Proceedings of the 2023 International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/8/20230102

Exploiting neural network for heart disease probability prediction

Songwei Liu[1, †], **Zhonghui Miao**[2, †], **Anqiao Zhang**[3, 4, †]

[1]School of Mathematics, Hunan University, Changsha, Hunan, 410082, China
[2]Liangjiang International College, Chongqing University of Technology, Chongqing, 401135, China
[3]Faculty of Sciences, Engineering and Technology, University of Adelaide, Adelaide, 5000, Australia

[4]anqiao.zhang@student.adelaide.edu.au
[†]These authors contributed equally

Abstract. In general, doctors determine the presence of heart disease through clinical evaluation and pathological data, and the diagnosis process is complex and inefficient. Based on the above situation, professionals are committed to researching efficient and accurate methods for predicting heart disease. After studying many literatures, this paper found that the existing heart disease prediction system has high requirements for clinical data. Based on the reality of the shortage of medical resources under the COVID-19 epidemic, this paper develops a simple heart disease prediction system, which predicts heart disease through simple and easy-to-measure data of patients, and then prevents heart disease. The method consists of two steps. First, collect the characteristics related to heart disease, and then select the most important 10 characteristics through correlation analysis and literature research, namely gender, age range, body mass index (BMI), smoking status, physical health index, walking difficulty, stroke status, skin cancer, diabetes, kidney disease. Second, an algorithm for heart disease based on artificial neural networks classification based on these features is developed. The prediction accuracy is close to 92%. In the future, the proposed model could be leveraged for heart disease recognition.

Keywords: heart disease, artificial intelligence, machine learning, neural network.

1. Introduction

In general, diagnosing whether a patient suffers from heart disease relies on a complex combination of pathological data and clinical findings. This intricate nature results in excessive healthcare expenses and strain on medical resources [1, 2]. This situation is also exacerbated under the influence of the COVID-19 epidemic. According to statistics from the World Health Organization, the world suffers from heart disease, and in 2019 one-third of the world's population died of heart disease [3, 4]. Recent years, the development of data science has promoted the research on heart disease prediction. In a large amount of clinical data, finding the undetected health information among healthy individuals and heart disease individuals is an effective method to predict heart disease. Statistics and machine learning are the two primary methodologies to predict cardiac status based on the representation of clinical data. Artificial neural network became a highly researched area within the realm of artificial intelligence around the 1980s. It draws parallels between the human brain's neural network and information processing, constructs networks with varying connections to form distinct structures, and develops a straightforward model [5]. Artificial neural networks (ANN) have high precision and high learning rates, making them worth a try as an algorithm for predicting heart disease [6, 7]. Existing prediction models have high requirements for clinical data. Based on the reality of the shortage of medical resources under the COVID-19 epidemic, it is necessary to reduce clinical data to reduce the pressure on medical resources and predict heart disease with high accuracy [8, 9]. In this study, a neural network is proposed for forecasting the condition of heart disease. In the future, it is expected to contribute to a more effective and accurate diagnostic system by leveraging a small number of simple and easy-to-measure features.

2. Method

2.1. Dataset

The data utilized in this project originates from the CDC, also known as the Center for Disease Control and Prevention, which contains 319,795 data [10]. The CDC survey uses a telephone questionnaire to collect information on the health status of US residents. The data from the CDC used in this project are based on surveys from 2020 (inclusive) to February 15, 2022, with the majority of data on basic physiological conditions of respondents, such as alcohol consumption and age. The survey included African American, American Indian, Alaska Native and white people who has answered the telephone in all 50 states, the District of Columbia, and three U.S. territories. In summary, the CDC data set can be considered a representative sample.

Table 1 shows the 10 clinical features and their descriptions.

Table 1. Clinical features and their descriptions.

Feature	Description
Gender	Gender of the person.
Age category	13 stages of age category: 18~24, 25~29, 30~34, 35~39, 40~44, 45~49, 50~54, 55~59, 60~64, 65~69, 70~74, 75~79, 80 and above
Body Mass Index (BMI)	Weight(kg) / Height(m)^2
Smoking	If the person smoked at least 100 cigarettes in his/her entire life.
Physical health	The number of days that the physical health of the person is bad.
Walking difficulty	If the person has difficulty in walking or climbing stairs.
Stroke	If the person ever had a stroke.
Skin cancer	If the person ever had skin cancer.
Diabetes	If the person has diabetes.
Kidney disease	If the person ever suffered from kidney disease (exclude urinary incontinence, kidney stones, or bladder infection)

2.2. Artificial neural network

This project uses Python and its toolkit to predict and classify heart disease. In this study, the team chooses the Deep Neural Network (DNN) algorithm. The DNN is a feature-evolving learning algorithm. The shallow neurons directly learn basic low-level features from the input data, such as edge and texture features, etc.; the deep features will build upon the learned shallow features and continue to learn more sophisticated features, acquiring in-depth semantic information as viewed by a computer. More hidden layers could increase the number of non-linear transformations brought about by the activation function, which enables the algorithm to construct more complex mapping relationships.

The composition of a DNN includes three components, including input, output, and hidden layers. The input one feeds the data into the neural network. Every hidden layer consists of certain neurons, which are critical to the fitting function. The final output is produced by the output layer. Given k-1 layer and the k layer of the neural network, neurons in k-1 layers are connected to all of them in the k layer, that is, each neuron in the k layer is calculated by weighting and summing the neurons in the k-1 layer. Let the vector $\mathrm{L}_{\mathrm{k}-1} = [\mathrm{l}_{\mathrm{k}-1,1}, \mathrm{l}_{\mathrm{k}-1,2}, \ldots, \mathrm{l}_{\mathrm{k}-1,\mathrm{Nk}}]$representing the output of layer k-1 in the DNN be designated as, so the output of a neuron in layer k could be denoted as $\mathrm{l}_{\mathrm{k,i}} = (\mathrm{W}_{\mathrm{k}}(\mathrm{i}))^{\mathrm{T}}\mathrm{l}_{\mathrm{k}-1,\mathrm{i}} + \mathrm{b}_{\mathrm{k}}(\mathrm{i}), \mathrm{i} = 1, 2, \ldots, \mathrm{N}_{\mathrm{k}}$, where the weight matrix and bias vector of the kth hidden layer are represented by Wk and bk, respectively. Therefore, the corresponding output is $\mathrm{l}_{\mathrm{k}} = \mathrm{W}_{\mathrm{k}}\mathrm{L}_{\mathrm{k}-1} + \mathrm{b}_{\mathrm{k}}$ is activated by ReLU, which could be denoted as:

$$f(x) = \begin{cases} x & if\ x > 0 \\ \lambda x & if\ x \leq 0 \end{cases} \tag{1}$$

As a result, the output produced by the entire model is a non-linear mapping from the input space, which could be denoted as $\mathrm{y} = \mathrm{f}(\mathrm{x}, \theta) = \mathrm{f}_{\mathrm{L}-1}(\mathrm{W}_{\mathrm{L}-1}\mathrm{f}_{\mathrm{L}-2}(\ldots \mathrm{f}_1(\mathrm{W}_1\mathrm{x} + \mathrm{b}_1)) + \mathrm{b}_{\mathrm{L}-1})$, where L denotes the number of layers, and θ stands for all learnable parameters. Model parameters are made up of weights and biases in all nodes. They are randomly initialized and iteratively refined during the learning process.

The neural network designed in this project is shown in Figure 1. The input layer is made up of 10 neurons, which has the same dimension with the features. It is followed by a hidden layer with 50 nodes. The second hidden layer consists of 30 nodes. It shrinks the features learned from the previous level to make them more descriptive. The output layer has two nodes, denotes the likelihood that a patient suffers from heart disease.

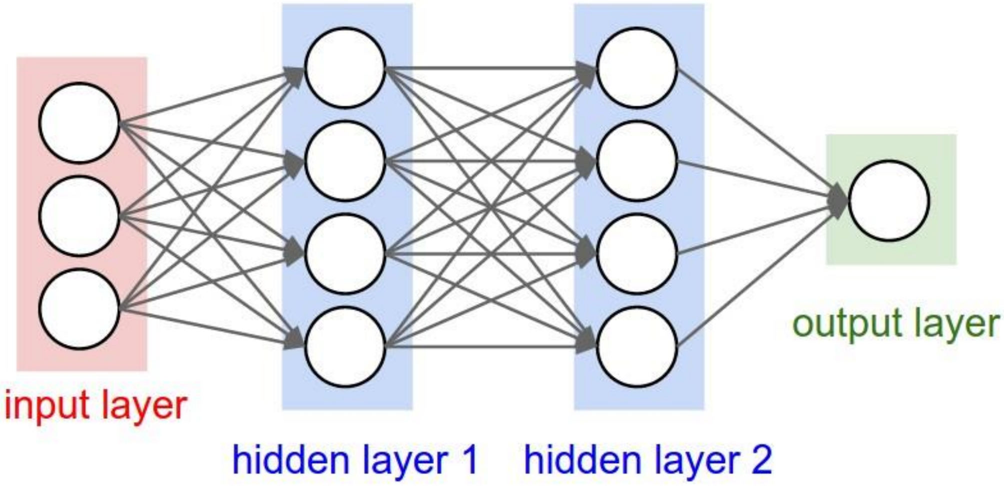

Figure 1. Architecture of the neural network.

2.3. Logistic regression

Logistic regression (LR) is widely leveraged in medical research, and it has its own origin of statistics. Logistic regression is a discriminative model with form P(y|x) = f(x, α), where the f is function and its corresponding parameters α. The α is usually determined by maximum-likelihood estimation, which is on the basis of the data set D. The functional form of f is considered a parametric algorithm, and the parameters (coefficients and intercept) used in f generally have interpretable contributions that can be described in plain language.

2.4. Random forest regression

Random Forest Regression is an ensemble learning method. It conducts regression calculations under the supervision of ground truth labels. The ensemble technique produces prediction results by combining the outputs from several machine learning algorithms, which is comparably more accurate rather than that of the output of one single model.

The prediction for a random forest algorithm is obtained by computing the mean of the predictions from multiple decision trees, which are constructed during the training phase. It usually has good performance on many aspects, featuring non-linear relationships among variables. However, it has shortages as the following: non-interpreted, easily over-fitted, etc.

2.5. Evaluation matrix

The performance of the neural network algorithm implemented in this case will be assessed in three respects.

Accuracy measures the overall correctness of the model's predictions. The accuracy measures the percentage of the number of correctly recognized samples over that of all samples. Mathematically, it is:

$$\text{Accuracy} = (\text{True Positives} + \text{True Negatives}) / \text{Total Predictions} \tag{2}$$

The recall rate can be defined as:

$$\text{Recall} = \text{True Positives} / (\text{True Positives} + \text{False Negatives}) \tag{3}$$

The precision can be written as:

$$\text{Precision} = \text{True Positives} / (\text{True Positives} + \text{False Positives}) \tag{4}$$

3. Result

It turns out that the method proposed in this paper for predicting heart disease was reasonable and Table 2 summarises the results generated by the aforementioned artificial neural network algorithm.

Table 2. Classification of heart disease performance calculated from 10 characteristics.

Accuracy(%)	Recall(%)	Precision(%)
91.65%	91.71%	91.59%

The classification has a 91.65% correct rate, a 91.59% accuracy rate, and a 91.71% recall rate.

In addition, there are other models, shown in Table 3, used in this paper for performance comparison.

Table 3. Categorical performance of heart disease calculated by different models.

Model / Evaluation matrix	Accuracy(%)	Recall(%)	Precision(%)
Deep Neural Network	91.65%	91.71%	91.59%
Logistics Regression	91.56%	91.56%	92.15%
Random Forest	88.89%	95.52%	92.57%

This project also incorporates a loss function to test the performance of the designed and performed ANN algorithm. The loss is calculated by Mean Square Error (MSE). It can be discovered that Figure 2 shows the different predicted outputs calculated by MSE loss function.

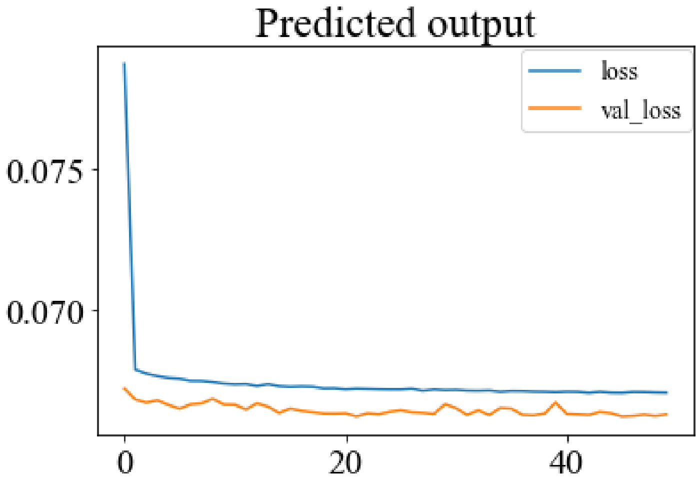

Figure 2. Loss functions for the training and test sets.

In comparison with the logistic regression algorithm, the deep neural network has a better overall performance in accuracy rate and regression rate. The correct rate of the random forest algorithm is less than 90%, and its performance is slightly inferior to that of DNN. In summary, the performance of the DNN algorithm is good, and it has a high accuracy rate for classifying and predicting heart disease.

In accordance with the results of data analysis, people should control their weight reasonably and reduce smoking in terms of preventing heart disease. As people grow older, it is of great necessity to strengthen the protection of the body and pay more attention to the impact of other diseases on heart disease.

4. Conclusion

This project completes an algorithm for an artificial neural network for predicting heart disease, which can be leveraged to predict heart disease from easily accessible data and thus prevent it. The prediction method consists of two steps, starting with the selection of 10 important characteristics, namely gender, age category, body mass index (BMI), smoking status, body mass index, walking difficulties, stroke status, skin cancer, diabetes, and kidney disease. After the data analysis on the data set used in this study, a concise conclusion can be drawn that one of the most efficient methods to prevent heart disease is to manage to lead a balanced life. The activities that mentioned as follows can positively help with the issues of heart disease prevention: to adopt a healthy diet with more vegetables, to take active part in physical exercise regularly, to reduce intake of carcinogens, such as nicotine and alcohol. What a healthy diet is can be simply described as the diet that is rich in vitamins, proteins, and whole grains but low in salt or bad quality fats. Various kinds of fresh fruits and vegetables can not only make the meals on our tables richer, but also provides different vitamins that are necessary for the daily metabolism in human body. Regular physical activities, such as brisk walking, jogging, or climbing, can help to keep health for cardiovascular condition, and significantly make the risk of heart disease reduce. Smoking cigarettes and exposure to second-hand smoke is a common risks factors for heart disease, and thus in order to reduce the damage from this, it is suggested that to quit smoking and to avoid the second-hand smoke can play a positive role in heart disease prevention. Besides, diseases such as high blood pressure, high cholesterol, and diabetes should also be given considerable attention from the perspective of preventing heart disease. In this project, the ANN algorithm that is developed and performed for heart disease classification, in the basis of the characteristics mentioned above, has an accuracy rate of approximately 92% for heart disease prediction. These data are simple and easy to obtain, and do not require the use of specialist equipment, instruments, or the guidance of a doctor. Thus, the method is useful in relieving the pressure on medical resources and easing the potential conflicts between doctors and patients in the light of the strain on medical resources caused by the Covid-19 epidemic. However, since the data used in this project comes entirely from the CDC, where though the populations involved are diverse, errors may occur when the algorithm is applied in other countries data set, and further improvement is needed in this case.

References

[1] Chen, A. H., Huang, S. Y., Hong, P. S., Cheng, C. H., & Lin, E. J. (2011). HDPS: Heart disease prediction system. In 2011 computing in Cardiology, 557-560.

[2] Loizeau, V., Kilpatrick, K., Bertrand, D. P., & Rothan-Tondeur, M. (2023). Exploring Strategies for Developing Enabling Environments for People with Chronic Heart Disease: An Ethnographic Study Protocol. International Journal of Environmental Research and Public Health, 20(3), 2680.

[3] Nowbar, A. N., Gitto, M., Howard, J. P., Francis, D. P., & Al-Lamee, R. (2019). Mortality from ischemic heart disease: Analysis of data from the World Health Organization and coronary artery disease risk factors From NCD Risk Factor Collaboration. Circulation: cardiovascular quality and outcomes, 12(6), e005375.

[4] Finegold, J. A., Asaria, P., & Francis, D. P. (2013). Mortality from ischaemic heart disease by country, region, and age: statistics from World Health Organisation and United Nations. International journal of cardiology, 168(2), 934-945.

[5] Krogh, A. (2008). What are artificial neural networks? Nature biotechnology, 26(2), 195-197.

[6] Palaniappan, S., & Awang, R. (2008). Intelligent heart disease prediction system using data mining techniques. In 2008 IEEE/ACS international conference on computer systems and applications, 108-115.

[7] Soni, J., Ansari, U., Sharma, D.,& Soni, S. (2011). Predictive data mining for medical diagnosis An overview of heart disease prediction. International Journal of Computer Applications, 17(8), 43-48.

[8] Xie, L., Yang, H., Zheng, X., Wu, Y., Lin, X., & Shen, Z. (2021). Medical resources and coronavirus disease (COVID-19) mortality rate: Evidence and implications from Hubei province in China. PLoS One, 16(1), e0244867.

[9] Alhalaseh, Y. N., Elshabrawy, H. A., Erashdi, M., Shahait, M., Abu-Humdan, A. M., & Al-Hussaini, M. (2021). Allocation of the “already” limited medical resources amid the COVID-19 Pandemic, an iterative ethical encounter including suggested solutions from a real-life encounter. Frontiers in medicine, 7, 616277.

[10] Kamil, P. (2022). Personal Key Indicators of Heart Disease, URL: https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease