ORIGINAL ARTICLE

INTERNATIONAL JOURNAL OF RESEARCH IN PHARMACEUTICAL SCIENCES

Published by JK Welfare & Pharmascope Foundation Journal Home Page: www.pharmascope.org/ijrps

Diagnosis of Type II Diabetes Based on Feed forward Neural Network Techniques

Laman R. Sultan*

Department of Power Mechanics, Basra Technical Institute, Southern Technical University, Al-Basrah, Iraq

Article History:

Received on: 04.09.2019
Revised on: 15.12.2019
Accepted on: 27.12.2019

Keywords:

Diabetes Type II, Prediction, Feedforward Neural Network

ABSTRACT

Diabetes is a disease caused by an increase in blood glucose levels due to insulin secretion deficiency (type I diabetes) or impaired insulin activity (type II diabetes). More than 90% of people with this condition are diagnosed with type II diabetes. Due to the sharply prevalence of type 2 diabetes in recent years, the prognosis and early diagnosis of the disease have become even more important. In this study, a model for diagnosis of type II diabetes was developed using Artificial Neural Network (ANN) method. The execution of the Frequent Pattern Growth algorithm on medical data is difficult. Association rule-based classification is an interesting area focused that can be utilized for early diagnosis. The discretization phase is necessary to transform numerical characteristics. Pima Indians Diabetes Data Set is taken as an input. The execution time, a number of rules generation and the detection of outlier percentage are analyzed. The CFP-growth algorithm utilizes for finding frequent patterns where constructing the Minimum Item Support (MIS)-tree, CFP-array and producing frequent patterns from the MIS-tree. From the set of frequent itemsets found, create all the association rules that have confidence exceeding the minimum confidence. In this study, we aim to build a model that helps Physicians in predicting Diabetes early and accurately. Data were collected from the PIMA Indian data set. Consisted of 768 samples (268 diabetic and 500 non-diabetics). Levenberg-Marquardt back-propagation algorithm was used to train the network, and the accuracy of the prediction of whether a person is diabetics or not was 88.8%.

*Corresponding Author

Name: Laman R. Sultan
Phone:
Email: laman.radi@stu.edu.iq

ISSN: 0975-7538

DOI: https://doi.org/10.26452/ijrps.v11i1.1943

Production and Hosted by

Pharmascope.org

INTRODUCTION

Diabetes is a disease that continues a long life. Treating it sometimes is difficult. One of the functions of the pancreas is to support the body with insulin. Failing to do this causes diabetes. Another cause is producing insulin but the failure in using it by the body efficiently. Insulin is connected with the level of sugar in the blood. It controls sugar to maintain the body healthy. Uncontrolling causes Hyperglycemia or hyperglycemia, which, over time, causes severe damage to many organs, particularly nerves and blood vessels (Temurtas *et al.*, 2009).

Statically, the number of Diabetes patients increases. They have reached 200 million people, which have doubled in the last ten years worldwide. It increases by about seven percent in the annual predominance of diabetes in the world. In 2013, 3.8 million people had died, whether because of diabetes or and high blood glucose. Diabetes patients More than are

infected and about. People for a long time suffered from different diseases that in some cases have been able to diagnose diseases and offer them the solution in order to enhance it, but unfortunately, sometimes, due to the lack of diagnosis of symptoms in patients for a long time may even threaten the life of the patient. Therefore, many studies have been done in the field of predicting for several diseases to the extent that today's human take advantage of decision supports models and smart method to predict. One of the decision support models application is in the medical field and diagnosis of illnesses such as diabetes (Motka *et al.*, 2013). Deferment in the diagnosis and prediction of diabetes due to insufficient control of blood glucose increases macrovascular and Capillaries difficulties risk, ocular diseases and kidney failure (Manzella *et al.*, 2005; Mohammed *et al.*, 2019).

Diabetes are of 2 types. The first is type I, which called insulin-dependent, and the second is type II diabetes, which called relative insulin deficiency (Morteza *et al.*, 2013). There are two categories of protracted complications of diabetes. The first category is called vascular complications of diabetes and the second is named non-vascular complications of diabetes.

The former category includes microvascular (eye disease, neuropathy, nephropathy) and macrovascular complications (coronary artery disease, peripheral vascular disease, cerebrovascular disease). The second category includes gastroparesis, sexual dysfunction, and skin changes (Chavey *et al.*, 2014).

Using Computers in the field of medical Diagnosis is increasing dramatically, and the medical industry research in this area is open. The quality and accuracy of disease diagnosis using a technique called Machine learning are improved by recent researchers. And to classify the data sets, the researchers have developed many types of this technique which have been successfully applied to clinical data. For instance, they have been used in the prediction of patient progress and length of stay (Elzamly *et al.*, 2015).

Several neural network models have been applied for T2DM diagnosis prediction, summarized in Table 1. Multi-Layer Perceptron models were applied on various datasets. (Motka *et al.*, 2013) and (Polat and Güneş, 2007) used Artificial Neural Fuzzy Inference Systems (ANFIS). Genetic Algorithms (GA) with a Back-propagation Neural Network were also applied (Karegowda *et al.*, 2011). It is important to note that the majority of these models were applied to the Pima Indian Diabetes Data (PIDD) (Lichman, 2013) and used small datasets that had no temporal information with a small number of features. (Ali *et al.*, 2014) developed a model for classifying diabetic patient control level based on historical medical records and using used Data Mining.

The techniques used for data mining in this study are Naïve Bayes, Logistic and J48, and for implementing it, the WEKA application was in use (Mohammed and Lomte, 2020).

The result in the following table showed the average, recall, F-measure and accuracy of the three techniques.

As it is noticed in the above comparison, the logistic algorithm was more accurate than the other two.

In Sudan, a new model was developed by (Naser *et al.*, 2015). It was used to classify diabetes type II treatment plans. In his study, the J48 algorithm was applied on 318 medical records of diabetes patients. The basic control information showed the following results, as in the following Table 2.

WEKA application was used for evaluation. The research work did not consider diabetes type I patients, which could have been included with additional attributes (Meng *et al.*, 2013). Also, the nutrition system and exercise could have been included to increase the accuracy of the system.

We proposed an ANN model, which uses a Levenberg-Marquardt backpropagation algorithm to predict diabetes that can be useful and helpful for doctors and practitioners. In this research, we used the following attributes: Number of pregnancies, PG Concentration (Plasma glucose at 2 hours in an oral glucose tolerance test), Diastolic BP (Diastolic Blood Pressure (mm Hg))), Tri-Fold Thick (Triceps Skin Fold Thickness (mm)), Serum Ins (2-Hour Serum Insulin (mu U/ml)), BMI (Body Mass Index: (weight in kg/ (height in m)^2)), DP Function (Diabetes Pedigree Function), Age (years), Diabetes (Whether or not the person has diabetes) see Table 3.

OBJECTIVES

1. Prediction and categorization of health status.
2. Identify some appropriate factors affecting health conditions.
3. Design an artificial neural network that can be used to predict health performance based on some predefined data for a particular health condition.

Table 1: The performance of the Naïve Bayes, Logistic and J48 techniques

Techniques	Average	recall	F-measure	accuracy
Logistic	0.73	0.744	0.653	74.4%
Naïve Bayes	0.717	0.742	0.653	74.2%
J48	0.54	0.735	0.623	73.5%

Table 2: Classification of diabetes type II using the J48 algorithm

Type	Oral Hypoglycemic	Insulin	Diet
Rates of record	59.1%	35.5%	5.3%

Table 3: Comparison of diabetes diagnostic accuracy- Pima dataset

Reference	Proposed model / Method	Accuracy (%)
(Bagrecha *et al.*, 2019)	convolution neural network	84
(Kaur and Kumari, 2018)	Radial Basis Kernel SVM	84
	k-Nearest Neighbors	88
	Artificial Neural Network	86
(Massaro *et al.*, 2019)	Long Short-Term Memory neural network	84
(Kapoor and Krishna, 2018)	K-Nearest Neighbors	83.12
(Aminul and Jahan, 2017)	Super Vector Machine (SVM)	77.08
(Mortajez and Jamshidinezhad, 2019)	Artificial Neural Network	84.5
	Proposed Method	88.8

METHODOLOGY

The following subsections show the process of diabetes II prediction,

Procedures

For this study, data were collected from the PIMA Indian data set and select suitable attributes from diabetic patients. It is extracted and applied to the models of a feed-forward network using the selected data. The data were entered into the feed-forward network, determined the value of each of the variables using a feed-forward network (the most influential factor on diabetes), then the data were trained, validated, and tested see Figure 1.

Dataset and Analysis

The sample was collected from 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 aged 21 years, 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. 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 preprocessing task was to normalize the data between -1 and +1.

In Table 2 para-clinical features were used as input features to an ANN, while as output feature of the health status based on the condition of diabetes of the patients was used. Input Features are,

The Output Variable

The output variable represents whether a person has diabetes or not (Sick, Healthy).

Table 7 shows the classification of the selected output variable, which is consistent with the classification system in the identification of disease cases.

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)

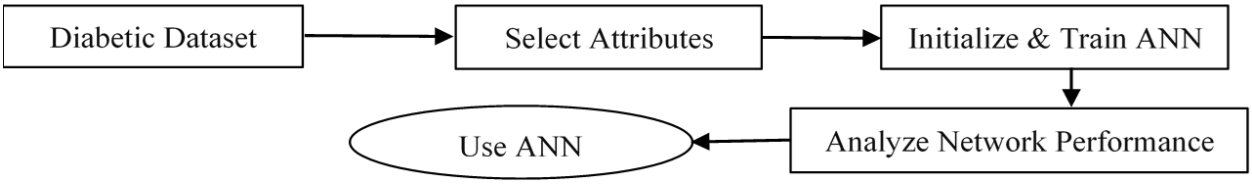

Figure 1: Proposed Workflow

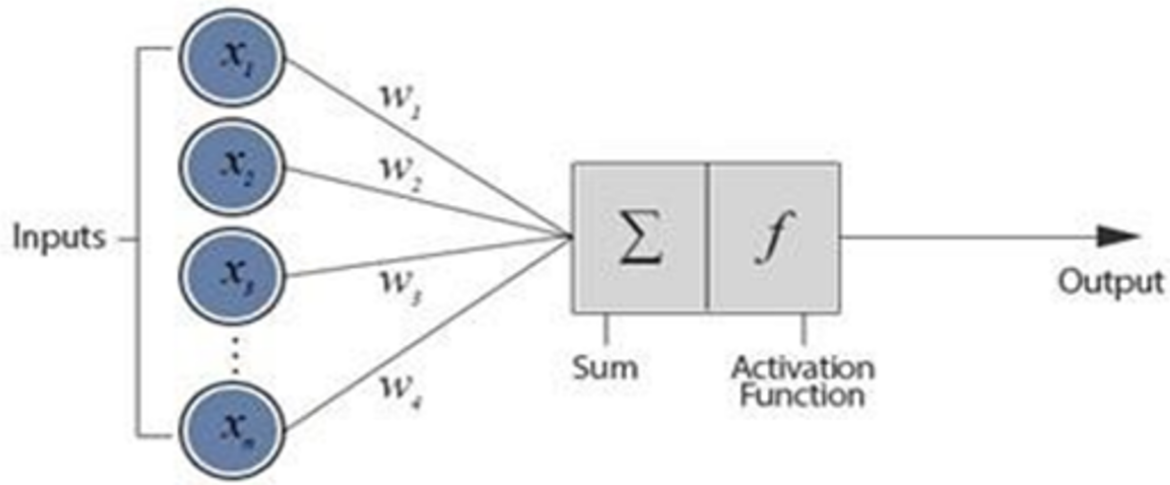

Figure 2: Neural Network

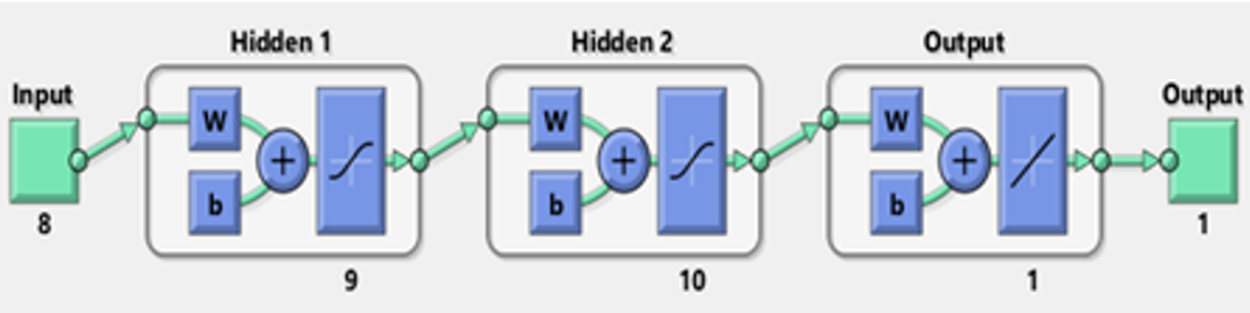

Figure 3: Feedforward network

Table 4: Attributes in the Data Set

Sl	Risk factors		Range
	Attribute	Description	
1	Pregnancy	Frequency	0-17
2	Plasma glucose Concentration	2 hours in an oral test (mm Hg)	0-199
3	Diastolic BP	(mm Hg)	0-122
4	Triceps SFT	(mm)	0-99
5	Serum-Insulin	2-Hour serum insulin (mu U/ml)	0-846
6	BMI: Body Mass Index	(how much does the body weigh in kg/ (how high is it in m)^2)	0-67.1
7	DP Function	Diabetes pedigree function	0.078 –2.42
8	Age	Age (years)	21-81
9	Class	Diabetes class variable	0-1

Table 5: Class Distribution

Class value	No. of instances
0	500
1	268

Table 6: 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

Table 7: Output Data Transformation

S/N	Output Variable	Diabetes
1	Healthy "1 "	The person does not have diabetes
2	Sick "0"	The person has diabetes

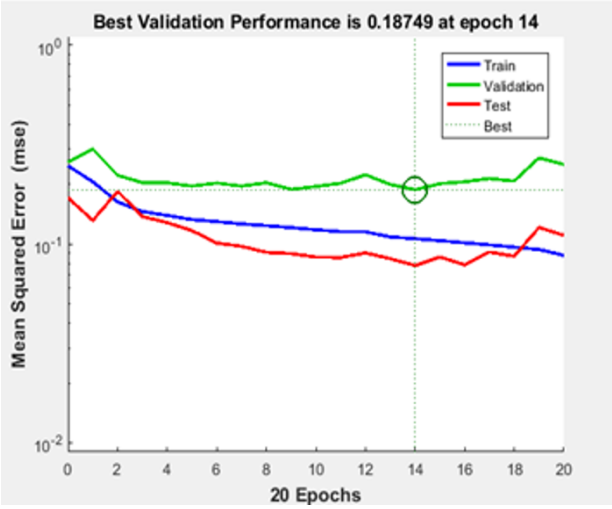

Figure 4: Performance graph

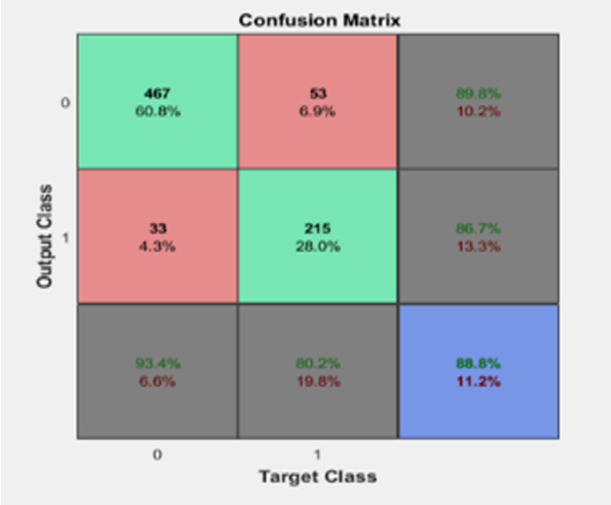

Figure 5: Confusion Matrices

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 (see Figure 2).

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

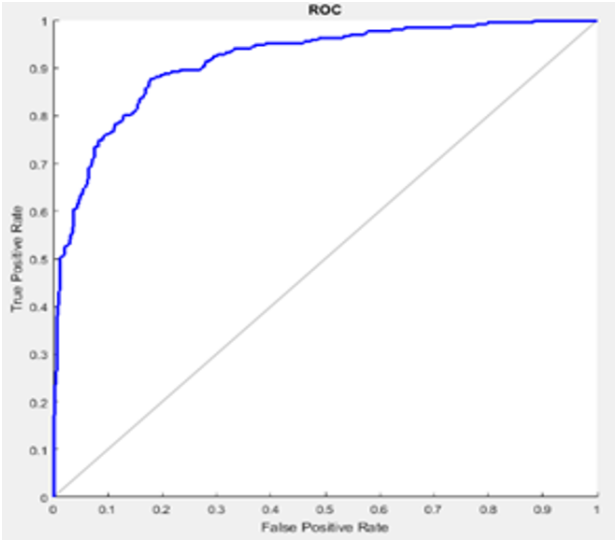

Figure 6: ROC Curves

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 preprocessing enables the data to be ready and the training to be more efficient. During this data preprocessing process, data must be divided into three different sets, namely for training, validation, and testing purposes. 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 (Mohammed *et al.*, 2020). 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.

Network Architecture

A multi-layered feed-forward network with 8 input nodes Figure 3, two hidden layers that composed of 9, 10 nodes respectively 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 backpropagation 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 (Venkatesan and Anitha, 2006). 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.

RESULTS AND DISCUSSION

The experimental analysis was done using MATLAB R2016a with the neural network toolbox to implement the proposed algorithm. For this, 768 records of the dataset were imported and divided into training, validation, and testing datasets (90%, 5%, and 5%, respectively). Input data (input weight) were normalized to be transformed into the range -1 to +1 before use (from Tables 4, 5, 6 and 7). 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.

Figure 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. 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. The following matrix has columns and rows, in which the former was representing the target class and the later was representing the output class. Correctly classified inputs are shown

along the diagonal of the matrix, and off-diagonal cells indicate misclassified inputs

The confusion matrix in Figure 5 offers that the network achieved 88.8% accuracy.

That is, 467 samples are correctly classified as non-diabetics. This corresponds to 60.8% of all 768 samples. Similarly, 215 samples are correctly classified as diabetic. This corresponds to 28.0% of all samples.53 of the non-diabetics are incorrectly classified as diabetic and this corresponds to 6.9% of all 768 samples in the data. Similarly, 33 of the diabetics are incorrectly classified as non-diabetics and this corresponds to 4.3% of all data.

Out of 520 non-diabetics predictions, 89.8% are correct and 10.2% are wrong. Out of 248 diabetics predictions, 86.7% are correct and 13.3% are wrong. Out of 500 the non-diabetics, 93.4% are correctly predicted as non-diabetics and 6.6% are predicted as diabetics. Out of 268 diabetics, 80.2% are correctly classified as diabetics and 19.8% are classified as non-diabetics.

Overall, 88.8% of the predictions are correct and 11.2% are wrong classifications.

The ROC curves in Figure 6 depicted that the training ROC curve was closer to the idle ROC and were identical. Figures 4 and 5 and Figure 6 show that the proposed model gave better prediction accuracy, suggesting its applicability to DM prediction.

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. Figure 6 shows ROC curves for different datasets.

CONCLUSIONS

In this paper, the diabetes was predicted using the artificial neural network model. By using the ANN model, we can use software to design and implement complex medical processes. It can be used in predicting, diagnosing, treating and helping the surgeons, physicians, and the general population. These systems can be implemented in a parallel way and are distributed in different measures. In general, an artificial neural network is a parallel processing system that is used to detect complex patterns in the data. This study aimed to determine the impact of some effective variables on diabetes. The proposed model was implemented in MATLAB R2016a. The diabetes dataset contains 768 samples with 8 attributes. This model was first used to determine the value of each of the variables using a feed-forward network (the most influential factor on diabetes). After training, validating, and testing the dataset, we got (88.8%) accuracy, the average error was (0.187), number of epochs was (14).

REFERENCES

Ali, R., Siddiqi, M. H., Idris, M., Kang, B. H., Lee, S. 2014. Prediction of Diabetes Mellitus Based on Boosting Ensemble Modeling BT - Ubiquitous Computing and Ambient Intelligence. Personalization and User Adapted Services. Cham. Springer International Publishing.

Aminul, M., Jahan, N. 2017. Prediction of Onset Diabetes using Machine Learning Techniques. International Journal of Computer Applications, 180(5):7–11.

Bagrecha, J. N., Chaithra, G. S., Jeevitha, S. 2019. Diabetes Disease Prediction using Neural Network. International Journal for Research in Applied Science & Engineering Technology, 7:3888–3893.

Chavey, A., Kioon, M., Bailbé, D. 2014. programming of beta-cell disorders and intergenerational risk of type 2 diabetes Diabetes. Maternal Diabetes, 40(5):323–353.

Elzamly, A., Naser, S. A., Hussin, B., Doheir, M. 2015. Predicting Software Analysis Process Risks Using Linear Stepwise Discriminant Analysis: Statistical Methods. 38:108–115.

Kapoor, S., Krishna, P. S. 2018. Optimizing Hyper Parameters for Improved Diabetes Prediction. International Research Journal of Engineering and Technology (IRJET), (05):5–5.

Karegowda, A. G., Manjunath, A., Jayaram, M. 2011. Application of Genetic Algorithm Optimized Neural Network Connection Weights for Medical Diagnosis of PIMA Indians Diabetes. International Journal on Soft Computing, 2(2):15–23.

Kaur, H., Kumari, V. 2018. Predictive modelling and analytics for diabetes using a machine learning approach. Applied Computing and Informatics. Pp, pages 1–5.

Lichman, M. 2013. UCI machine learning repository. Center for Machine Learning and Intelligent Systems.

Manzella, D., Grella, R., Abbatecola, A. M., Paolisso, G. 2005. Repaglinide Administration Improves Brachial Reactivity in Type 2 Diabetic Patients. Diabetes Care, 28(2):366–371.

Massaro, A., Maritati, V., Giannone, D., Convertini, D., Galiano, A. 2019. LSTM DSS Automatism and Dataset Optimization for Diabetes Prediction. Applied Sciences, 9(17):3532–3532.

Meng, X. H., Huang, Y. X., Rao, D. P., Zhang, Q., Liu,

Q. 2013. Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. The Kaohsiung Journal of Medical Sciences, 29(2):93–99.

Mohammed, N. M., Lomte, S. S. 2020. Secure and Efficient Outsourcing of Large Scale Linear Fractional Programming. In *Computing in Engineering and Technology*, pages 277–286. Springer.

Mohammed, N. M., Sultan, L. R., Hamoud, A. A., Lomte, S. S. 2020. Verifiable, secure computation of linear fractional programming using certificate validation. International Journal of Power Electronics and Drive Systems, (1):11–11.

Mohammed, N. M., Sultan, L. R., Lomte, S. S. 2019. Privacy-preserving outsourcing algorithm for two-point linear boundary value problems. Indonesian Journal of Electrical Engineering and Computer Science, 16(2):1065–1069.

Mortajez, S., Jamshidinezhad, A. 2019. An Artificial Neural Network Model to Diagnosis of Type II. Diabetes Journal of Research in Medical and Dental Science, 7(1):66–70.

Morteza, A., Nakhjavani, M., Asgarani, F., Carvalho, F. L. F., Karimi, R., Esteghamati, A. 2013. Inconsistency in albuminuria predictors in type 2 diabetes: a comparison between neural network and conditional logistic regression. Translational Research, 161(5):397–405.

Motka, R., Parmarl, V., Kumar, B., Verma, A. R. 2013. Diabetes mellitus forecasts using different data mining techniques. 4th International Conference on Computer and Communication Technology (ICCCT), pages 99–103.

Naser, S. A., Zaqout, I., Ghosh, M. A., Atallah, R., Alajrami, E. 2015. Predicting Student Performance Using Artificial Neural Network: in the Faculty of Engineering and Information Technology. International Journal of Hybrid Information Technology, 8(2):221–228.

Polat, K., Güneş, S. 2007. An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to a diagnosis of diabetes disease. Digital Signal Processing, 17(4):702–710.

Temurtas, H., Yumusak, N., Temurtas, F. 2009. A comparative study on diabetes disease diagnosis using neural networks. Expert Systems with Applications, 36(4):8610–8615.

Venkatesan, P., Anitha, S. 2006. Application of a radial basis function neural network for the diagnosis of diabetes mellitus. Current Science, 91(9):1195–1199.

