Heart Disease Prediction Method Based On ANN

Qizhong Chen [1], Ningyao Ma [2, *]

[1] Nanjing University of Information Science and Technology, Nanjing, China

[2] Dalian University of Technology, Dalian, China

* Corresponding Author Email: mny011116@mail.dlut.edu.cn

Abstract. Heart disease is one of the main health problems worldwide, and heart disease prediction based on artificial neural networks (ANNs) has become a research hotspot, aiming to improve the accuracy of early diagnosis. At present, heart disease prediction still faces problems such as data complexity, feature selection, overfitting, etc., which need to be overcome in ANN models. The heart disease prediction based on ANN has potential importance, which can assist medical decision-making, reduce the incidence rate of heart disease, and provide support for personalized treatment and health management. This study proposes a heart disease diagnosis method relied on machine learning, which balances the dataset using the SMOTE (Synthetic Minority Over Sampling Technique) algorithm as well as ENN (Edited Nearest Neighbors) algorithm. Then, research was conducted based on a variety of machine learning models, such as LR (logistic regression), DT (decision tree), RF (random forest), GBDT (gradient enhancement decision tree), XGBoost (extreme gradient enhancement), SVM (support vector machine), and ANN (deep neural network), applied to predict heart disease datasets. Patients can measure the risk of heart disease through past medical history, household equipment, and personal habits. Research has shown that the auc and recall based on the ANN model are as high as 0.808 and 0.81, respectively.

Keywords: Artificial Neural Network, Heart Disease Prediction, Machine Learning.

1. Introduction

Nowadays, one of the most severe causes of death and health issues throughout the whole world is heart disease. According to statistics, about 17.9 million victims die from cardiovascular disease every year, accounting for 31% of the total deaths all over the world [1]. This serious cardiovascular disease will affect people of all ages and become an important challenge in the field of public health. If the disease can be found as soon as possible and active preventive measures can be taken, the mortality can be greatly reduced [2]. Therefore, the early and exact foresight for heart disease is of great significance for prevention, intervention and treatment.

Artificial neural networks have been implemented extensively in the medical industry recently, which shows that great promise in areas like disease prediction, medical photograph interpretation and so on. Spontaneously, the application of ANN to evaluate the likelihood and risk of developing heart disease has grown into a specialized research area. Nevertheless, the existing research mainly rely on professional medical equipment to measure the data which is needed, running out of multiple resources as well as labor and causing unnecessary inconvenience for patients.

At present, in the field of intelligent diagnosis of heart disease, the mainstream methods are generally random forest, Ann and so on. For example, Garg et al. uses two supervised machine learning algorithms, k-NN and random forest. By considering some attributes, such as chest pain, cholesterol level, age, etc., people with heart disease and people without heart disease are classified. The prediction accuracy of nearest neighbor method is 86.885% [3], and the prediction accuracy of random forest method is 81.967%. As another example, sun et al. Proposed an improved sparse automatic encoder based artificial neural network to help forecast heart disease. In the initial phase, a sparse self-encoder (SAE) is trained to determine the optimal way to represent data; In the following phase, according to the learning records, artificial neural network (ANN) is utilized to forecast the health state. Then use Adam algorithm to optimize SAE and apply batch normalization. The two-stage method effectively improves the classification effect of neural network and has stronger robustness than other methods. The accuracy of the model to the test data is 90% [4]. Most of the data

they use are from the Kaggle website, such as the Kaggle Framingham Heart dataset [5], due to the fact that there exist more samples in these databases.

Research has shown that fundamental health actions such as smoking, physical activity, diet and weight have a certain impact on heart disease [6]. Other factors including hypertension, high blood cholesterol, physical illness, stress levels, alcohol consumption, and irregular diet are also potential causes of heart disease [7].

2. Method

2.1. Pipeline

The study used the "SMOTE" algorithm as well as "ENN" algorithm to balance the dataset. Then, research is conducted on the prediction of heart disease datasets depended on all kinds of machine learning models, such as LR (logistic regression), DT (decision tree), RF (random forest), GBDT (gradient lifting decision tree), XGBoost (extreme gradient lifting), SVM (support vector machine), and ANN (deep neural network). Among them, the activation function of the hidden layer in the ANN model uses Relu, and the loss function uses binary_ Crosstropy, Batch_ Size=50, Epochs=100, with 50 neural units in the input layer, 30 neural units in the hidden layer, and 1 neural unit in the output layer. At the same time, a method based on AUC monitoring, optimal model callback, and dynamic optimization of learning rate is adopted. Overall flow chart as shown in Fig 1.

Fig1. Experiment workflow

2.2. Data balancing method

The most prevailing oversampling technique,which can be combined with plenty of various undersampling techniques as well, is highly likely to be "SMOTE" [8].

Synthetic Minority Oversampling Technology, regularly known as "SMOTE", is an enhanced method built on random oversampling algorithm. Notwithstanding the information the model learns is too specific and not sufficiently generalized, random oversampling, which uses a straightfoward mehod of copying samples to increase minority class samples, can easily result in model overfitting issues. The core principle of the "SMOTE" method is to assess minority class samples and manually

synthesize new samples based on minority class samples to add to the dataset, as shown in the figure. The following process is how the algorithm functions.

Step1: Determine k-nearest neighbor of each sample x by calculating its distance from every other sample in the minority class sample set Smin using the Euclidean distance as the reference.

Step2: Calculate the sampling rate N by setting a sampling ratio based on the sample imbalance ratio.Supposing the chosen k-nearest neighbor is x_n,randomly pick up multiple samples from each minority sample's k-nearest neighbors.

Step3. Harnessing the initial sample and each k-nearest neighbor x_n that was selected casually,create a new sample utilizing the formula below.

$$x_{new} = x + rand(0,1) * |x - x_n| \quad (1)$$

Another very popular undersampling method is the "Edit Nearest Neighbor" or "ENN" rule [9]. The EDIT NEAREST NEIGHBOR (ENN) algorithm is an algorithm used for data dimensionality reduction, which improves the performance of classifiers by removing redundancy and noise from sample data. This algorithm determines which samples should be deleted by comparing the distance between each sample and its nearest neighbor. Specifically, the "ENN" algorithm first uses the "ENN" algorithm to find the K nearest neighbors of each sample, and then compares whether the class labels between each sample and its nearest neighbors are the same. If the class labels of most of the nearest neighbors are different from the sample, remove the sample from the dataset.

The "ENN" algorithm works on the following principles. For each sample in the category to be undersampled, calculate its nearest neighbor sample. If the nearest neighbor sample does not match the category of the current sample, the current sample will be deleted from the dataset.

Through this process, the "ENN" algorithm can edit the dataset and delete samples that are not "enough" consistent with their neighboring samples. This algorithm determines whether samples should be retained by considering the neighboring samples around them. In terms of selection criteria, "ENN" provides two optional strategies. Majority class selection (kind_sel='mode '): The current sample will only be retained if all nearest neighbor samples belong to the same category as the current sample. Select All (kind_sel='all '): As long as one of the nearest neighbor samples does not match the category of the current sample, the current sample will be deleted.

2.3. MLP Classifier framework

MLPClassifier [10]is a classifier in the scikit learn library that applies backpropagation algorithms in multi-layer neural networks to train models that can be used for classification tasks.

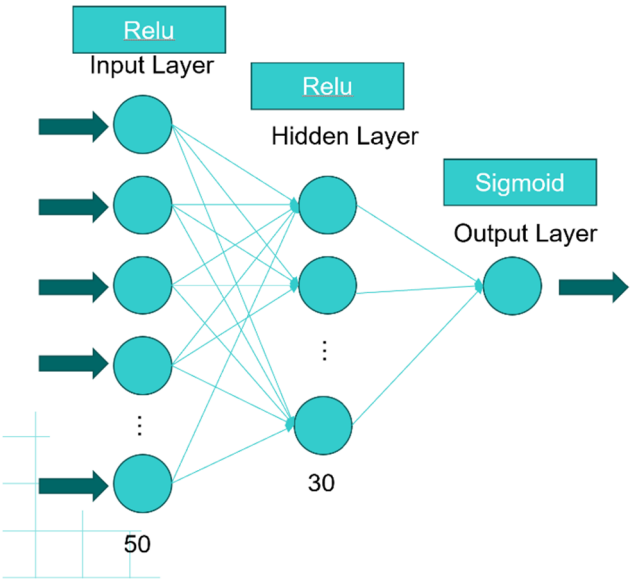

Fig 2. Framework

Multilayer perceptron (MLP), which is called a feedforward artificial neural network (FFANN) as well, is frequently applied in different areas including audio processing, natural language processing, autonomous vehicles, computer vision, etc.

The basic working principle of MLP Classifier is to classify input data, gather input data through the connections of multiple neurons to the output layer, and ultimately determine the prediction results. In neurons, activation functions are used to determine whether the input signal should activate the neuron. Fig 2 shows the whole framework.

3. Experiment

3.1. Dataset Introduction

Source of the dataset: The Kaggle data platform provides the information for this study. In detail, this survey is composed of 319795 people who are at least 18 years old. In addition, the proportion of samples without heart disease to ones with that is approximately 10.7 to 1. The distribution shows as figure 3. Initially, this dataset came from a telephone survey conducted by the Centers for Disease Control and Prevention in the United States in 2020, which collected data on the health state of American residents. In the end, this study uses it for heart disease prediction to assist professional doctors to diagnose heart disease preliminarily.

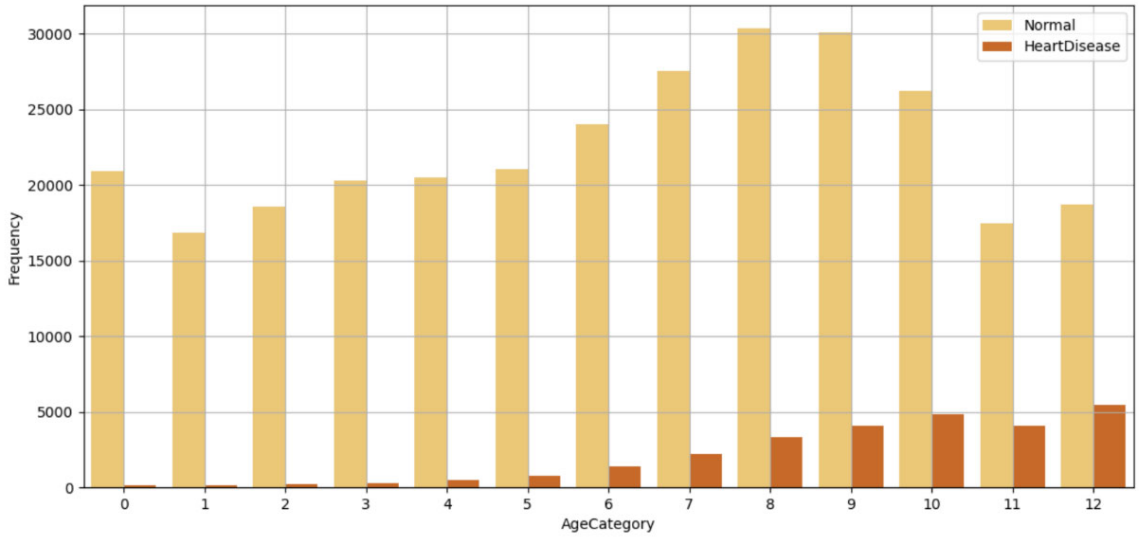

Fig 3. Distribution of Cases with Yes/No heart disease according to Age

3.2. Data preprocessing

Data preprocessing: The dataset in this study is extremely unbalanced, so the "SMOTE" algorithm as well as "ENN" algorithm is used to balance the dataset. "SMOTE", also known as "artificial minority oversampling method," generates composite samples of minority classes through interpolation, thereby increasing the number of minority class samples; ENN is a way of removing misclassified samples that entails finding misclassified items in the dataset by utilizing k=3 nearest neighbors and subsequently deleting them. Subsequently, a Pearson correlation coefficient threshold of 0.02 was selected. The filtered correlation coefficient table will only include attributes with a correlation coefficient value greater than or equal to 0.02 or less than or equal to -0.02 related to heart disease. All correlation coefficients that do not meet this threshold will be deleted, and only those that meet the criteria will be retained. This process will screen out features in the dataset that are weakly associated with heart disease, and only retain features that are strongly associated with heart disease.

3.3. Preprocessed data situation

After balancing the samples, all evaluation indicators of all models improved, with ANN, GBDT, and XGBoost performing well, but ANN showed better recall values than GBDT and XGBoost.

After balancing the samples, the samples were screened with a Pearson correlation coefficient threshold of 0.02. Except for f1 score, all models showed varying degrees of decline in accuracy, precision, recall, and auc, indicating that we screened out some important features during the screening process, so we cannot remove them.

3.4. Training and testing parameter settings for classifiers

Pre sampling training set x: 223856 samples, y: 223856 samples
After sampling, the training set x: 338279 samples, y: 338279 samples
Test set: x: 95939, y: 95939
Remote+enn sampling is independent of the test set.

In this study, we used sampled training data (x: 338279, y: 338279) as the training set, shown as Table1. Perform feature selection, processing, and transformation on the training set for use by the model. Subsequently, we select traditional machine learning models that are suitable for the problem, such as random forests, support vector machines, logistic regression, etc. Set appropriate parameters for the selected model. Train the selected model using sampled training data. After the training, we used test data (x: 95939, y: 95939) as the test set. Perform the same feature processing and transformation on the test set as the training set to ensure compatibility with the trained model. Then use the trained model to predict the test set. Based on the predicted results and real labels, calculate the evaluation indicators of the model, like accuracy, precision, recall, F1 score, etc., to assess the performance of the model on the test set.

Table 1. set samples.

	health	Heart disease	Total
Traning	148832	189447	338279
Test	87791	8148	95939
Total	236623	197595	434218

4. Experimental result

Accuracy is a commonly used indicator to assess the performance of classification models, reflecting the ratio of the number of samples which the model accurately predicted to the total number of samples.

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \tag{2}$$

F1 score is a universally employed metric to measure the performance of classification models, which combines the accuracy and recall of the model to balance the relationship between the two.

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{Precision + Recall} \tag{3}$$

Precision measures the proportion of samples predicted by the model as positive categories, which is actually positive categories. The calculation formula for accuracy is as (4).

$$\text{Precision} = \frac{\text{True Positivesl}}{\text{True Positives} + \text{False Positives}} \tag{4}$$

In the analysis of the dataset, it was found that this dataset is relatively complex, and the absolute values of the Pearson correlation coefficients between various attributes and heart disease (HeartDisease) are relatively low, making it unsuitable to use models based on linear relationship algorithms for prediction, for example, LR (logistic regression), etc. For the raw data, the degree of

imbalance is high. When directly training the machine learning model, it was found that the accuracy index is abnormally high, while other indicators are generally low, resulting in poor performance. After balancing the dataset with "SMOTE+ENN", all evaluation indicators of the models improved, with auc and recall based on ANN models reaching 0.808 and 0.81, respectively, shown as Table2. Moreover, ANN has high potential and good robustness. Therefore, it is recommended to use the ANN model as a model for predicting heart disease.

Table 2. Result

	precision	recall	f1-score	support
health	0.97	0.67	0.79	87791
Heart disease	0.18	0.81	0.30	8148
accuracy			0.68	95939
macro avg	0.58	0.74	0.55	95939
weighted avg	0.91	0.68	0.75	95939
AUC:0.808				

5. Conclusion

In the data set analysis, this study found that stroke, physical health, diffwalking, AgeCategory, and diabetes had a greater impact on heart disease prediction. In the analysis of the dataset, it was found that this dataset is relatively complex, with low Pearson correlation coefficients between various attributes and HeartDisease, making it unsuitable to use models based on linear relationship algorithms for prediction, such as LR (logistic regression). This study uses deep neural network (ANN) models to predict heart disease problems and compares them with models trained by machine learning algorithms such as logistic regression, decision tree, Xgboost, RF, GBDT, SVM, KNN. By comparing the performance of these models using five indicators: accuracy, f1 receiver, recall, precision, and auc, it is found that ANN has certain advantages over traditional learning models. This study used a Pearson correlation coefficient threshold of 0.02 to screen and found that all models, except for the f1 score, showed varying degrees of decline in accuracy, precision, recall, and auc, indicating that self-health rating and average sleep time also have important roles in predicting heart disease and cannot be discarded. The dataset on Kaggle used in this study was from the Centers for Disease Control and Prevention in the United States in 2020, which has certain regional and timeliness. Therefore, we hope to collect the latest data from people of all ethnic groups around the world in the future and achieve better performance by updating and expanding the dataset.

Authors Contribution

All the authors contributed equally, and their names were listed in alphabetical order.

References

[1] E. Bathrellou, M. D. Kontogianni, E. Chrysanthopoulou et al., “Adherence to a dash-style diet and cardiovascular disease risk: the 10-year follow-up of the Attica study,” Nutrition and Health, vol. 25, no. 3, pp. 225–230, 2019.

[2] S. Mohan, C. Thirumalai and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," in IEEE Access, vol. 7, pp. 81542-81554, 2019, doi: 10.1109/ACCESS.2019.2923707.

[3] Garg A, Sharma B, Khan R. Heart disease prediction using machine learning techniques[J]. IOP Conference Series: Materials Science and Engineering, 2021.DOI:10.1088/1757-899X/1022/1/012046.

[4] MIENYE D., SUN Y.X., WANG Z. H. Improved sparse autoencoder based artificial neural network approach for prediction of heart disease [J]. Informatics in medicine unlocked, 2020,18: 100307.

[5] Framingham Heart study dataset [Online]. Available, https://kaggle.com/amana jmera1/framingham-heart-study-dataset. [Accessed 24 January 2020].

[6] Heart Disease and Stroke Statistics—2022 Update: A Report from the American Heart Association

[7] S. K. Jonnavithula, A. K. Jha, M. Kavitha, and S. Srinivasulu, "Role of Machine Learning algorithms over heart diseases prediction," AIP Conference Proceedings, Vol. 2292, 2020, pp. 40013–40013.

[8] N. V. Chawla,K. W. Bowyer,L. O. Hall,W. P. Kegelmeye, "SMOTE: Synthetic Minority Over-sampling Technique," in Journal of Artificial Intelligence Research,vol. 16, 2002, DOI: https://doi.org/10.1613/jair.953

[9] D. L. Wilson, "Asymptotic Properties of Nearest Neighbor Rules Using Edited Data," in IEEE Transactions on Systems, Man, and Cybernetics, vol. SMC-2, no. 3, pp. 408-421, July 1972, doi: 10.1109/TSMC.1972.4309137.

[10] Windeatt, T. (2008). Ensemble MLP Classifier Design. In: Jain, L.C., Sato-Ilic, M., Virvou, M., Tsihrintzis, G.A., Balas, V.E., Abeynayake, C. (eds) Computational Intelligence Paradigms. Studies in Computational Intelligence, vol 137. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-79474-5_6