Machine Learning Research
2019; 4(2): 33-38
http://www.sciencepublishinggroup.com/j/mlr
doi: 10.11648/j.mlr.20190402.12
ISSN: 2637-5672 (Print); ISSN: 2637-5680 (Online)

Predicting Diabetes Mellitus Using Artificial Neural Network Through a Simulation Study

Shehu Usman Gulumbe[1], Shamsuddeen Suleiman[1], Shehu Badamasi[1], Ahmad Yusuf Tambuwal[2], Umar Usman[1]

[1]Department of Mathematics Usmanu Danfodiyo University, Sokoto, Nigeria

[2]Department of Information and Communication Technology, Usmanu Danfodiyo University, Sokoto, Nigeria

Email address:
suleman.shamsuddeen@udusok.edu.ng (S. Suleiman)

To cite this article:
Shehu Usman Gulumbe, Shamsuddeen Suleiman, Shehu Badamasi, Ahmad Yusuf Tambuwal, Umar Usman. Predicting Diabetes Mellitus Using Artificial Neural Network Through a Simulation Study. *Machine Learning Research.* Vol. 4, No. 2, 2019, pp. 33-38.
doi: 10.11648/j.mlr.20190402.12

Received: July 24, 2019; **Accepted**: August 16, 2019; **Published**: September 2, 2019

Abstract: Diabetes mellitus (DM) is a diverse group of metabolic disorders that is frequently associated with a high disease burden in developing countries such as Nigeria. It also needs continuous blood glucose monitoring and self-management. This research is aimed to predict diabetes mellitus using artificial neural network. In this research, 100 patients were considered from Ahmadu Bello University Teaching Hospital who have undergone diabetes screening test and 29 risk factors were used. Back propagation algorithm was used to train the artificial neural network for the original and simulated data sets. The results show that the models achieved 98.7%, 57.0%, 73.3%, and 63.0% accuracy for training the original, simulated at 100, simulated at 150 and simulated at 200 data sets respectively. The results also shows that the areas covered under receiver operating curves are 0.997, 0.587, 0.849 and 0.706 for training the original, simulated at 100, simulated at 150 and simulated at 200 data sets respectively. The research therefore concludes that in order to predict diabetes mellitus in patients, the simulated data can be used in place of the original data since the simulated ANN models have been able to discriminate between diabetic and non-diabetic patients.

Keywords: Diabetes Mellitus, Backpropagation, Simulation, Prediction, Artificial Neural Network

1. Introduction

The World Health Organization (WHO, 2011) defines diabetes as a chronic disease that occurs either when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin it produces. Diabetes cases are on the increase all over the world and countries are struggling to fight the disease. The misconception that diabetes is "a disease of the wealthy" is still held by some people; but the evidence published in the Diabetes Atlas of the International Diabetes Federation (IDF, 2013) disproves that delusion: a staggering 80% of people with diabetes live in low and middle-income countries, and the socially disadvantaged in any country are the most vulnerable to that illness.

Today's emerging diabetes hotspots include countries in the Middle East, Western Pacific, sub-Saharan Africa and South-East Asia where economic development has transformed lifestyles. These rapid transitions are bringing previously unheard rates of obesity and diabetes; developing countries are facing a firestorm of ill health with inadequate resources to protect their population. Thus, it is necessary to increase awareness of the importance of a healthful diet and physical activity, especially for children and adolescents. Crucially though, environments have to be created that lay the foundations for healthy living. (IDF, 2013)

Nigeria has the largest population in Africa (about 170 million); and of this the adult population (aged 20–79 years), is approximately 79 million. One third of all the cases of diabetes are in the rural communities, while the rest are in the urban centres. About two million of the cases of diabetes in Nigeria are undiagnosed. Deaths related to diabetes in Nigeria in 2013 were estimated to be 105,091 cases [1].

Today, modern techniques such as Artificial Intelligence (AI) and data mining have been used by several researchers

to predict diabetes [2-4].

Simulations are the attempt to adjust real life situations in a controllable environment, so that results can be reproduced. Thus, if simulation methods are applied, it is the target to reconstruct real life situations, but the starting points can be different. If it is not exactly clear which starting point is the most realistic one, therefore a scenario analysis might be helpful. Although one special scenario can be very unlikely, it is sometimes useful to know what can happen in an extreme situation. The sample is drawn only once but there are many values to be estimated. Without a simulation it is not possible to evaluate the quality of an estimator.

Simulation studies are computer experiments which involve creating data by pseudo random sampling from known probability distributions. They are an invaluable tool for statistical research, particularly for the evaluation of new methods and the comparison of competing methods. Simulation studies are much used in the pages of Statistics in Medicine, but our experience is that many statisticians lack the necessary understanding to execute a simulation study with confidence [5]. Proper understanding of simulation studies would enable such people to run simulation studies themselves and to critically appraise published simulation studies. But, usually design and reporting issues lead to uncritical use or appraisal of simulation studies. In this context, better understanding of the rationale, design, execution, analysis and reporting of simulation studies is necessary to improve what researchers can learn from them. Simulation studies are used to obtain empirical results about the behaviour of statistical methods in certain scenarios, as opposed to analytic results, which may cover many settings. It is not always possible to obtain analytic results, or may be extremely difficult. Simulation studies come in to their own when methods make wrong assumptions or data are messy; that is, they can assess the resilience of methods. This is not always possible with analytic results, which may assume that data arise from a specific model.

Statistical Neural Network is a non-parametric method that can be used in the medical field to classify subjects based on input variables into sick or healthy. Classification and prediction of the patient's condition based on risk factors are an application of artificial neural networks [6].

Statistical neural networks mimic mixed structure of the human brain. Billion nerve cells (neurons) through the communication that with each other (synapses) creates a biological neural network in the human brain that is dedicated to human activities such as reading, comprehension, speaking, breathing, movement, voice recognition, face detection, also resolve issues and data storage. Artificial neural networks, in fact, simulate a part of brain functions [5, 7].

The study [8] predicted the onset of diabetes disease using Artificial Neural Network (ANN). In this paper they experimented and suggested an Artificial Neural Network (ANN) based classification model as one of the powerful method in intelligent field for classifying diabetic patients into two classes. For achieving better results, genetic algorithm (GA) is used for feature selection. The GA is used for optimally finding out the number of neurons in the single hidden layered model. Further, the model is trained with Back Propagation (BP) algorithm and GA (Genetic Algorithm) and classification accuracies are compared. The designed models are also compared with the Functional Link ANN (FLANN) and several classification systems like NN (nearest neighbor), kNN (k-nearest neighbor), BSS (nearest neighbor with backward sequential selection of feature, MFS1 (multiple feature subset), MFS2 (multiple feature subset) for Data classification accuracies. It is revealed from the simulation that our suggested model is performing better compared to NN (nearest neighbor), kNN (k-nearest neighbor), BSS (nearest neighbor with backward sequential selection of feature, MFS1 (multiple feature subset), MFS2 (multiple feature subset) and FLANN model and it can be a very good candidate for many real time domain applications as these are simple with good performances.

The researcher [9] predicted the onset of diabetes using machine learning. *The paper uses classification techniques, like logistic regression to predict the disease in its early stages*. The result shows that the simple anthropometric variables like waist circumference are better predictors, than the three month average blood glucose level.

The paper [10] review of model prediction in diabetes and of designing glucose regulators based on model predictive control for the artificial pancreas. The present work presents a comparative assessment of glucose prediction models for diabetic patients using data from sensors monitoring blood glucose concentration as well as data from in silico simulations. The models are based on neural networks and linear and nonlinear mathematical models evaluated for prediction horizons ranging from 5 to 120min. Several devices record blood glucose levels and insulin doses are used and there are many approaches to optimizing their implementation. The clinical studies focused on recording carbohydrates intake, physical activities and some recorded the psychological state of patients. They concluded that that models developed with exogenous inputs are more reliable than simpler models while the combination of prediction models with compartment models provide better estimations for short term predictions.

The research [11] predicted diabetes diagnosis using classification based data mining techniques, they used binary logistic regression, multilayer perceptron and k-nearest neighbor as classification for diabetes data. They compared classification accuracy for classifying data. They fund that the binary logistic regression accuracy is 0.69, multilayer perceptron accuracy is 0.71 and KNN gives the accuracy of 0.80 with shows k-nearest neighbour accuracy is higher than that of binary logistic regression and multilayer perceptron.

The author [12] establish an appropriate prediction model based on data mining techniques for predicting type 2 diabetes mellitus (t2dm). were the the accuracy of the prediction model were improved to make the model adaptive to more than one dataset. The model comprised of two parts, the improved k-means algorithm and the logistic regression

algorithm. It utilized Pima Indians diabetes dataset and the Waikato environment for knowledge analysis toolkit to compare the results with the results from other researchers. It shows that the model attained a 3.04% higher accuracy of prediction than those of other researcher. the model ensures that the dataset quality is sufficient. The model was applied to two other diabetes data set, with shows that the model is useful for realistic health management of diabetes.

The study [13] predicted diabetes mellitus with machine learning techniques In this study, they used decision tree, random forest and neural network to predict diabetes mellitus. The dataset is the hospital physical examination data in Luzhou, China. It contains 14 attributes. In this study, five-fold cross validation was used to examine the models. In order to verity the universal applicability of the methods, they chose some methods that have the better performance to conduct independent test experiments. They randomly selected 68994 healthy people and diabetic patients' data, respectively as training set. Due to the data unbalance, we randomly extracted 5 times data. And the result is the average of these five experiments. In this study, we used principal component analysis (PCA) and minimum redundancy maximum relevance (mRMR) to reduce the dimensionality. The results showed that prediction with random forest could reach the highest accuracy (ACC=0.8084) when all the attributes were used.

The author [14] used an artificial neural network model to diagnosis of Type II diabetes. In this study, a hybrid ANN-Genetic Algorithm model was developed for classification of diabetic patients. Therefore, the number of optimal neurons as well as hidden layers was determined to design the architecture of the ANN model. To reduce the mean square error of the MSE network and optimize the accuracy of the diagnostic system a Genetic Algorithm (GA) was combined with the proposed ANN model. For experiment process, the model was considered on a dataset included 768 samples to diagnose the patients with type II diabetes from other cases. The results showed a precision of 85% for diagnosing of type-2 diabetic patients. The proposed structure based on the lower mean square error of the MSE, indicated the best performance of the ANN with the MSE rate of 0.155. The developed intelligent model showed an effective performance in comparison with existing methods with a minimum error and maximum confidence in the diagnosis process of diabetic disease.

The paper [15] predicted diabetes based on Artificial Intelligence Technique. The objective of this paper is to develop a system that can offer a precise early prediction of diabetes for a patient with the help of artificial intelligence technique. The datasets consist of several medical predictor variables and one target variable. Independent variables include the Body Mass Index (BMI), insulin level, age, number of pregnancies the patient had (for female) and some others. Based on these parameters, prediction of diabetes, applying artificial intelligence technique, described in this article, seems quite satisfactory.

2. Multilayer Networks

Multilayer networks are universal approximators; the training of such networks means determining a procedure for selecting the network parameters (weights and biases) which will best approximate a given function. The procedure for selecting the parameters for a given problem is called *training* the network. In this research a training procedure called *Backpropagation*, which is based on gradient descent is used.

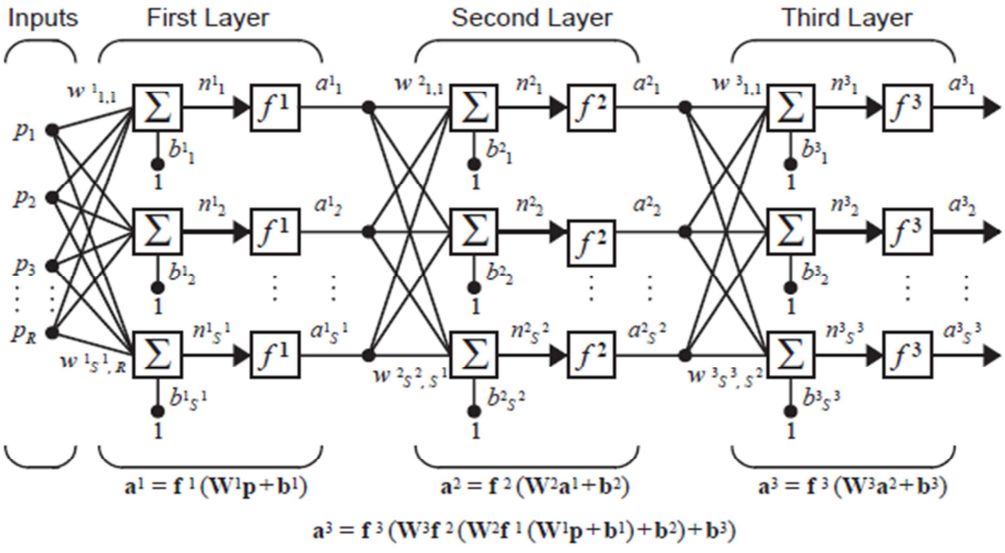

Figure 1. *Three layer Network.*

In multilayer networks the output of one layer becomes the input to the following layer. The equations that describe this operation are as follows:

$$a^{m+1} = f^{m+1}(w^{m+1}a^m + b^{m+1}) \quad (1)$$

For m=0, 1, ----, m-1

Where M is the number of layers in the network The neurons in the first layer receive external inputs:

$$a^0 = p \quad (2)$$

The outputs of the neurons in the last layer are considered the network outputs:

$$a = a^m \tag{3}$$

The algorithm should adjust the network parameters in order to minimize the sum squared error:

$$F(x) = \sum_{q=1}^{Q} e_q^2 = \sum_{q=1}^{Q} (t_q - a_q)^2 \tag{4}$$

Where x is a vector containing all of network weights and biases. If the network has multiple outputs.

Since the performance index in (4) is sum of squares of nonlinear function, the numerical optimization techniques for nonlinear least squares can be used to minimize this cost function. The Levenberg-Marquardt algorithm, which is an approximation to the Newton's method is said to be more efficient in comparison to other methods for convergence of the Backpropagation algorithm for training a moderate-sized feed forward neural network [2]. As the cost function is a sum of squares of nonlinear function, the Hessian matrix required for updating the weights and biases need not be calculated and can be approximated as:

$$H = J^T(x)J(x) \tag{5}$$

The updated weights and biases are given by:

$$x_{k+1} = x_k - [J^T(x)J(x) + \mu I]^{-1} J^T(x)e(x) \tag{6}$$

Where μ is a scalar and I is the identity matrix.

3. Results and Discussion

The neural network model was trained using Levenberg-Marquardt (LM) training algorithms. An Intel (R) Core (TM) i32310M CPU @ 2.10GHz processor was used to train the neural network model. Figure 2 is the architectural design of how neural network is trained. In order to train, validate and test the neural networks developed using the LM algorithms, we have divided the data set in the following way: 70% of it for the training process, 15% for the validation process and the remaining 15% for the testing process. In all the cases, the samples have been randomly chosen as to cover the specified percentages. In order to train the neural networks, we have used the mean square error (MSE) as an objective function. When training a network with this function, if there are multiple outputs having different ranges of values, the accuracy is optimized for the output element that has a wider range of values and is less optimized relative to the output element with a smaller range of values. Thus, the network will learn to fit the first output element very well, while the second output element is not fit as accurate as the first. In order to solve this issue, we have normalized the errors, by setting the normalization performance parameter to its 'standard' value. By using this method, the errors have been computed as if both of the output elements had values ranging from -1 to 1 and consequently, the two output elements have been fitted very well.

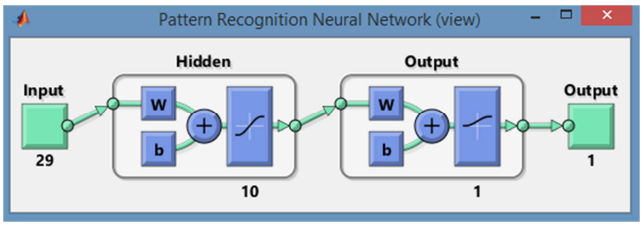

Figure 2. *Pattern Recognition Neural Network (view).*

Table 1. *Cofusion Matrix for for the original data Actual Classifiaction.*

Predicted classifaction	Positive (Diabetic patients)	Negative (Non-diabetic patients)	Total
Positive (Diabetic patients)	TP=182	FP=0	182
Negative (Non-diabetic patients)	FN=0	TN=93	93
Total	182	93	275

Table 2. *Cofusion Matrix for Simulation At N=100 Actual Classifiaction.*

Predicted classifaction	Positive (Diabetic patients)	Negative (Non-diabetic patients)	Total
Positive (Diabetic patients)	TP=33	FP=22	55
Negative (Non-diabetic patients)	FN=21	TN=24	45
Total	54	46	100

Table 3. *Cofusion Matrix for Simulated data at N=150 Actual Classifiaction.*

Predicted classifaction	Positive (Diabetic patients)	Negative (Non-diabetic patients)	Total
Positive (Diabetic patients)	TP=60	FP=24	84
Negative (Non-diabetic patients)	FN=16	TN=50	66
Total	76	74	150

Table 4. *Cofusion Matrix Simulated data at N=200 Actual Classifiaction.*

Predicted classifaction	Positive (Diabetic patients)	Negative (Non-diabetic patients)	Total
Positive (Diabetic patients)	TP=75	FP=42	117
Negative (Non-diabetic patients)	FN=32	TN=51	83
Total	107	93	200

Table 1 represents the confusion matrix for the original data. It indicates that 182 diabetic patients were correctly

classified by the neural network classifier; none of the diabetic patients was misclassified by the neural network classifier and also none of the diabetic patients was misclassified by the neural network classifier. Finally 93 non diabetic patients were correctly classified by the neural network classified.

Table 2 represents the confusion matrix for simulation at n=100. It indicates that 33 diabetic patients were correctly classified by the neural network classifier; 22 of the diabetic patients were misclassified by the neural network classifier and also 21 of the diabetic patients were misclassified by the neural network classifier. Finally 24 non diabetic patients were correctly classified by the neural network classified.

Table 3 represents the confusion matrix for simulation at n=150. It indicates that 60 diabetic patients were correctly classified by the neural network classifier; 24 of the diabetic patients were misclassified by the neural network classifier and also 16 of the diabetic patients was misclassified by the neural network classifier. Finally 50 non diabetic patients were correctly classified by the neural network classified.

Table 4 represents the confusion matrix for simulation at n=200. It indicates that 75 diabetic patients were correctly classified by the neural network classifier; 42 of the diabetic patients were misclassified by the neural network classifier and also 32 of the diabetic patients were misclassified by the neural network classifier. Finally 51 non diabetic patients were correctly classified by the neural network classified.

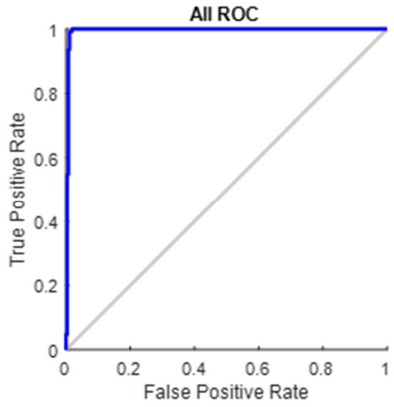

Figure 3. *ROC of the original data.*

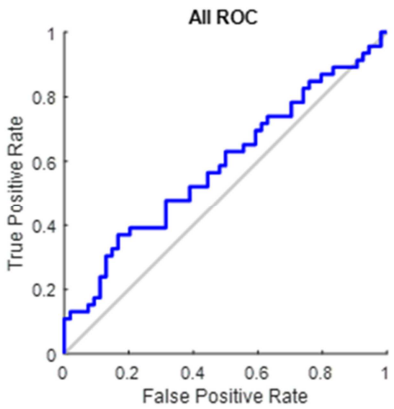

Figure 4. *ROC of the simulated data at n=100.*

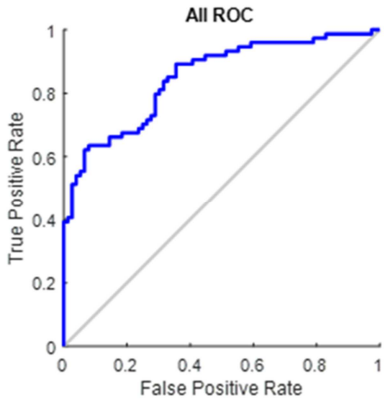

Figure 5. *ROC of the simulated data at n=150.*

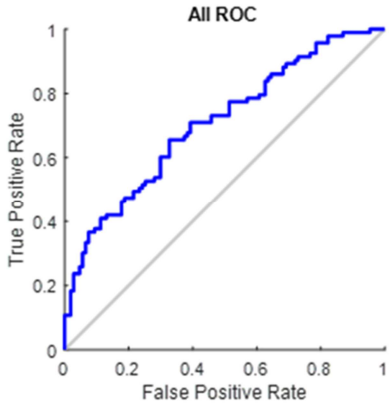

Figure 6. *ROC of the simulated data at n=200.*

Table 5. *Performance Metrics.*

Data	Indices	
Original	Accuracy	98.7%
	Sensitivity	98.1%
	Specificity	100%
	Positive Predictive Value	100%
	Negative Predictive Value	96.2%
Simulated at 100	Accuracy	57.0%
	Sensitivity	52.0%
	Specificity	61.1%
	Positive Predictive Value	60.0%
	Negative Predictive Value	53.3%
Simulated at 150	Accuracy	73.3%
	Sensitivity	54.8%
	Specificity	78.9%
	Positive Predictive Value	71.4%
	Negative Predictive Value	75.8%
Simulated at 200	Accuracy	63.0%
	Sensitivity	54.8%
	Specificity	70.1%
	Positive Predictive Value	64.1%
	Negative Predictive Value	61.4%

The diagonal joining the point (0, 0) to (1, 1) divides the square in two equal parts and each has an area equal to 0.5. When ROC is this line, overall there is 50-50 chances that test will correctly discriminate the diabetic and non-diabetic subjects. The minimum value of AUC should be considered 0.5 instead of 0 because AUC=0 means test incorrectly classified all subjects with disease as negative and all non-disease subjects as positive. If the test results are reversed

then area=0 is transformed to area=1; thus a perfectly inaccurate test can be transformed into a perfectly accurate test. It is very clear that all the ROCs in figures 3 to 6 have shown that the area covered are far greater than 0.5 because the lines of the curves are at the top left of the graphs. The areas covered by these figures are also shown in table 6 below. This therefore suggests that the models have been able to discriminate between the diabetic and non-diabetic patients.

Table 6. *Area under receiver operating curve (AUC) of the models.*

Data	AUC
Original	0.997
Simulated at 100	0.587
Simulated at 250	0.849
Simulated at 200	0.706

4. Conclusion

The automatic diagnosis of diabetes is an important real-world medical problem. Detection of diabetes in its early stages is the key for treatment. This study shows how artificial neural network is used to predict actual diagnosis of diabetes patients visiting Ahmadu Bello University Teaching Hospital (ABUTH) Zaria for local and systematic treatment, along with presenting related work in the field. The original data sets generated from ABUTH Zaria was simulated at 100, 150 and 200 sample sizes. In this paper, Artificial Neural Network (ANN) models were trained using the Liebenberg-Marquardt algorithm for both the original and simulated data sets. The results show that the models achieved 98.7%, 57.0%, 73.3%, and 63.0% accuracy for training the original, simulated at 100, simulated at 150 and simulated at 200 data sets respectively. The results also shows that the areas covered under receiver operating curves are 0.997, 0.587, 0.849 and 0.706 for training the original, simulated at 100, simulated at 150 and simulated at 200 data sets respectively. The research therefore concludes that in order to predict diabetes mellitus in patients, the simulated data can be used in place of the original data since the simulated ANN models have been able to discriminate between diabetic and non-diabetic patients.

References

[1] Oputa R. N. and Chinenye S., (2015). Diabetes in Nigeria – a translational medicine approach. African Journal of Diabetes Medicine, 23 (1), 7-10.

[2] Indoria P. and Rathore Y. K., (2018). ASurvey: Detection and Prediction of Diabetes Using Machine Learning Techniques, International Journal of Engineering Research & Technology (IJERT), 7 (3), 287-291.

[3] Mamman M. and Saratha S. (2017), Predicting the survival of diabetes using neural network, AIP Conference Proceedings 1870, 040046; doi: 10.1063/1.4995878.

[4] Natchiar S. U. and Baulkani S. (2018). Review of Diabetes Disease Diagnosis Using Data Mining and Soft Computing Techniques, International Journal of Pure and Applied Mathematics, 118 (10), 137-142.

[5] Hagan, M. T., & Menhaj, M., (1994) "Training feed-forward networks with the Marquardt algorithm", IEEE Trans. Neural Networks, Vol. 5, No. 6, pp 989-993.

[6] Livingstone, D., Totowa, NJ (2008). Artificial Neural Networks Methods and Application. 1st ed.: Hummana Pres.

[7] Dunne, RA., Wiley, J., Inc, S. (2007) "A Statistical Approach to Neural Networks for Pattern Recognition", New Jersey: John Wiley & Sons Inc.

[8] Pradhan M. and Sahu R. K., (2011) Predict the onset of diabetes disease using Artificial Neural Network (ANN), International Journal of Computer Science & Emerging Technologies, 2 (2), 303-311.

[9] Mishra V., Samuel C. and Sharma S. K. (2015). Use of Machine Learning to Predict the Onset of Diabetes. International Journal of Recent advances in Mechanical Engineering, 4 (2), 9-14.

[10] Saiti K., Macas M., Stechova K., Pithova P., and Lhotska L. (2017). A review of model prediction in diabetes and of designing glucose regulators based on model predictive control for the artificial pancreas, Biomedical Engineering Education, 2-4.

[11] Selvakumar S., Kannan K. S. and Nachiyar S. G. (2017). Prediction of Diabetes Diagnosis Using Classification Based Data Mining Techniques, International Journal of Statistics and Systems, 12 (2), 183-188.

[12] Wu H., Yang S., Huang Z., He J., and Wang X. (2018). Type 2 diabetes mellitus prediction model based on data mining, Informatics in Medicine Unlocked, 10, 100-107.

[13] Quan Z., Kaiyang Q., Yamei L., Dehui Y., Ying J., and Hua T. (2018). Predicting Diabetes Mellitus With Machine Learning Techniques, Frontiers in genetics, 9, 1-10.

[14] Sareh M. and Amir J. (2019). An Artificial Neural Network Model to Diagnosis of Type II Diabetes, Journal of Research in Medical and Dental Science, 7 (1), 66-70.

[15] Shameem H. (2018). Prediction of Diabetes Based on Artificial Intelligence Technique, International Research Journal of Engineering and Technology (IRJET), 5 (11), 11-15.