Smart Home Elderly Care System for the Elderly Living Alone. *Internet of Things Technology*, 2022,12(08):112-116.

[4] Cai Zengyu, Cui Mengmeng, Hou Jialin, etc. Design and Implementation of Heart Disease Prediction System Based on Residual Neural Network. *Modern Electronic Technology*, 2022, 45(03): 121-124.

[5] Lu Zheng, Wei Kai, You Zhijun, etc. Risk assessment method for cable adsorption card based on logistic regression. *Science Technology and Engineering*, 2023, 23(16): 6862-6869.

[6] Myles A J, Feudale R N, Liu Y, et al. An introduction to decision tree modelling. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 2004, 18(6): 275-285.

[7] Ogunleye A, Wang Q G. XGBoost model for chronic kidney disease diagnosis. *IEEE/ACM transactions on computational biology and bioinformatics*, 2019, 17(6): 2131-2140.

[8] Lin Feng, Hai Chunlong, Mei Liquan. Multi-source aerodynamic data fusion modeling based on XGBoost. *Journal of Aerodynamics* 2021: 1-9

[9] Bansal A, Singhrova A. Performance analysis of supervised machine learning algorithms for diabetes and breast cancer dataset. *2021 International Conference on Artificial Intelligence and Smart Systems*, 2021: 137-143.

[10] Wang Q, Pan L, Shen J, et al. Power plant data filtering based on Gaussian naive Bayesian classification and prediction error method. *2019 Chinese Automation Congres*, 2019: 1490-1495.

[11] Huang J, Ling C X. Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on knowledge and Data Engineering*, 2005, 17(3): 299-310.