Research on Heart Disease Prediction Method based on Neural Network

Qiaoyun Xue[1,*]

[1]Smart-technology Business Department, Qingdao Hisense Hitachi Air Conditioning System Co. LTD, Qingdao, 266555, PR China
[*]Corresponding E-mail: xueqiaoyun@hisensehitachi.com

Abstract. With the emergence of the deep learning method, many tasks have benefited from this technique, with satisfying performance and reasonable robustness. Recently, the deep learning-based method has been applied to medical diagnosis. However, medical image processing and data analysis need high accuracy and an extremely low error rate, the method needs to pay more attention. In this paper, the author aims to apply a neural network to help diagnose heart disease. The author first processes the dataset and calculates the feature correlation. Then a multi-layer neural network was designed and evaluated on the public dataset. Furthermore, a K-Nearest Neighbor method is also implemented for comparison. Among the existing method, KNN and random forest, the neural network achieves the best results, which achieve 75% accuracy. This paper aims to validate the effectiveness of deep learning methods and provides a brief attempt at medical diagnosis. The author wishes the deep learning-based method can be applied to real-world medical applications as soon as possible.

Keywords: Deep learning, Medical diagnosis, Heart disease prediction.

1. Introduction

Heart disease has always been a severe factor that threatens people's life [1]. This study developed a model that reads patients' Physiological characteristics to determine which feature is the most important in causing heart disease to happen and predict if the patients get heart disease or not. The study first processes the dataset, which will remove outliers and excessive data to maintain the prediction's outcome at a balanced level. Then four distinct models were used to predict whether a patient has heart disease based on a specific scenario, Eventually, the ROC curve was used to find out which model is the most accurate in predicting patient's heart disease situation [2,3]. S. Radhimeenakshi proposed a classification system for the prediction of heart disease. The work he carried out focuses on two algorithms namely Support Vector Machine (SVM) and Decision Tree. The model achieved more accuracy when Decision Tree classification algorithm is used. The accuracy of the model was evaluated using a confusion matrix and the model achieved an accuracy of 55% for SVM.

SENTHILKUMAR MOHAN reported a novel method that aims at finding significant features by applying machine learning techniques resulting in improving the accuracy in the prediction of cardiovascular disease. The prediction model is introduced with different combinations of features and several known classification techniques. They produce an enhanced performance level with an accuracy level of 88.7% through the prediction model for heart disease with the hybrid random forest with a linear model (HRFLM) [4].

2. Method

2.1 Data preprocessing

This study starts data preprocessing by eliminating the outliers, First Q1 and Q3 are calculated by using 'np.percentile' and IQR is found by calculating Q3-Q1. Eventually, this project traverses through the entire dataset and eliminates the data that is either greater or equal to Q3+1.5* interquartile range or smaller or equal to Q1-1.5* interquartile range. The uneven distribution of this dataset severely affects the training of our model, So the author tries to remove excessive data until the data size of patients who have heart disease and patients who don't are even. The author manually removed some of the patients' data while maintaining that each race occupies a reasonable proportion of our training data, for example, some of the data were removed until their number remained roughly equal to that of other groups because the majority of the data concern white people. After trimming the excessive data, the data size has come to 40000x23, With twenty thousand patients' samples that have heart disease while the other half do not[5,6]. Furthermore, it's crucial to find the correlation between different features and whether the patient has heart disease or not. Table 1 contains the correlation data that was obtained using the 'np.corrcoef' function. Some characteristics, such as whether patients had asthma in the past, are removed since they are not strongly associated to patients' heart disease problems. Furthermore, patients' general health level and their stroke history are the two most important features in determining whether they have heart disease.

Table 1. Feature correlation on the utilized dataset.

Factors	Correlation
BMI	0.052520417
Smoking	0.108470801
AlcoholDrinking	-0.03273086
Stroke	0.197026225
Physical Health	0.169366629
MentalHealth	0.026541485
DiffWalking	0.1997398
SexMale	0.070485298
SexFemale	-0.07048529
White	0.040522771
Black	-0.010472308
Hispanic	-0.03586567
Other	-0.003292365
Asian	-0.030835228
American,Indian/AlaskanNative	0.008832572
AgeCategory	0.234251767
PhysicalActivity	-0.09848231

2.2. Heart Disease Classification Using NN

There are many types of neural networks, like Multilayer Perceptron (MLP), Convolution Neural Networks (CNN), and Recurrent Neural Networks (RNN) [7]. Multilayer Perceptron was chosen for two reasons. First off, classification analysis is required for the investigation, and MLPs are appropriate for classification prediction problems. Secondly, the dataset is a tabular dataset and MLPs are suitable for tabular datasets. Two methods were used to prevent overfitting. The first technique used is L2 regularization. Since "Reasonable values of lambda [regularization hyperparameter] range between 0 and 0.1", 0.01, 0.001, and 0.0001 were used as possible strengths of the L2 regularization. We initially select 0.01 as the strength of the L2 regularization because we cannot tell whether overfitting occurs before we train the model [8]. The second tactic is called a dropout. We can avoid overfitting by removing units from a neural network's hidden and visible layers. Hyperparameter selection and parameter grid filling manually. Using GridSearchCV to search through each

combination of parameters from the parameter grid and find out which combination of hyperparameters has the best performance. This project chooses 2-4 layers as the number of hidden layers because during training the author found that 2-4 layers have better performance than a single layer. For the number of neurons in each hidden layer, Heaton's rules of thumb were followed: “The number of hidden neurons should be between the size of the input layer and the size of the output layer. The number of hidden neurons should be 2/3 the size of the input layer, plus the size of the output layer. The number of hidden neurons should be less than twice the size of the input layer”.

Based on the above rules, (22,22), (15,15), (40,40), (22,14,4), (15,10,3), (40,26,177), (22,14,4,2), (40,26,17,12), (22,22,22,22), (40,40,40,40) were chosen as possible hidden layer sizes. This study chooses ReLU for the activation function because this paper is using an MLP. A constant learning rate is used and values are selected based on the below rules. “typical values for a neural network with standardized inputs (or inputs mapped to the (0,1) interval) are less than 1 and greater than 10^−6”(Bengio, 2012). “default value of 0.01 typically works for standard multi-layer neural networks but it would be foolish to rely exclusively on this default value”(Bengio, 2012). 0.1, 0.01, 0.001 and 0.0001 were selected as the possible learning rates. Adam and SGD were chosen as potential optimizers. Since Adam is the best choice in most cases and “SGD can bring better results if combined with a good learning rate”(Giordano, 2022). Finally, this study finds hyperparameters with the best performance by using GridSearchCV. The study found that the size of (22,22,22,22), the learning rate of 0.01, and the optimizer of Adam have the best performance. From the classification report, we can know that the accuracy for the training set is 75.46% and the accuracy for the test set is 75.26%. The accuracy doesn’t vary a lot in the different datasets and the variance is low. It means that our model has similar performance on different datasets and overfitting doesn’t occur.

2.3. Heart Disease Classification Using KNN

Through this part, kNN was used to train machine learning models to best predict cases of cardiac problems. The k-Nearest-Neighbours (kNN) is a simple but effective method for classification. It is used for classification and regression. The input consists of the k closest training examples in a data set. The output depends on whether k-NN is used for classification or regression. In k-NN classification, the output is a class membership. An object is classified by a plurality vote of its neighbours. The object is assigned to the class most common among its k nearest neighbours (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbour. In k-NN regression, the output is the property value for the object. This value is the average of the values of k's nearest neighbours. Our data set is binary classification, and kNN is an effective method for classification. And kNN is also suitable for pure numerical data. So this part chooses kNN model [9].

The process can be summarized as follow: Load data and Statistical exploration: The findings indicate that all 22 variables, with the exception of BMI, are integers. Also, the data collection does not contain a null value. Preprocessing before modelling: First, this study separates the data set into an X input set and a y target set. And then the data was divided into train and test groups in a 9:1 ratio. Making certain that the random number is the same each time is important. In order to achieve this, the random state was set to a fixed value. In order to make all the data sets in a system homogeneous in terms of content and format, the data set was finally standardized. Overfitting can be prevented by transforming the data into a standard normal distribution using the theorem of large numbers**.** Choose the best hyperparameters for the model: The most important factor for the kNN model is the hyperparameters which include K, Metric, and Weights. The model will go through cross-validation to define the best hyperparameters. The “GridSearchCV” (Grid search and cross-validation) was used to adjust the parameters. In this model, the author makes the weights equal to uniforms which means do not include weight influence in our model. And if weights are equal to the distance, the distance weight will be taken into account. And in order to increase efficiency, n_jobs was set equal to -1 which means the number of jobs is set to the number of CPU cores. This study tries the k from 1 to 50 and tries the metrics in three types of metric. And cross-validation parameters using 10-fold cross-

validation. And then fit the model using the train data. The results are that the best k is 33 and the best metric is Minkowski in this model. Training and Performance Measurement [10]: The best hyperparameters are used to train the model and test the model using the test data. The results show that the model's accuracy is 0.74 and its AUC is 0.81.

3. Results and Discussion

With kNN, NN these four models. Also, we introduce two existing methods, including SVM, Random Forest. ROC curves were implemented to determine which model has the highest performance. The true positive rate lies on the y-axis and the false positive rate lies on the x-axis of the ROC curve. Plotting ROC curves is an effective way in finding the most accurate model since this project is training a binary classification model. First, this study plots the ROC curve with the hyperparameters and then ranks each model's performance based on its area under the curve, which is the rate of accuracy. As shown in Table 2, The performance of our model ranks like this: 1) NN, 2) KNN, 3) Random Forest, and 4)Support Vector Machine. The difference in the performance of our models is not big. Overall, the most accurate model for predicting heart disease seems to be a neural network which is not surprising. As shown in Table 2.

Table 2. Comparison results among the existing methods.

Method	Accuracy
Neural network	0.75
K-nearest neighbor	0.74
Random forest	0.73
Support Vector Machine	0.70

4. Conclusion

In this project, the author successfully created four models for predicting whether a patient has heart disease or not based on machine learning methods. The study used Neural Network, kNN as our prediction model and found that the neural network is the most accurate model in predicting the patient's heart disease condition. Also, among these different features in the dataset, General health level and history of stroke have great importance in determining a patient's heart disease condition. This research may have a great impact on heart disease prevention. These models can be implemented in health detection apps that Both medical institutions and regular patients can use to check whether their patients (or themselves) have heart disease. People can also use the models to prevent heart disease by comparing their own physical conditions with the features this study provides. With more medical knowledge, this model can be more explicit in determining the specific kind of heart disease a patient is having given enough information about. From then on, our model will become a multiclass classification and greatly reduces the time it takes to predict a patient's heart disease condition.

References

[1] Ali L, Bukhari S A C. An approach based on mutually informed neural networks to optimize the generalization capabilities of decision support systems developed for heart failure prediction[J]. Irbm, 2021, 42(5): 345-352.

[2] Ortiz J, Ghefter C G M, Silva C E S, et al. One-year mortality prognosis in heart failure: a neural network approach based on echocardiographic data[J]. Journal of the American College of Cardiology, 1995, 26(7): 1586-1593.

[3] Mienye I D, Sun Y, Wang Z. Improved sparse autoencoder based artificial neural network approach for prediction of heart disease[J]. Informatics in Medicine Unlocked, 2020, 18: 100307.

[4] ÇELİK G. Prediction of Heart Failure Disease with a Proposed Method Based on Deep Neural Networks[J]. Ankara/Turkey, 2022: 279.

[5] Zhang D, Chen Y, Chen Y, et al. An ECG heartbeat classification method based on deep

convolutional neural network[J]. Journal of Healthcare Engineering, 2021, 2021: 1-9.
[6] Ali L, Rahman A, Khan A, et al. An automated diagnostic system for heart disease prediction based on statistical model and optimally configured deep neural network[J]. Ieee Access, 2019, 7: 34938-34945.
[7] Zhu X, Xu H, Zhao Z, et al. An Environmental Intrusion Detection Technology Based on WiFi[J]. Wireless Personal Communications, 2021, 119(2): 1425-1436.
[8] Khan M U, Samer S, Alshehri M D, et al. Artificial neural network-based cardiovascular disease prediction using spectral features[J]. Computers and Electrical Engineering, 2022, 101: 108094.
[9] Waghulde N P, Patil N P. Genetic neural approach for heart disease prediction[J]. International Journal of Advanced Computer Research, 2014, 4(3): 778.
[10] Chitra R, Seenivasagam V. Review of heart disease prediction system using data mining and hybrid intelligent techniques[J]. ICTACT journal on soft computing, 2013, 3(04): 605-09.