Proceedings of the 3rd International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/4/2023434

Diabetes classification and prediction using artificial neural networks

Jiajie Wu

The School of Mathematics, University of Leeds, Leeds, United Kingdom, LS2 9JT

mm20j2w@leeds.ac.uk

Abstract. Diabetes is a chronic disease which threatens the global human health. Over time it could lead to other serious medical problems and does not have a permanent cure. Hence if we could diagnose or predict it early, then it might be possible to prevent it. Several researches have shown that computer technology can effectively assist in the diagnosis of diseases. And neural network could be used for classification and prediction. In order to identify the most significant element that has the greatest impact on the classification and prediction, this paper applies artificial neural networks to the PIMA Indian diabetes dataset. This study's neural network is a fully connected neural network. We used the fully connected neural network discussed above to the dataset that deleted one factor column per time, and we compared their accuracy to determine which factor was the most significant in this dataset. Finally, this study discovers that "Glucose" level, followed by "BMI," is the most crucial component for diabetes in this dataset, with a fully connected neural network having an accuracy of 84%.

Keywords: Deep Learning, Diabetes Classification and Prediction, Fully connected neural networks, PIMA dataset.

1. Introduction

We are all aware that diabetes is a long-term disease that results from insulin resistance or inappropriate insulin utilization. Type I diabetes, Type II diabetes, gestational diabetes, impaired glucose tolerance, and impaired fasting glycemia are the four main types of diabetes [1]. Two of these four types of diabetes are far more common than the others, and those are types I and II. Insufficient production of the hormone insulin causes type I diabetes, also known as insulin-dependent diabetes; insufficient use of insulin causes type II diabetes, also known as non-insulin-dependent diabetes. Several serious complications, such as lower limb amputation, kidney failure, heart attacks, and stroke, can emerge in people with diabetes over time [2]. This is because diabetes causes long-term damage to many biological systems, but especially the nerves and blood vessels. The World Health Organization estimates that there are 422 million people with diabetes worldwide, and that the disease is responsible for 1.5 million deaths each year [3]. And the prevalence and incidence of diabetes have been increasing during the past few decades [3].

Although diabetes cannot be cured, it can be managed with early detection and preventative care. Several studies from the past have shown that computers can be useful in the medical diagnosis process. Three machine learning models (neural networks, random forests, and decision trees) were used by Zou et al. to analyze data from the Luzhou and Pima Indian populations and make predictions about the

DOI: 10.54254/2755-2721/4/2023434

prevalence of diabetes mellitus. They eventually got the best results, with an accuracy of 80.84% on the Luzhou dataset and 77.21% among the Pima Indians [4]. Huma Naz and Sachin Ahuja present a strategy for diabetes prediction that makes use of a number of machine learning algorithms and the PIMA dataset. Accuracy in the 90-98% range can be achieved using operational classifiers like Artificial Neural Networks (ANNs), Naive Bayes (NBs), Decision Trees (DTs), and Deep Learning (DLs) [5]. Long short-term memory (LSTM) [6, 7], convolutional neural network (CNN) [6, 7], feedforward network [7], pattern network [7], cascade forward architecture [7], and Gaussian Process (GP) [8] are just a few examples of the deep learning (DL) architectures that have been the focus of related research to predict diabetes early.

In this research, we used a fully connected neural network to classify and forecast diabetes, and we applied this network to the PIMA Indian dataset to determine which predictability factor is most important. The purpose of this study is to see if very accurate predictions can be made using a relatively straightforward artificial neural network architecture. It also provides suggestions for which factors are most important for using in diabetes forecasting models. Therefore, clinicians should pay more attention to the factors like glucose, body mass index, and age that are more important for prognosis.

This paper is divided into six parts. To recap, the introduction is the first part. In part II, the preprocessing of the data is described together with the dataset used in this study. Section III displays the fully connected neural network after showing the dataset. In section IV, we use the aforementioned fully connected neural network to identify the key variable that has the greatest impact on this dataset's prediction accuracy.

2. Dataset

This research used the PIMA Indian dataset to classify and predict diabetes, which was compiled by the National Institute of Diabetes and Digestive and Kidney Diseases [9]. This dataset contains data on 768 individuals, all of whom are over the age of 21, on nine different topics. In these 768 records, 268 people are identified as having diabetes, whereas 500 others are not diabetic. Number of pregnancies, glucose levels, blood pressure, skin thickness, insulin levels, body mass index (BMI), diabetes family history, and age are just some of the eight components (the first eight columns in the dataset) and one outcome (the ninth) that make up the dataset (the last column in the dataset). For the outcome column, it only has two outcomes (1 or 0), which shows the person have diabetes or not (1 for having diabetes and 0 for not).

Before using this data directly, data preprocessing is necessary. There are two steps in the preprocessing of this dataset: the first one is rejecting any outliers; and the second one is filling in any zero values. Above all, to reject any outliers, we used the interquartile range (IQR) method. That means we use IQR for each column to determine if a point is an outlier or not. Firstly, we calculate the IQR for the data by the formula:

$$IQR = Q_3 - Q_1 \tag{1}$$

where Q_1 is the first quartile for its column, and Q_3 is the third quartile. And then any number greater than $(Q_3 + 1.5 * \text{IQR})$ or less than $(Q_1 - 1.5 * \text{IQR})$ is an outlier. As we can see, here are eight box plots for the eight factors in this dataset in order, respectively.

DOI: 10.54254/2755-2721/4/2023434

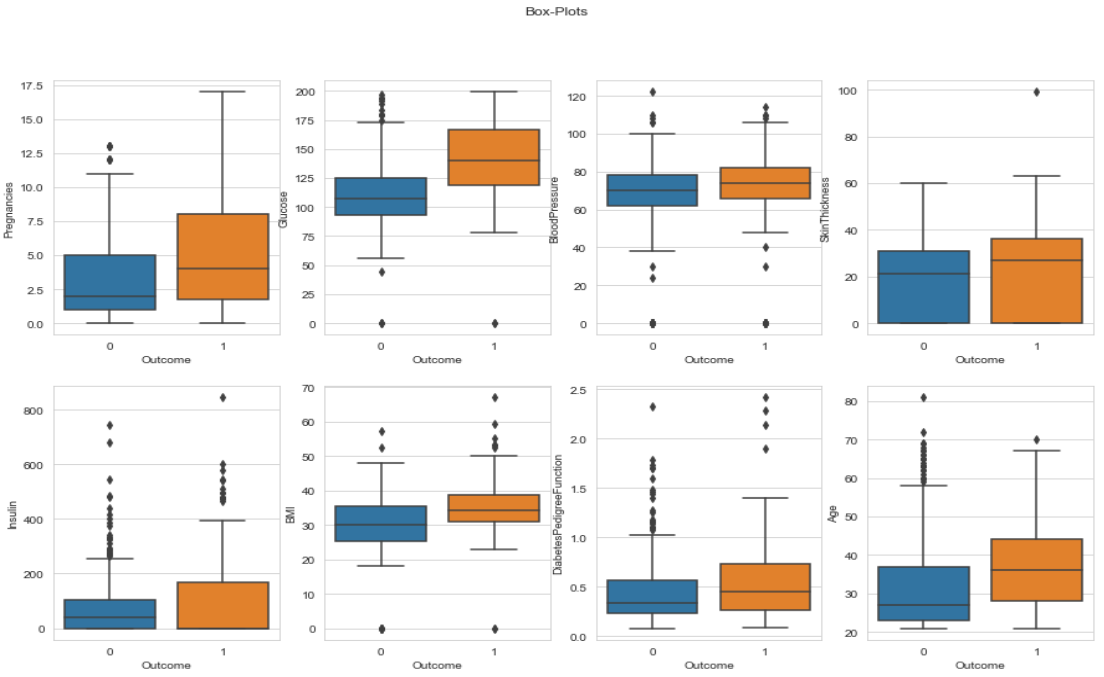

Figure 1. Eight box-plots for 'Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin', 'BMI', 'DiabetesPedigreeFunction' and 'Age' columns respectively [10].

In these box plots above, these black points outside the range from $(Q_1 - 1.5 * IQR)$ to $(Q_3 + 1.5 * IQR)$ are removed from this dataset.

In the PIMA dataset, there are lots of zero values in 'Glucose', 'Blood Pressure', 'Skin Thickness', 'Insulin' and 'BMI' columns, which mean that they are missing records since these number should not be zero. Hence, to fill these zero values, we could use their median values of their factor respectively to replace them instead of dropping these records. After the data preprocessing, the original 768 records are reduced to 636. The PIMA dataset is then used to classify and forecast diabetes by splitting it into a training set (445 records) and a testing set (191 records).

3. Fully connected neural network

A fully connected neural network is a type of artificial neural network that has one input layer, one output layer, and many hidden layers. Further, all neurons within a given layer are connected to neurons in subsequent layers [11]. The following image displays the fully linked neural network's architecture.

DOI: 10.54254/2755-2721/4/2023434

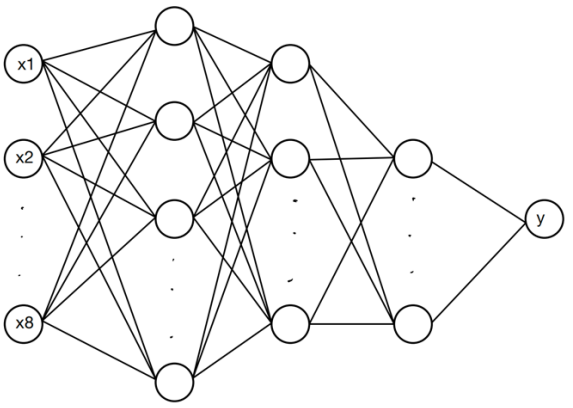

Figure 2. The architecture of the fully connected neural network.

In this paper, the inputs (x_1 to x_8) for the fully connected neural network are the 8 influence factors, and the output y is the 'Outcome' column where 1 for having diabetes and 0 for not. In order to find a model with greater accuracy, this fully connected neural network altered the number of neurons and hidden layers, which is limited to 100 and no more than 3. The activation functions for hidden layers are all ReLU, whose formula is shown as following:

$$f(x) = x^+ = \max(0, x) \tag{2}$$

and for output layer are Sigmoid:

$$S(x) = \frac{1}{1 + e^{-x}} \tag{3}$$

since this is a binary classification model, the output is either 0 or 1. The loss function is 'binary_crossentropy':

$$Loss = -\frac{1}{output\ size} \sum_{i=1}^{output\ size} y_i \cdot log\hat{y}_i + (1 - y_i) \cdot \log(1 - \hat{y}_i) \tag{4}$$

4. The most important factor in this dataset

To find the most important factor, we drop one factor column each time, use the fully connected neural network mentioned above applying to the rest of dataset to do the training and testing, and calculate its accuracy and loss. For example, we drop 'Pregnancies' column firstly, and the rest of dataset includes 7 influence factors and 1 outcome column. Repeat this for 8 times and compare their accuracy and loss with the original one, which is obtained without dropping any factors. If its accuracy decreased most and loss increased most by dropping one factor, then that factor is the most important factor in this PIMA dataset.

5. Results

Last but not least, the fully connected neural network that offered the best accuracy is composed of 3 hidden layers that each contain 64, 32, and 16 neurons. And its accuracy is 84%. Here is the confusion matrix and accuracy and loss function plots.

DOI: 10.54254/2755-2721/4/2023434

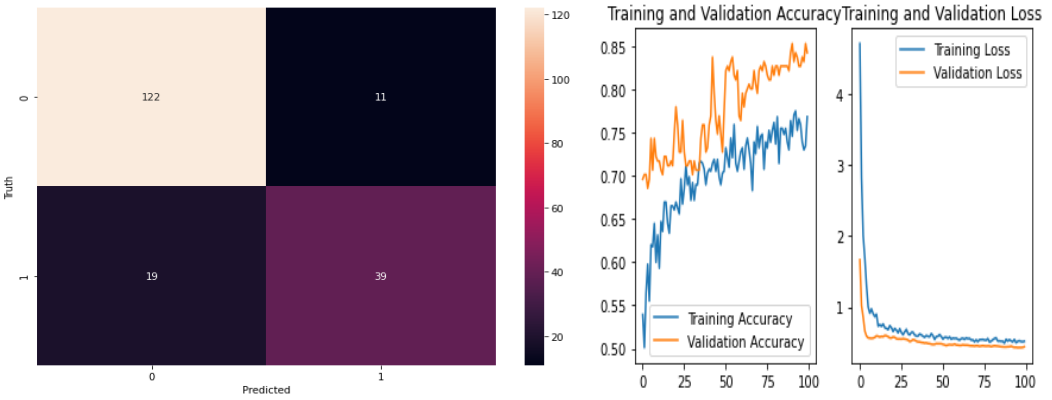

Figure 3. The picture on the left hand side (Picture 1) is the cunfusion matrix, and the right hand side one (Picture 2) shows how the accuracy and loss function change during 100 epochs.

To find the most important factor for the PIMA dataset, we obtained 8 pictures of accuracy and loss function after dropping one factor each time in order.

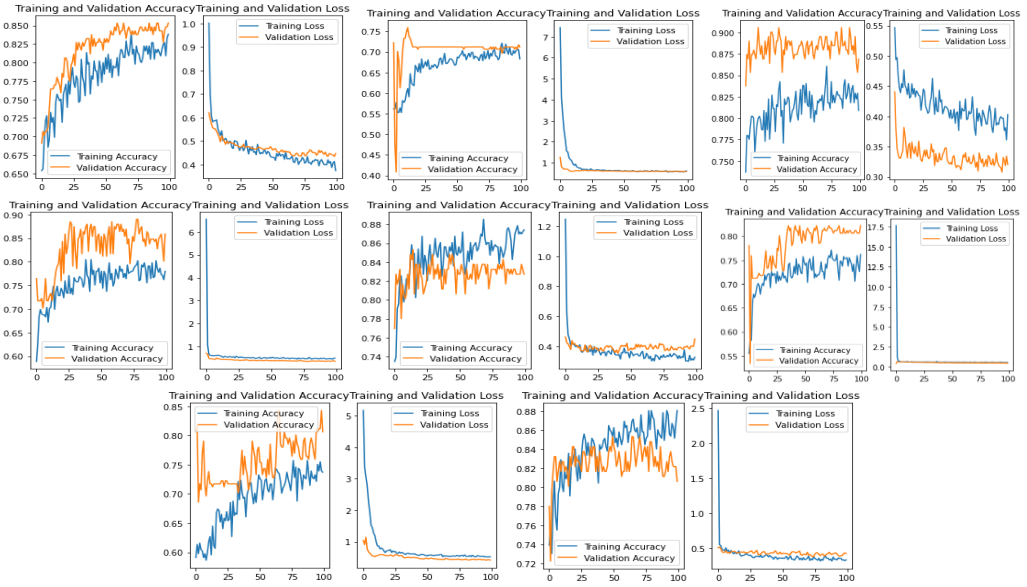

Figure 4. These pictures being provided after dropping ‘Pregnancies’ (Picture 3), ‘Glucose’ (Picture 4), ‘Blood Pressure’ (Picture 5), ‘Skin Thickness’ (Picture 6), ‘Insulin’ (Picture 7), ‘BMI’ (Picture 8), ‘Diabetes Pedigree Function’ (Picture 9) and ‘Age’ (Picture 10) respectively.

By comparing these eight accuracy and loss function plots above (Picture 3 to 10) with the original one (Picture 2) separately, we could find that “Glucose” level is the most important factor in this dataset, since the accuracy for dropping glucose column dropped most, from 84% to about 71%. The reason why glucose is the most important factor might be that high glucose is one of the symptoms of diabetes. Diabetes can lead to excess sugar in the blood, which would directly reflect on the glucose level [2-3]. After ‘Glucose’, ‘BMI’ and ‘Age’ are also important. Hence, when diagnose a person have diabetes or not, doctors could pay more attention on patients’ glucose level, BMI and age.

DOI: 10.54254/2755-2721/4/2023434

6. Conclusion

In this paper, the author used a fully connected neural network applying to PIMA diabetes dataset to classify and predict diabetes, and also find the most important factor in this dataset. 70:30 is the ratio between the training set and the testing set. Eight variables from the dataset were employed as inputs by the fully connected neural network, which then produced whether or not the subject had diabetes. To obtain the best model, the research changed the number of layers and neurons, dropping one factor at a time and comparing their accuracy to the original (no factor dropped) to determine the most important factor in this dataset. The model with the highest accuracy, 84%, is composed of three hidden layers, each with 64,32,16 neurons. In this dataset, the most important factor is "Glucose" level, following the 'BMI' and 'Age'.

In the future work, we could explore more data preprocessing techniques to make the dataset more effective., such as K-nearest neighbour (KNN) method, normalization, etc. We could also use other types of neural networks to improve the accuracy of prediction.

References

[1] Bdcc.pro. The Better Diabetes Care China Project. [online] Available at: <http://bdcc.pro/learning/get-to-know/types> [Accessed 29 August 2022].

[2] Who.int. n.d. Diabetes. [online] Available at: <https://www.who.int/news-room/fact-sheets/detail/diabetes> [Accessed 29 August 2022].

[3] Who.int. n.d. Diabetes. [online] Available at: <https://www.who.int/health-topics/diabetes> [Accessed 29 August 2022].

[4] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju and H. Tang, Predicting Diabetes Mellitus With Machine Learning Techniques, Front. Genet., vol. 9, pp. 1-10, 2018.

[5] H. Naz and S. Ahuja, Deep learning approach for diabetes prediction using PIMA Indian dataset, Journal of Diabetes & Metabolic Disorders, vol. 19, no. 1, pp. 391-403, 2020.

[6] S. G., V. R. and S. K.P., Diabetes detection using deep learning algorithms, ICT Express, vol. 4, no. 4, pp. 243-246, 2018.

[7] M. S. Diab, S. Husain and A. Jarndal, On Diabetes Classification and Prediction using Artificial Neural Networks, 2020 International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI), 2020, pp. 1-5, doi: 10.1109/CCCI49893.2020.9256621.

[8] M. Maniruzzaman, N. Kumar, M. M. Abedin et al., "Comparative approaches for classification of diabetes mellitus data: Machine learning paradigm", Comput. Methods Programs Biomed., vol. 152, pp. 23-34, Dec. 2017.

[9] Kaggle.com. n.d. Pima Indians Diabetes Database. [online] Available at: <https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database> [Accessed 30 August 2022].

[10] Kaggle.com. n.d. Diabetes-Prediction-using-NN. [online] Available at: <https://www.kaggle.com/code/alexreddy/diabetes-prediction-using-nn> [Accessed 7 September 2022].

[11] Liu M Y, Wu L J, Liang H, et al. A kind of high-precision LSTM-FC atmospheric contaminant concentrations forecasting model. Comput Sci, 2021, 48(6A): 184-189.