A Feed-Forward Neural Network Model For The Accurate Prediction Of Diabetes Mellitus

Yinghui Zhang, Zihan Lin, Yubeen Kang, Ruoci Ning, Yuqi Meng

Abstract: Diabetes mellitus is a group of metabolic diseases showing high blood sugar levels over prolonged periods. It is one of the deadly diseases growing at rapid rates in developing countries. Diabetes has affected over 246 million people worldwide. According to the World Health Organization (WHO) report, this number is expected to rise to over 380 million by 2025. If untreated, diabetes can lead to long-term complications such as heart disease and kidney failure. Therefore, there is a great need for the timely diagnosis of diabetes for people around the world. In particular, diabetes has been identified to be a very serious threat to younger generations and working individuals. Diabetes can be managed if it can be predicted during the early stages with changes in the diet and lifestyle of the patient. Therefore, this paper proposes a model for the early prediction of diabetes by considering major risk factors. An artificial neural network model with the Levenberg-Marquardt training algorithm is built using the PIMA Indian Diabetes dataset. The objective of the study is to predict the occurrence of diabetes mellitus using known risk factors based on feed-forward artificial neural network.

Index Terms: ANN, Diabetes, feed forward network, Levenberg-Marquardt training, Matlab, Neural Networks, Prediction of Diabetes.

1 Introduction

Diabetes mellitus (DM), generally referred to as diabetes, represents metabolic diseases with high blood sugar levels over prolonged periods. Typical symptoms include frequent urination and increased thirst, and if untreated, DM can cause serious long-term complications such as heart disease and kidney failure. DM can arise from the pancreas not producing sufficient amounts of insulin or the body not properly responding to insulin. There are two types of DM. Type 1 DM results from the pancreas's failure to produce sufficient amounts of insulin. Type 2 DM involves insulin resistance in which the body fails to properly respond to insulin. Prevention and treatment methods for DM typically involve a healthy diet and physical exercise. Type 1 DM requires insulin injections, and Type 2 DM may be treated with medication with or without insulin. As of 2015, there were about 415 million people with DM worldwide, with Type 2 accounting for 90%.

Prevalence of Diabetes and its complications

In the last decades, DM prevalence has sharply increased along with aging populations. DM is a significant contributor to mortality rates, reducing life expectancy in older DM patients [1]. DM was responsible for 4.5 million deaths worldwide in 2015 [2]. DM patients progressively develop several complications from hyperglycemia, and these can be broadly categorized as macro-vascular complications (e.g., coronary artery disease) and micro-vascular complications (e.g., diabetic neuropathy) [3]. Potential DM patients tends to be unware of health risks, but late diagnosis or a lack of treatment can exacerbate chronic vascular complications [4]. However, screening and detecting individuals can help delay the progression of the disease and the prevention of additional complications [1] and allow for greater control over treatment processes and reduced treatment cost [4]. Insulin administration is the principal treatment option for Type 1 DM, allowing the saving of life, alleviating disease symptoms, and preventing long-term complications. The most common medications include sulfonylurea and peptide analogs [5]. However, most have many side effects, and insulin treatment may result in weight gain and sudden hypoglycemia. Therefore, effective anti-diabetic drugs represent an urgent research topic. Although it is possible to prevent and limit DM complications through physical exercise and a better lifestyle, there are unmanageable risk factors including demographic characteristics such as age, sex, and race. Previous studies have examined the relationships between such factors with respect to DM prevalence. Men are more likely to develop DM than women (about 8% higher), and there is a positive correlation of DM and age. Non-demographic risk factors in DM include physical inactivity and obesity [6]. The rest of this paper is organized as follows: Section II provides a literature review. Section III presents the artificial neural network, and Section IV describes the dataset and analysis. Section V shows the network architecture, and Section VI discusses the results. Section VII concludes.

2 Literature Review

Ramesh et al. [7] implemented a deep learning neural network using predictive analytics based on the PIMA Indian diabetes dataset and applied logistic regression or data obtained using the restricted Boltzmann machine (RBM) to produce 81% accuracy. Yasaswi and Prajna [8] proposed a new architecture, applying the C4.5 classification algorithm for the prediction of DM using the PIMA Indian diabetes dataset. Ramesh et al. [9] proposed an optimal prediction model using the RBM in deep learning. The deep neural network was trained and tested with the tensor flow. Kambleand Patil [10] implemented the RBM in Java to classify DM and obtained 80% accuracy. Veena and Anjali [11] proposed a decision support system using the Adaboost algorithm with the decision stump as the base classifier and obtained 80.729% accuracy. They also compared the SVM, naïve Bayes, and decision tree classifiers. Radha etal. [12] compared five classification techniques C4.5, SVM, K-NN, PLR, and BLR to predict DM based on computing time and found BLR to have the shortest computing time with 75% accuracy and an error rate of 0.27, followed by the SVM. The BLR algorithm played a vital role in data mining techniques. Ayush and Divya [13] designed a DM prediction system based on daily lifestyle activities. They used the classification and regression tree (CART) algorithm for the prediction of DM and obtained 75% accuracy. Durairaj and Kalaiselvi [14] proposed a back-propagation algorithm for the prediction of DM using the PIMA Indian dataset. The classification accuracy of the back-propagation network was better than other methods. Sapna and Kumar [15] proposed the diagnosis of DM using clinical big data based on the neural network. They compared Quasi-Newton back-propagation and resilient back-propagation methods and found the latter to

perform better. Zhilbert etal. [16] developed an intelligent system for DM prediction by focusing on the joint implementation of the support vector machine (SVM) and NaïveBayes statistical modeling and found the joint implementation to perform better than individual implementation.

3 Artificial Neural Network

The artificial neural network (ANN) is a computational machine learning model based on the structure and function of a biological neural network. The information flowing through the network affects the ANN structure since the network changes (learns) through the input and output. The ANN is considered a nonlinear statistical data modeling tool in which complex input-output relationships are modeled for patterns.

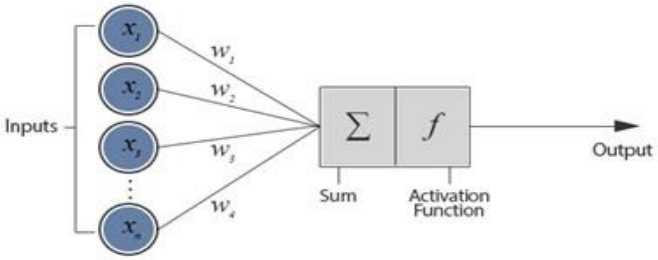

Fig . 1. *Neural Network*

The ANN has several advantages, but one of the most recognized is its ability to learn from datasets. As a result, the ANN can be used as a random function approximation tool. Such tools can help estimate optimal methods for arriving at solutions while defining computing functions and distributions. The ANN uses data samples instead of whole datasets for solutions, reducing time and cost. The ANN is considered a relatively simple mathematical model for enhancing existing data analysis methods. Such networks can learn through examples (training data), remember past experiences, and perform parallel processing. This kind of learning is possible because the "neurons" that can receive and process information are similar to those in the human brain. The input layer receives input from sensors and gives it to the subsequent layer for processing. The processing layer consists of summation and activation functions where the activation is checked using the threshold and the output signal is generated from the network.

Training the Neural network

Training the ANN is an iterative process that starts with the collection of data. Then data pre-processing enables the data to be ready and the training to be more efficient. During this data pre-processing process, data must be divided into three different sets, namely for training, validation, and testing purposes.

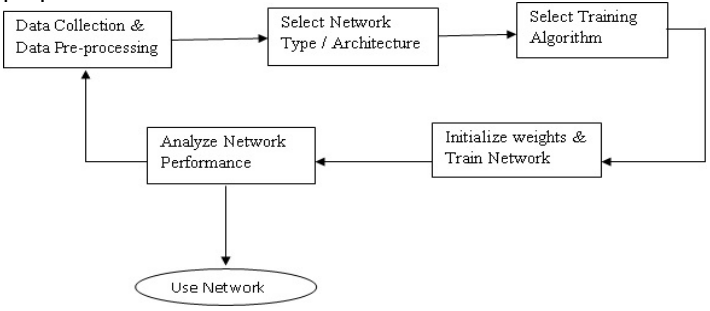

Fig. 2. *Neural network training flow chart.*

After data pre-processing, an appropriate type of network such as multi-layer, competitive, and dynamic, among others, must be selected, and the network architecture needs to be configured in terms of numbers of layers and neurons. Then the next step is the selection of the training algorithm. A training algorithm should be selected such that it is appropriate for the network and the problem at hand. After the network is trained, the network's performance is analyzed, which allows for the identification of any necessary changes required for the data, the network architecture, and the training algorithm. These changes are made in the next iteration. In this way, the whole process is iterated until satisfactory results are obtained for the network.

4 Dataset and Analysis

The dataset used in this study was the PIMA Indian Diabetes dataset. These people showed the highest risk of DM. The individuals in this dataset were under continuous study since 1965 by NIDDK because of the high risk of DM occurrence. This dataset was obtained from the UCI Machine Learning Repository, which consisted of 768 samples (268 diabetic and 500 non-diabetics). All individuals were Pima Indian women at least 21 years of age who lived near Phoenix, Arizona (USA). Among several attributes, eight were considered to be linked to DM (Table 1). In the dataset, a value 1 for the class indicates "tested positive for DM" and a value 0, "tested negative for DM."All these women had DM diagnosis tests. This dataset was one of the most popular DM datasets for DM researchers. The dataset had nine variables (eight input variables and one target variable). Since all eight variables were risk factors to be considered, the only data pre-processing task was to normalize the data between -1 and +1.

Table. 1
Attributes of DM Dataset

Sl	Risk factors		Range
	Attribute	Description	
1	Pregnancy	Number of times pregnant	0-17
2	Plasma glucose	Plasma glucose concentration 2 hours in an oral glucose tolerance test (mm Hg)	0-199
3	Diastolic BP	Diastolic blood pressure (mm Hg)	0-122
4	Triceps SFT	Triceps skin fold thickness (mm)	0-99
5	Serum-Insulin	2-Hour serum insulin (mu U/ml)	0-846
6	BMI	Body mass index (weight in kg/(height in m)^2)	0-67.1
7	DPF	Diabetes pedigree function	0.078 –2.42
8	Age	Age (years)	21-81
9	Class	Diabetes class variable	0-1

Table. 2
Class Distribution

Class value	No. of instances
0	500
1	268

Table. 3
STATISTICAL ANALYSIS OF DM DATASET

Attribute #	Mean	Standard deviation
1	3.8	3.4
2	120.9	32.0
3	69.1	19.4
4	20.5	16.0
5	79.8	115.2
6	32.0	7.9
7	0.5	0.3
8	33.2	11.8

5 NETWORK ARCHITECTURE

A multi-layered feed-forward network with 8 input nodes, 10 hidden nodes, and 1 output node was considered. The number of input nodes was the number of risk factors in the dataset. Since this was a small dataset with few attributes, all input attributes in Table 1 were fed to the network as input. The number of neurons in the hidden layer was based on the problem dataset, and the expected model performance was such that the training of the model was fast and provided the optimal output. Weights and biases of the neural network were initialized using the MATLAB configuration. For this, the Levenberg-Marquardt back-propagation algorithm was used for training and learning. The algorithm "trainlm" was the fastest back-propagation algorithm. A simple training operation on the network is not likely to result in optimal performance because of the possibility of reaching a local minimum. Therefore, training was restarted using different initial conditions, and the network that provided the best performance was selected. In addition, the network architecture was adjusted according to network performance. During the training stage, network performance was evaluated, and if the result was not satisfactory, the network configuration was adjusted by either increasing or decreasing the node and layers and changing the training algorithm. The network was configured to train based on the maximum number of epochs less than the default value of 100. The learning stopped when the predefined minimum error was reached or the number of epochs was completed.

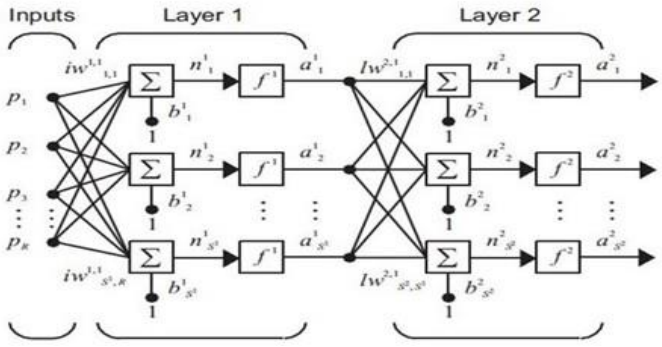

Fig. 3. *A Two layered feed forward network*

6 RESULTS AND DISCUSSIONS

The experimental analysis was done using MATLAB R2017a with the neural network tool box to implement the proposed algorithm. For this, 768 records of the dataset were imported and divided into training, validation, and testing datasets (70%, 15%, and 15%, respectively). Input data (input weight) were normalized to be transformed into the range -1 to +1 before use. These weights, together with random biases, were passed to the input layer of the neural network for training purposes. The performance measure of the mean square error (MSE) was used to evaluate network performance. The formula for mean square error was

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2$$

Fig. 4 shows the mean square error versus the epoch number. The green line indicates the validation error, and the blue line, the training error. In the target network, which had 10 neurons in the hidden layer, the minimum validation error occurred at epoch 6, as shown by the circle. The network parameters were saved at this point.

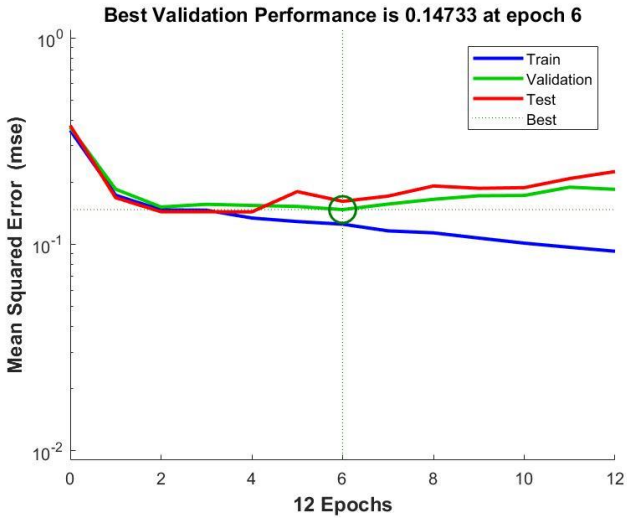

Fig. 4. *Performance graph*

Since the classification techniques had discrete target values, regression analysis was not useful for result validation. Therefore, a confusion matrix was used for validation purposes. The confusion matrix was constructed for training and validation, and testing as shown in Fig. 5. This matrix was a table with columns representing the target class and rows representing the output class. Correctly classified inputs are shown along the diagonal of the matrix, and off-diagonal cells indicate misclassified inputs.

Fig. 5. *Confusion Matrices*

Another important validation tool for classification problems is the receiver operating characteristic (ROC) curve. An ideal point for the ROC curve to pass through is (0,1), which corresponds to no false positives and only all true positives. Fig. 6 shows ROC curves for different datasets.

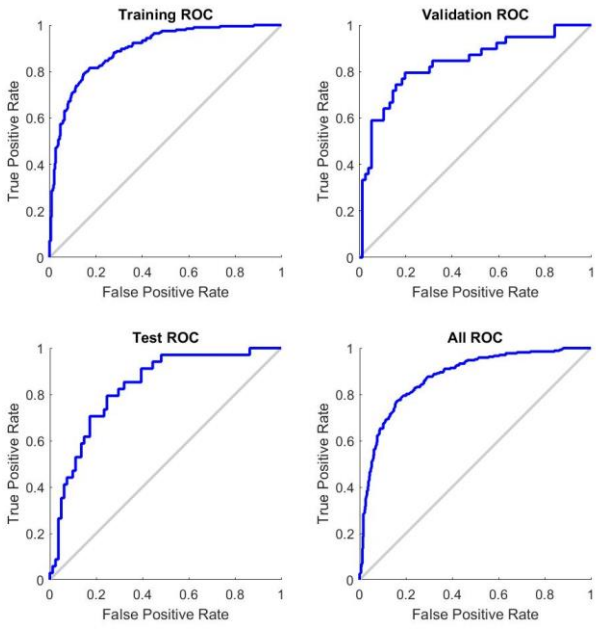

Fig. 5. *ROC Curves*

The confusion matrix in Fig. 5 indicates that the network achieved 81.9% accuracy. With training data, 82.7% accuracy was achieved. That is, 445 out of 538 were classified correctly, and only 93 were misclassified. In addition, with validation and testing data, accuracy rates of 82.6% and 77.4% were respectively achieved. The ROC curves in Fig. 6 indicate that the training ROC curve was closer to the idle ROC curve and the training curve and all ROC curve were identical. The figures Fig. 4, Fig. 5, and Fig. 6 show that the proposed model gave better prediction accuracy, suggesting its applicability to DM prediction.

7 Conclusion

This paper presents a multi-layer feed-forward neural network for the prediction of DM using risk factors in the PIMA Indian DM dataset. The Levenberg-Marquardt training algorithm was used to train the network, and the mean square error was used to measure performance. The network model was trained several times to obtain an accuracy value over 80%. Among the 768 individuals in the dataset, 268 were diabetic, and 500 were non-diabetic. With this non uniform dataset, the proposed model gave satisfactory results with 82% accuracy. Future research should extend this study by improving the training method and changing the activation function with a deep neural network to better predict DM in early stages.

References

[1] Rydeń L, Standl E, Bartnik M, Van den Berghe G, Betteridge J, De Boer MJ, et al. Guidelines on diabetes, pre-diabetes, and cardiovascular diseases: full text. European Heart Journal Supplements. 2007; 9 (suppl C):C3–C74. https://doi.org/10.1093/eurheartj/ehl261

[2] International Diabetes Federation, http://www.diabetesatlas.org.

[3] Cade WT. Diabetes-related microvascular and macrovascular diseases in the physical therapy setting. Phys Ther Nov 2008;88(11):1322–35.

[4] Habibi S, Ahmadi M, Alizadeh S. Type 2 Diabetes Mellitus Screening and Risk Factors Using Decision Tree: Results of Data Mining. Global journal of health science. 2015; 7(5):304. https://doi.org/10.5539/gjhs.v7n5p304 PMID: 26156928

[5] Krentz AJ, Bailey CJ. Oral antidiabetic agents: current role in type 2 diabetes mellitus. Drugs 2005;65(3):385–411.

[6] Alghamdi, Manal, et al. "Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford Exercise Testing (FIT) project." PloS one 12.7 (2017): e0179805.

[7] Sushant Ramesh, et al. “A Deep Learning Approach to Identify Diabetes”, Advanced Science and Technology Letter Vol. 145 (NGCIT 2017): Pp. 44-49

[8] BaratamYasaswi and BodapatiPrajna, “The Early Augmentation for Diabetes Diagnosis Using Data Mining Approaches”, IJCST Vol. 7, Issue 3, July-Sept 2016: pp 27-31.

[9] Sushant Ramesh et al. “Optimal Predictive Analytics of Pima Diabetics Using Deep Learning”, International Journal of Database Theory and Applicaation. Vol.10, No.9 (2017) pp. 47-62.

[10] Kamble and Patil, “Diabetes Detecting using Deep Learning Approach”, International Journal for Innovative Research in Science & Technology, Vol.2, Issue.12, May 2016. Pp. 342-349

[11] Veena Vijayan and Anjali, “Prediction and Diagnosis of Diabetes Mellitus – A Machine Learning Approach”, IEEE Recent Advances in Intelligent Computational Systems (RAICS), Dec-2015, pp: 122-127.

[12] Radha & Srinivasan, “Predicting Diabetes by Consequencing the various Data Mining Classification Techniques”, International Journal of Innovative Science, Engineering and Technology, Vol.1, Issue.6,Aug-2014.

[13] Ayush Anand., Divya Shakti.: Prediction of Diabetes Based on Personal Lifestyle Indicators.: IEEE 1st Intl. Conf. on Next Generation Computing Technologies (NGCT), Dehradun, pp. 673-676. (2015).
[14] Durairaj., Kalaiselvi.: Prediction of Diabetes Using Back Propagation Algorithm.: Intl. Journal of Emerging Technology and Innovative Eng., Vol.1, Issue 8, pp.21-25, (2015)
[15] Sapna S., Pravin Kumar M.: Diagnosis of Disease from Clinical Big Data Using Neural Network.: Indian Journal of Science and Technology, Vol 8(24)., pp. 1-7., (2015).
[16] ZhilbertTafa., NerxhivanePervetical., BertranKarahoda.: An Intelligent System for Diabetes Prediction: 4th Mediterranean Conf. on Embedded Computing.pp. 378-382. (2015).