The 4th International Conference on Computing and Data Science (CONF-CDS 2022)
DOI: 10.54254/2755-2721/2/20220665

Predicting Heart Disease Based on Wide and Deep Neural Network

Borui Pan

Southern Methodist University, Dallas, Texas, USA

boruip@smu.edu

Abstract. Nowadays, more and more people suffer from heart disease because of stressful life, irregular diet, lack of exercise and other reasons. The population affected by heart disease is also younger than ever. If heart disease can be diagnosed as early as possible, it will be of great help in the treatment of heart disease. Thus, this paper proposes models to predict heart disease based on wide and deep neural network, and the result shows that the current work has maintain good performance. Analysis is also provided in this paper to state factors that can affect performance.

Keywords: heart disease, wide and deep neural network, crossed feature, deep neural network.

1. Introduction

Heart disease has always been a difficult problem for people to solve and try to prevent. Nowadays, due to the increasingly fierce competition in society, young people suffer more and more pressure, and the population of heart disease is becoming younger and younger. Heart disease, or cardiovascular disease, has been one of the leading causes of death worldwide in the last decade. Early and accurate diagnosis of heart disease according to these factors is an important part of the prevention or treatment of heart disease.

There have been many studies using traditional machine learning methods to predict heart disease, like using Naive Bayes, Decision Tree, Random Forest, and K Nearest Neighbor. Some neural network methods have also been taken into heart disease prediction, like Artificial Neural Network, Deep Neural Network and Convolution Neural Network. However, studies about predicting heart disease using Wide and Deep Neural Network are rare.

Wide and Deep Neural Network was first introduced by Cheng et al., which was applied in Google Play recommender system [1]. Wide and Deep network has been proved to have good performance in classification problems in many studies. In this study, Wide and Deep network is used to do the prediction of heart disease.

2. Related works

Machine learning methods have been widely applied in medical field, and there is a lot of work on heart disease using machine learning-related methods. Dwivedi (2016) compared the prediction of six different machine learning methods on heart disease dataset [2]. Artificial neural network, support vector machine method for classification, Naive Bayes classifier, logistic regression classifier, k-nearest neighbor and classification trees were concluded in this work. Logistic regression outperformed over other machine learning techniques with 85 % accuracy on StatLog heart disease dataset. Gandhi et al. (2015) discussed the advantages and disadvantages for different classifiers to make the prediction on cardiovascular disease (CVD), and decision trees, neural networks and Naive Bayes Classifier were taken into discussion [3]. Based on the discussion they proposed how to obtain the hidden patterns for making decision in healthcare organizations using data mining techniques. Kumar et al. (2018) selected four popular machine learning methods and measured the performance on a heart disease dataset from UCI machine learning repository [4]. Naive Bayes had the best performance of 83.45 % accuracy compared to SVM, decision tree and k-nearest neighbor. Shah et al. (2020) applied four machine learning classification techniques, Naive Bayes, decision tree, random forest and k-nearest neighbor on the heart disease dataset from UCI machine learning repository to predict the diagnosis of heart disease status [5]. The result showed that the k-nearest neighbors obtained the best highest accuracy of 90.789 % when k is 7. Katarya et al. (2021) did a comparative study and analysis in heart disease among logistic regression, Naive Bayes, SVM, k-nearest neighbors, decision tree, random forest, ANN, DNN and MLP [6]. These algorithms were applied on a dataset with 76 columns from UCI repository, but they choose only 14 columns in experiment. Random forest obtained the best performance with 95.60 % accuracy.

Wide and deep network is a neural network with great performance, and it is widely used in recommendation algorithms and classification problems. Wide and deep neural network was first introduced by Cheng et al [1]. (2016) in a recommender system. The wide component enables the network better to memorize sparse feature interactions by using cross-product, and the deep component in the network could help with generalization, which means this model can provide both memorization and generalization. Du et al. (2018) proposed a wide and deep learning model and compared its performance with other machine learning models [7]. The comparison in this work showed that the proposed wide and deep learning model was superior to logistic regression, random forests, SVM and deep neural network with accuracy of 85.74 % and deep neural network provided the second highest accuracy of 81.42 %. Imran et al. (2018) applied Wide and Deep Learning in classification of Chronic Kidney disease [8]. CKD dataset from UCI Machine Learning Repository was used in this research, which contains 24 features. In this research, Logistic Regression and Feedforward Neural Network were used to compare with Wide and Deep Learning. In the comparison, Feedforward Neural Network provided the best performance with 99 % AUC, 99 % F1-Score, 0.97 % Precision and 99 % Recall. Wide and Deep Learning provided very close scores to Feedforward Neural Network, however logistic regression provided the lowest scores. Jais et al. (2019) compared the performance of Wide and Deep Neural Network with and without Adam optimization algorithm [9]. Breast Cancer Wisconsin dataset from UCI Machine Learning Repository was adopted in the experiment. The result showed that Wide

Age: age of the patient [years]
Sex: sex of the patient [M: Male, F: Female]
ChestPainType: chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
RestingBP: resting blood pressure [mm Hg]
Cholesterol: serum cholesterol [mm/dl]
FastingBS: fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]
RestingECG: resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]
MaxHR: maximum heart rate achieved [Numeric value between 60 and 202]
ExerciseAngina: exercise-induced angina [Y: Yes, N: No]
Oldpeak: oldpeak = ST [Numeric value measured in depression]
ST_Slope: the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]
HeartDisease: output class [1: heart disease, 0: Normal]

Figure 1. Heart Failure Prediction Dataset Features.

DOI: 10.54254/2755-2721/2/20220665

and Deep neural network with Adam optimization provided the better performance with accuracy of 99.1 % and that without Adam optimization provided accuracy of 94.7 %.

3. Methodology and Results

3.1 Dataset

The dataset used in this work is Heart Failure Prediction Dataset which is collected from Kaggle. This dataset consists of five different independent datasets which have never been combined before. It makes the Heart Failure Prediction Dataset contain more available data for research purposes and there are 11 features and 918 instances in this dataset. The following figure shows each feature in detail [10].

3.2 Data Preprocessing

Data preprocessing is an important part to clean the dataset and improve the performance of experiment using data mining and machine learning methods. In this work, there is no duplicated and missing data in the dataset, so that there is no need to extra data cleaning in the preprocessing.

For the data to be better used on the designed network, numeric features and categorical features are preprocessed separately. Numeric features are standardized to restrict the preprocessed data to a certain range to eliminate the influence caused by outliers. Categorical features are encoded as integers for the embedding.

3.3 Wide and Deep Neural Network

Wide Component. The wide component is a linear model:

$$y = w^T x + b \tag{1}$$

y is the prediction of the model, x is the input and $x = [x_1, x_2, x_3, \dots, x_d]$ where x is a vector of features and d is the number of features, w is the parameters of the model and w = [w1, w2, w3, ..., wd] and b is the bias of the model.

In this work, the inputs of this model are the crossed features, and the crossed features are come from the categorical features. For example, after the categorical features are encoded as integers, the column Sex is [0, 1] where 0 represents male and 1 represents female, and the same as the column ExerciseAngina, 0 represents No and 1 represents Yes. If Sex and ExerciseAngina are going to be crossed, the crossed feature is Sex_ExerciseAngina and the value is [0, 1, 2, 3]. To be specific, (Sex = 0, ExerciseAngina = 0) is 0, (Sex = 1, ExerciseAngina = 0) is 1, (Sex = 0, ExerciseAngina = 1) is 2, and (Sex = 1, ExerciseAngina = 1) is 3.

Deep Component. The deep component is a deep neural network:

$$a^{l+1} = f(W^l a^l + b^l) \tag{2}$$

a is the activations of the layer, l is the number of the layer, w is the weights of the layer and b is the bias. f is the activation function, ReLU function is adopted as the activation function in this work. In this work, the input of the deep component is the numeric features and encoded categorical features.

3.4 Experiment and Result

The experiment uses three models on the Heart Failure Prediction Dataset and compares the performance of these models. The first model is a Wide and Deep Neural Network (WDNN) as the following figure.

DOI: 10.54254/2755-2721/2/20220665

There are eight crossed features as the input of the wide component, [['Sex','ChestPainType'], ['Sex','RestingECG'], ['Sex','ExerciseAngina'], ['Sex','ST_Slope'], ['ExerciseAngina','ChestPainType'], ['ExerciseAngina','ST_Slope'], ['ExerciseAngina','ChestPainType','ST_Slope'], and ['RestingECG','ST_Slope']]. And there are three layers in the deep component with 100 units in the first layer, 50 units in the second layer and 25 units in the third layer.

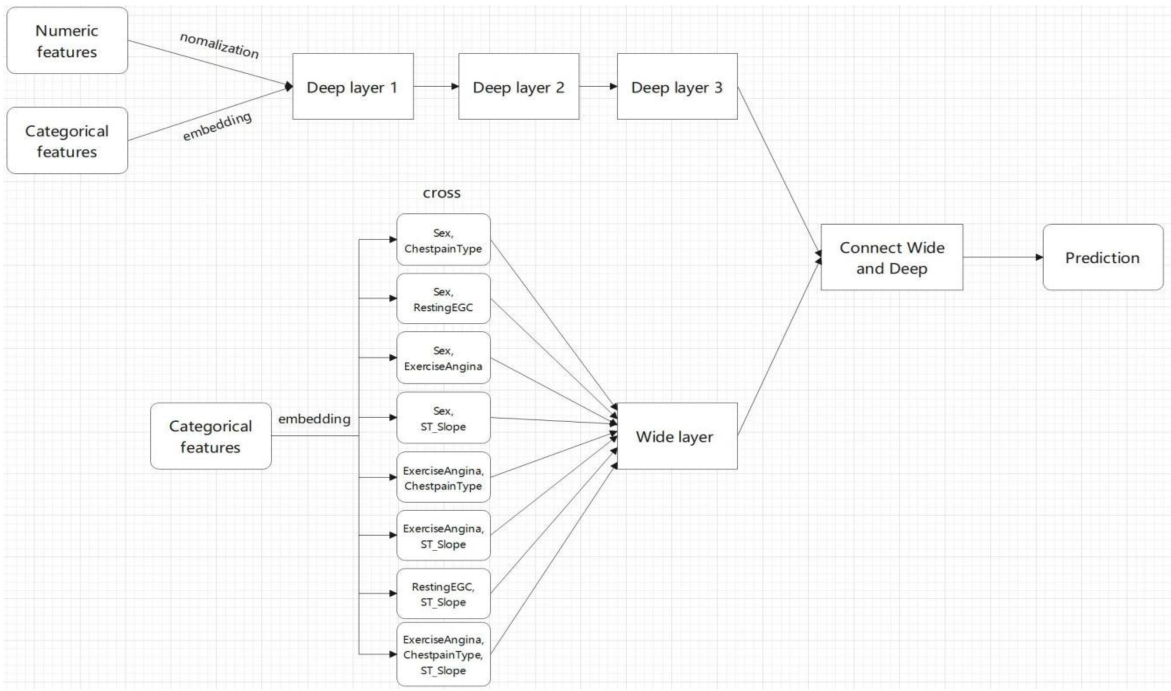

Figure. 2. Architecture of WDNN.

The second model is a no-cross Wide and Deep Neural Network (NCWDNN) and it has the same deep component with the first model. For the wide component, the categorical features without cross are the input. In this work, they are [[‘Sex’], [‘ChestPainType’], [‘RestingECG’], [‘ExerciseAngina’], [‘ST_Slope’]]. The third model is a Deep Neural Network which is the only deep component of the proposed Wide and Deep Network. The following figures show the second and the third model.

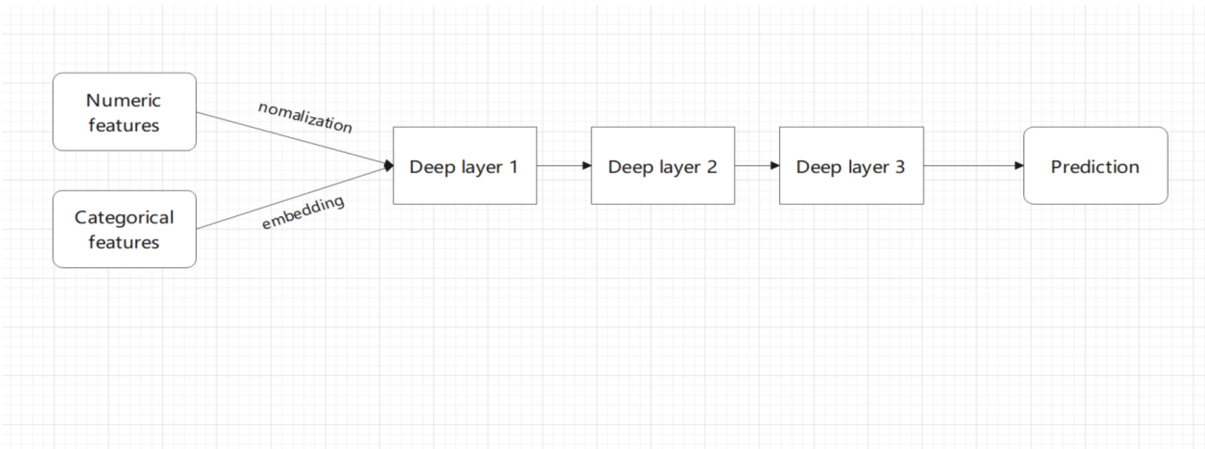

Figure 3. Architecture of DNN.

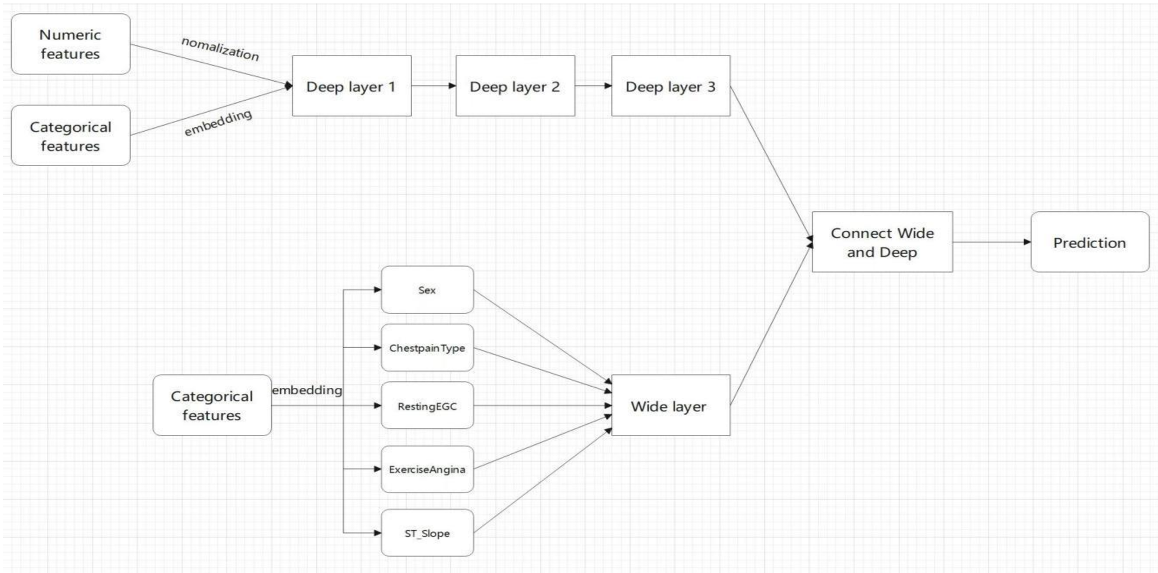

Figure 4. Architecture of NCWDNN.

As the following table shows, this work has tested and compared the performance of these different three models based on accuracy. WDNN obtained the best performance with accuracy of 84.24 %. NCWDNN got the better performance than DNN with accuracy of 82.61 %, and DNN obtained lowest performance with accuracy of 79.35 %.

Table 1. Comparison of three models.

	DNN	NCWDNN	WDNN
Deep component	Yes	Yes	Yes
Wide component	\	Yes	Yes
Crossed features in Wide component	\	\	Yes
Accuracy	79.35 %	82.61 %	84.24 %

4. Discussion

The experiment estimated the performance of three models by predicting heart disease using the 11 features in the dataset. As the result shows, NCWDNN got accuracy of 82.61 % which is better than DNN did with the accuracy of 79.35 %. The two model have the same deep component with three layers and the same parameters in the deep component. The difference is that NCWDNN has a wide component with the input of embedded categorical features which is the same with that of deep component. Thus, this indicates that these categorical features may be the more influential features in this dataset. In another hand, WDNN with both deep component and wide component got the best performance, and it has the same deep component structure with the other two models. The difference between WDNN and NCWDNN is that I chose eight different combinations from the embedded categorical features which are more relevant to do the cross. For example, chest pain type might have relation with exercise-induced angina, so I crossed [‘ChestPainType’] and [‘ExerciseAngina’]. And the result turned out that the crossed features enable the model to get better performance than NCWDNN. The performance might be better if particular algorithms used here to explore the correlation between features and cross-features are done based on the correlation.

5. Conclusion

Heart disease is a difficult problem to solve, and it needs to be discovered as soon as possible. Machine learning method can aid in heart disease diagnosis and research. Thus, this paper proposed three models which are related to wide and deep learning to make prediction on a heart disease dataset, compared the performance and discussed the difference of structure among these three models. The result demonstrates that WDNN outperformed DNN and NCWDNN with accuracy of 84.24 %, and it also shows that WDNN can contribute to the process in heart disease discovering with both nonlinear network component and linear network component. The future work can be focused on both wide component and deep component. On the one hand, the different combination of crossed features in the wide component can be study, on the other hand, different structure in the deep component can also lead to better performance.

References

[1] Cheng, Heng-Tze, et al. "Wide & deep learning for recommender systems." Proceedings of the 1st workshop on deep learning for recommender systems. 2016.
[2] Dwivedi, Ashok Kumar. "Performance evaluation of different machine learning techniques for prediction of heart disease." Neural Computing and Applications 29.10 (2018): 685-693.
[3] Gandhi, Monika, and Shailendra Narayan Singh. "Predictions in heart disease using techniques of data mining." 2015 International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE). IEEE, 2015.
[4] Kumar, M. Nikhil, K. V. S. Koushik, and K. Deepak. "Prediction of heart diseases using data mining and machine learning algorithms and tools." International Journal of Scientific Research in Computer Science, Engineering and Information Technology 3.3 (2018): 887-898.
[5] Shah, Devansh, Samir Patel, and Santosh Kumar Bharti. "Heart disease prediction using machine learning techniques." SN Computer Science 1.6 (2020): 1-6.
[6] Katarya, Rahul, and Sunit Kumar Meena. "Machine learning techniques for heart disease prediction: a comparative study and analysis." Health and Technology 11.1 (2021): 87-97.
[7] Du, Yingyi, et al. "Predicting drug-target interaction via wide and deep learning." Proceedings of the 2018 6th International Conference on Bioinformatics and Computational Biology. 2018.
[8] Al Imran, Abdullah, Md Nur Amin, and Fatema Tuj Johora. "Classification of chronic kidney disease using logistic regression, feedforward neural network and wide & deep learning." 2018 International Conference on Innovation in Engineering and Technology (ICIET). IEEE, 2018.
[9] Jais, Imran Khan Mohd, Amelia Ritahani Ismail, and Syed Qamrun Nisa. "Adam optimization algorithm for wide and deep neural network." Knowledge Engineering and Data Science 2.1 (2019): 41-46.
[10] fedesoriano. (September 2021). Heart Failure Prediction Dataset. Retrieved [Date Retrieved] from https://www.kaggle.com/fedesoriano/heart-failure-prediction.