A feedforward artificial neural network model for classification and detection of type 2 diabetes

Enoch A. Frimpong¹, Ariyo Oluwasanmi¹, Edward Y. Baagyere¹,², Qin Zhiguang¹

¹School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu, China, No.4, Section 2, North Jianshe Road, 610054, Chengdu, P.R.China
²Department of Computer Science, University for Development Studies, Box TL 1350, Tamale, Ghana
E-mail: ybaagyere@uds.edu.gh and frimpongadienoch@ymail.com

Abstract. Efforts to enhance accuracy in medical diagnostics in molecular medicine have contributed to the wide use of artificial neural network (ANN) algorithms for disease detection due to its ability to process large medical datasets and integrate them into characterized outputs to avoid misdiagnosis. Typically, the application of ANNs have proven useful in sample analyses of patients with diabetes and in decision support systems. Over the years, various ANN models have been utilized in medical diagnostics; however, these approaches still maintain certain levels of error and have lesser training and testing accuracies in disease detection. In this study, we propose a Feedforward Artificial Neural Network (FFANN) model with a dense neural network architecture suitable for processing numeric and textual dataset. We carefully designed our model structure to have the ability to maximize the number of layers and nodes required to learn every feature of the dataset and also to perform effective computations but avoiding model under fitting and overfitting which occurs when less or more layers are used respectively. This approach puts our model ahead of other state-of-the-art prediction models which have been proposed in terms of performance as it achieved 97.27% and 96.09% training and testing accuracies, respectively, for type 2 diabetes detection on Pima Indian Diabetes dataset.

1. Introduction
Diabetes is a condition where the sugar level in the blood is raised, for the sugar in our blood is tightly controlled by insulin. Diabetes occurs when control mechanisms are defected and the pancreas is unable to produce insulin or when the produced insulin cannot be used effectively [1]. The high sugar is poisonous particularly to the internal organs most especially the blood vessels in the body, at the initial period, the complications begin to build up and may not be felt by the patient for a very long time. High blood sugar is associated with high blood pressure, high cholesterol and excess weight, lack...
of exercise, unhealthy diet, Increasing age, impaired glucose tolerance (IGT), poor nutrition during pregnancy, history of gestational diabetes and cigarette smoking. The high blood sugar is harmful to the body and can result in severe health problems like cardiovascular disease, kidney disease and sight loss [2].

It has also been known that the prevalence of this disease is due to urbanization and economic development [3]. Diabetes happens to be one of the world’s most common and deadly diseases which affects people of all ages and has long been a serious health challenge in both developing and developed countries [4]. According to the International Diabetes Federation (IDF), the population of adults living with diabetes from year 2000 to 2019 has grown from 151 million to 463 million and it is estimated that by the year 2045, there will be about 700 million adults living with diabetes [5]. It is therefore a worldwide problem with devastating social, economic and human impact [6]. Proceeding many years of research and developments in diabetes treatments, there are still no definite cure for diabetes. There are two types of diabetes mellitus which are type 1 and type 2 [7].

In this paper, our primary aim is to develop an artificial neural network model to detect diabetes at an early stage. A timely diagnosis of such deadly disease and putting patient on the right treatment prevents the disease from advancing into complex health problems and also cuts down cost in further therapeutic treatments [8], [9]. Early intervention associated with lifestyle modifications or pharmacotherapy has proven to effectively prevent or delay type 2 diabetes and complications in adults [10]. There have been diverse solutions proposed for all kinds of diseases [7]. Expert systems and artificial intelligence methods such as artificial neural networks (ANN) are extensively used to aid diabetes diagnosis [11], this is due to the fact that ANN has a high classification proficiency [12] and its architecture which makes it easy for users to build varying types of networks [13].

In our research, we used Pima Indiana Diabetes datasets [14] (PIDD) to evaluate our model for type 2 diabetes detection due to the fact that it has been used extensively over the years by other researchers for diabetes detection which guarantees the authenticity of the dataset[15][16].

2. Related Works

Choubey et al. [17] used Genetic Algorithm (GA) for selection of variables on PIDD, they then applied Naïve Bayes (NB) on the variables selected for classification of diabetes disease which obtained 78.69% accuracy.

Olaniyi, et al. [18] diagnosed onset diabetes using multi-layer feedforward trained with back-propagation algorithm. Back-propagation is considered as one of the simplest techniques used for supervised training algorithms of multilayered neural networks and its error correction based. It approximates the non-linear relationship between the input and output by internally adjusting the weight values with respect to the calculated error so that at the end of each cycle, the obtained error size will be less than the previous cycle value. The authors considered using PIDD due to the fact that it had adequate attributes to improve training and classification test of patients. The model classifies patient’s status as either binary 1 or binary 0, indicating whether the patient is tested positive or negative respectively. Their neural network gained 82% success rate on test which is much better than C4.5, ADAP and EM algorithms.

H. Roopa et al. [19] proposed a Principal Component Analysis with Linear Regression Method (PCA-LRM) where features of the data are extracted and projected to a different space by means of principal component analysis, which is then modelled by employing linear regression method on the newly created attributes. The accuracy attained by this method is 82.1% for diabetes prediction.

S. Kumar et al. [20] used data obtained from 250 diabetic patients both male and female from ages 25-78 from Pusat Perubtan Universiti Kebangsaan in Kuala Lampur. They trained the dataset using MATLAB neural network toolbox and attained an accuracy rate of 88.80%.

R.B Lukmanto et al. [21] proposed a fuzzy logic system to classify and detect patients with diabetes.
They identified the highly discriminative features in the PIDD using F-Score feature selection. In an instance where a dataset’s feature has a missing value of about 5%, it terminated, they only considered features with high F-score as informative feature. They further used fuzzy support vector machine techniques for classification analysis and to generate the fuzzy rules and fuzzy inference. Their proposed method attained 89.02% accuracy.

Z. Soltani et al. [7] also took advantage of PIDD. They processed the data in MATLAB using a probabilistic artificial neural network (PNN) to diagnose diabetes due to its high speed in training. PNN is extensively used in pattern recognition problems and classification. Its operations are organized into a multi layered feedforward network comprising of an input layer, pattern layer, summation layer and output layer. Their approach also attained an accuracy level of 89.56%.

N. Barakat et al. [22] proposed an Intelligible SVM for diabetes diagnoses. They used datasets of 4682 subjects from age 20 and above in the Sultanate of Oman collected by the 1991 National Survey of Diabetes. They chose a majority sample of non-diabetic subjects using sub-sampling to overcome extensively skewed classification problems. This is because class imbalance may prevent the learning algorithm from learning good models. They also used K-means clustering algorithm to devise five clusters from the original data and to sort out sample (subjects) within each cluster with respect to their Euclidean distance from the center of the cluster. Their model achieved a prediction accuracy of 94%.

M.F. Ijaz et al. [23] proposed a Hybrid Prediction Model (HPM), their model consists of Density-based Spatial Clustering of Application which includes a Noise-based outlier detection (DBSCAN) which eliminates the outlier data, a Synthetic Minority Over-Sampling Technique (SMOTE) used for balancing the class distribution and Random Forest (RF) for diseases classification. They considered three dataset benchmarks to predict both diabetes and hypertension risk at the early stage. When the model was applied to diabetes dataset I, it attained a 92.555% prediction accuracy. To authenticate the prediction accuracy, they applied the HPM to hypertension dataset II, which also gained 76.419% accuracy.

3. Proposed Methodology

The proposed model implements a dense neural network with three hidden layers alongside the input and output layers as shown in figure 1. The input layer has eight nodes indicating each columns of the dataset. Then, the three hidden layers have 200, 200 and 150 neurons respectively. The model follows the multilayer perception approach which is an architecture of feedforward artificial neural network (ANN) consisting of multiple layers of perceptrons or dense nodes. A dense network is selected because it is more suitable for this task being a numeric dataset rather than other models such as convolutional neural network and recurrent neural network which are preferable for images and sequence task respectively. Each of the layers is activated using Rectified Linear Unit (ReLu) responsible for transforming the summed weight of a layer’s input into its output, making training easier and also enhancing performance. The model output layer has two nodes with Sigmoid activating function to determine the existence or non-existence of diabetes.

The model objective function is to minimize the cost function computed from the data label and the model’s prediction. The cost function is calculated by finding the binary cross entropy (BCE) loss which is represented in (1).

\[
BCE = -\frac{1}{N} \sum_{i=0}^{N} y_i \cdot \log(\hat{y}_i) + (1-y_i) \cdot \log(1-\hat{y}_i)
\]

(1)

Where, \(N\) is the total number of data points, \(y\) is the value of the dataset label, and \(\hat{y}\) is the corresponding probability of the model’s prediction. The model is trained by back propagating the loss to find the global minimum with a learning rate of 0.001 using Adam optimizer.
4. Experimental Analysis

4.1. Dataset
In this paper, we used the Pima Indian Diabetes dataset to test and train our model for diagnosing type 2 diabetes. The dataset consists of 768 type 2 diabetes data samples each of these with eight attributes. The dataset was collated from the population of Pima Indian females in Arizona [24]. According to [14] this population has the highest diabetes rate as compared to any other population all over the world. They also record about 70% hypertension and obesity rate. The statistical summary of the dataset is shown in table 1.

\begin{table}[h]
\centering
\begin{tabular}{|l|c|c|c|}
\hline
Feature Name & Quantiles & Mean & Standard Deviation \\
& min & max & \\
\hline
Number of times pregnant & 0 & 17 & 3.85 & 3.37 \\
Glucose tolerance test & 0 & 199 & 121 & 121 & 32 \\
Diastolic blood pressure (mm Hg) & 0 & 122 & 69.1 & 19.3 \\
Triceps skin fold thickness (mm) & 0 & 99 & 20.5 & 15.9 \\
2Hr Serum Insulin (mu U/ml) & 0 & 846 & 79.8 & 115 \\
Body mass index (Kg/m^2) & 0 & 67.1 & 32 & 7.88 \\
Diabetes pedigree function & 0.08 & 2.42 & 0.47 & 0.33 \\
Age (yrs.) & 21 & 81 & 33.2 & 11.8 \\
\hline
\end{tabular}
\caption{Mean and Standard deviation analysis [14]}\end{table}

4.2. Evaluation Metrics
In this work, we report the efficiency of the model by evaluating its accuracy and F1 score metrics as shown in Equation (2) and (5), where TP, FP, TN and FN represents true positives, false positives, true negatives and false negatives respectively. To compute the model’s F1 Score, we calculate the precision as the number of TP divided by the sum of the number of TP and FP, and recall is calculated as TP divided by the sum of TP and FN as presented in Eq. (3) and Eq. (4) respectively.

\begin{equation}
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}
\end{equation}
\[ \text{Precision} = \frac{TP}{TP + FP} \]  

\[ \text{Recall} = \frac{TP}{TP + FN} \]  

\[ F1\text{Score} = 2 \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \]

4.3. Model Training

The training dataset was divided into 70% for training, 20% for testing, and 10% for validation, as done normally in literature. The data was trained on our model with three hidden layers. The model’s prediction loss was computed using cross-entropy loss function and minimized using back-propagation techniques. The model’s parameters are updated using Adam optimizer with learning rate of 0.001, and a batch size of 500 while the dataset is trained on 500 epochs. In optimizing our model design appropriately, we varied several of the hyperparameters such as number of epochs, number of neurons and number of layers in the model architecture. Furthermore, to confirm the parameter analysis of our model, and to eliminate forms of overfitting and underfitting, we run our model for different epoch values for 100, 200, 300, 400 and 500 epochs respectively. It was realized that the model began overfitting above 500 epochs, causing huge gap between the training and validation scores. Also, after testing various optimizers and learning rates, we found that the Adam optimizer and a learning rate of 0.001 gave the best model performance in terms of training and testing accuracy. The effect of the training epochs on the model are shown in Table 2. The best results were obtained at 500 epochs, thus our three-layer network produced the best result with these parameter settings compared to other designs in both training and testing accuracy.

| Epochs | Train Accuracy % | Test Accuracy % | Precision % | Recall % | F1 Score % |
|--------|------------------|----------------|-------------|----------|------------|
| 500    | 97.27            | 96.09          | 96.53       | 93.28    | 94.88      |
| 400    | 89.09            | 83.75          | 86.48       | 83.57    | 82.38      |
| 300    | 82.64            | 77.74          | 76.63       | 75.44    | 73.29      |
| 200    | 77.04            | 60.05          | 65.66       | 71.80    | 64.27      |
| 100    | 67.53            | 55.72          | 53.97       | 61.82    | 57.62      |

4.4. Result and Analysis

In this section, we discuss our model’s performance against other existing models of the same research focus, as well as the model accuracy and loss against the epoch cycle. As depicted in Figure 2, the graph shows the model accuracy increasing over the epoch. This indicates that the model is able to learn the gradient of the loss function, such that the global minimum of the model is reached easily. At the same time, while the accuracy increases, the loss decreases without any sign of underfitting or overfitting as depicted in Figure 3. Also, Figure 4 shows the ability of our model to generalize as both our train and test accuracies show high percentage of accuracy respectively. In comparison to other approaches, our FFANN model has the highest test accuracy as illustrated in Table 3, recording about 2.18% improvement over the SVM [22], 3.67% over the HPM [23], 6.80% over the PNN [7], 7.36% over the FS-FSVM [21], 7.59% over the Bayesian-Regulation [20], 14.56% over the PCA-LRM [19], 14.66% over the Back-Propagation [18], and 18.11% over the GA-NBs [17] models test accuracies.
Figure 2. Model’s accuracy against the epoch cycle.

Figure 3. Model’s loss against the epoch cycle.

Figure 4. Accuracy in training and testing our model with Pima Indian Diabetes dataset.

Table 3. Accuracy comparison between proposed model and other models.

| Previous works          | Model        | Accuracy  |
|-------------------------|--------------|-----------|
| Choubey et al. [17]     | GA-NBs       | 78.69%    |
| Olaniyi et al. [18]     | Back propagation | 82.00%    |
| H. Roopal et al. [19]   | PCA-LRM      | 82.1%     |
| S. Kumar et al. [20]    | Bayesian regulation | 88.80%    |
| Lukmanto et al. [21]    | FS-FSVM      | 89.02%    |
| Z. Soltani et al. [7]   | PNN          | 89.56%    |
| M. F Ijaz et al. [23]   | HPM          | 92.56%    |
| N. Barakat et al. [22]  | SVM          | 94.00%    |
| Our study               | FFANN        | 96.09%    |

5. Conclusion

Our proposed Feedforward Artificial Neural Network with three layers was designed to effectively learn every attribute of the dataset, also avoiding under and overfitting issues leaving less room for error in the detection of similar features from unknown or test dataset. This technique makes the model applicable to other medical datasets for detection of different kinds of disease in relation to that dataset. Our proposed model is well optimized and achieves better results than other models in literature.
References

[1] Bhojane, D.Y., Palhal, A.P., Daphal, G.T. and Bairagi, J.H., (2018). A comparison of mechanisms of action of hypoglycemic principles of momordica cymbalaria and synthetic antidiabetic drug.

[2] Centers for Disease Control and Prevention, (2011). National diabetes fact sheet: national estimates and general information on diabetes and prediabetes in the United States, 2011. Atlanta, GA: US department of health and human services, centers for disease control and prevention, 201(1), 2568-2569.

[3] Forouzanfar, M.H., Afshin, A., Alexander, L.T., Anderson, H.R., Bhutta, Z.A., Biryukov, S., Brauer, M., Burnett, R., Cercy, K., Charlson, F.J. and Cohen, A.J., (2016). Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015. The lancet, 388(10053), 1659-1724.

[4] Lieberman, L.S., (2003). Dietary, evolutionary, and modernizing influences on the prevalence of type 2 diabetes. Annual review of nutrition, 23(1), 345-377.

[5] Iwanegbe, I., Suleiman, M., & Jimah, A. (2019). Treatment of Diabetic Condition in Streptozotocin Induced Diabetic Wister Rats Using Food Blends Such as Unripe Plantain, Soybean and Ginger. International Journal of Biochemistry Research & Review, 1-12.

[6] Zhang, Y., Dall, T.M., Mann, S.E., Chen, Y., Martin, J., Moore, V., Baldwin, A., Reidel, V.A. and Quick, W.W., (2009). The economic costs of undiagnosed diabetes. Population health management, 12(2), pp.95-101.

[7] Soltani, Z. and Jafarian, A., (2016). A new artificial neural networks approach for diagnosing diabetes disease type II. International Journal of Advanced Computer Science and Applications, 7(6), pp.89-94.

[8] Sa’di, S., Maleki, A., Hashemi, R., Panbechi, Z. and Chalabi, K., (2015). Comparison of data mining algorithms in the diagnosis of type II diabetes. International Journal on Computational Science & Applications (IJCSEA), 5(5), pp.1-12.

[9] Zebardast, B., Rashidi, R., Hasanpour, T. and Gharehchopogh, F.S., (2014). Artificial neural network models for diagnosing heart disease: a brief review. International Journal of Academic Research, 6(3), pp.73-78.

[10] Knowler, W.C., Barrett-Connor, E., Fowler, S.E., Hamman, R.F., Lachin, J.M., Walker, E.A. and Nathan, D.M., (2002). Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. The New England journal of medicine, 346(6), pp.393-403.

[11] Erkaymaz, O., Ozer, M. and Perc, M., (2017). Performance of small-world feedforward neural networks for the diagnosis of diabetes. Applied Mathematics and Computation, 311, pp.22-28.

[12] Haykin, S., (1994). Neural networks: a comprehensive foundation. Prentice Hall PTR.

[13] Gurney, K., (1997). An introduction to neural networks. CRC press.

[14] Pima Indians Diabetes Database. https://www.kaggle.com/uciml/pima-indians-diabetes-database.

[15] Han, J., Rodriguez, J.C. and Beheshti, M., (2008), December. Diabetes data analysis and prediction model discovery using rapidminer. In 2008 Second international conference on future generation communication and networking (Vol. 3, pp. 96-99). IEEE.

[16] Al Jarullah, A.A., (2011), April. Decision tree discovery for the diagnosis of type II diabetes. In 2011 International conference on innovations in information technology (pp. 303-307). IEEE.

[17] Choubey, D. K., Paul, S., Kumar, S., & Kumar, S. (2017, February). Classification of Pima indian diabetes dataset using naive bayes with genetic algorithm as an attribute selection. In Communication and Computing Systems: Proceedings of the International Conference on Communication and Computing System (ICCCS 2016) (pp. 451-455).
[18] Olaniyi, E.O. and Adnan, K., (2014). Onset diabetes diagnosis using artificial neural network. *Int. J. Sci. Eng. Res.*, 5(10), pp.754-759.

[19] Roopa, H., & Asha, T. (2019). A Linear Model Based on Principal Component Analysis for Disease Prediction. *IEEE Access*, 7, 105314-105318.

[20] Kumar, S. and Kumaravel, A., (2013). Diabetes diagnosis using artificial neural network. *International Journal of Engineering Sciences & Research Technology*, pp.1642-1644.

[21] Lukmanto, R.B., Nugroho, A. and Akbar, H., (2019). Early Detection of Diabetes Mellitus using Feature Selection and Fuzzy Support Vector Machine. *Procedia Computer Science*, 157, pp.46-54.

[22] Barakat, N., Bradley, A.P. and Barakat, M.N.H., (2010). Intelligible support vector machines for diagnosis of diabetes mellitus. *IEEE transactions on information technology in biomedicine*, 14(4), pp.1114-1120.

[23] Ijaz, M.F., Alfian, G., Syafrudin, M. and Rhee, J., (2018). Hybrid prediction model for type 2 diabetes and hypertension using DBSCAN-based outlier detection, synthetic minority over sampling technique (SMOTE), and random forest. *Applied Sciences*, 8(8), p.1325.

[24] Asuncion, A., & Newman, D. (2007). UCI machine learning repository.