Diabetes Mellitus Prediction using Hybrid Artificial Neural Network

Banibrata Paul
Department of Computer Science & Engineering
Birla Institute of Technology
Mesra, Ranchi, Jharkhand, India
Email: paul.banibrata1@gmail.com

Dr. Bhaskar Karn
Department of Computer Science & Engineering
Birla Institute of Technology
Mesra, Ranchi, Jharkhand, India
Email: bhaskar@bitmesra.ac.in

2021 IEEE Bombay Section Signature Conference (IBSSC) | 978-1-6654-1758-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/IBSSC53889.2021.9673397

***Abstract*— Diabetes mellitus is prominent diseases throughout the world. As human society has become economically and socially isolated, the pervasiveness of diabetes is rapidly expanding. It is the metabolic diseases where a patient has high blood sugar either due to the failure of the body to produce adequate insulin or the failure of the cells to respond to the produced insulin. It can be identified by studying on several readings taken from a patient such as albumin, creatinine, fasting, glucose, potassium, sodium and much more. This study evaluates artificial neural network based algorithms for detection of diabetes. It presents a study on the prediction of diabetes disease through the scaled conjugate gradient back propagation of artificial neural networks using k-fold cross validation. The PIMA Indian Diabetes (PID) dataset has been used from Kaggle. The network is trained using data from 768 diabetes patients aged 21 to 81 years. The accuracy of the results depends on the number of neurons in the hidden layer. The proposed system uses 8 input attributes and provides a minimum percentage of accuracy of 77% and a maximum accuracy of 100% by estimating the presence and absence of diabetes during testing.**

Keywords*— *Artificial Neural Network, Back propagation, Diabetes mellitus, PIMA Indian Diabetes (PID) dataset, Scaled Conjugate Gradient Algorithm.

I. INTRODUCTION

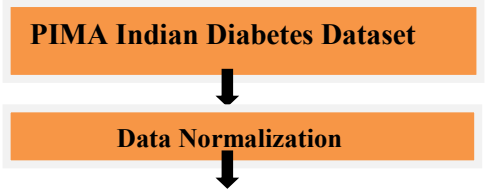
At present, diabetes is the common disease affecting people of all ages. Diabetes mellitus is a combination of metabolic diseases which cannot control the measurement of sugars. Due to diabetes glucose in the blood stream becomes excess. In this disease, failure to regulate blood glucose level can lead to high sugar levels as well as low sugar levels. In both cases late detection of the imbalanced sugar levels can lead to serve health problems such as stroke or kidney failure. Diabetes is caused by many factors such as lifestyle, obesity and genetics. It is classified as Type1, Type2, and Gestational diabetes [1-13]. Type1 is when the pancreas is unable to secret enough insulin. It is more common in children and younger adults under the age of 20 years. Whereas Type2 diabetes results in insulin levels high. It occurs mainly in older and obese people, which usually occurs at an age 40 and more. The Gestational diabetes is found in women within six month pregnancy. Diabetes is said to be a chronic disease, but at the same time can be avoided through proper monitoring of health Symptoms. In this paper we have applied artificial neural network based scaled conjugate gradient back propagation algorithm. By applying this method medical diagnosis expert system produce more accurate results. The rest of the work is arranged as follows: in section 2 we present the survey of related research work; the proposed methodology is presented in section 3; section 4 and 5 determine experimentation and experimental results with observation of the proposed methods respectively and section 6 concludes the work.

II. LITERATURE REVIEW

Works are being carried out with different varieties of methodology and reached various classification accuracies. Using an Artificial Back propagation S C G Neural Network algorithm, Muhammad Masher Bukhari et al. [2021] [1] developed an improved ANN model and obtained 93% accuracy. Abdur Redman et al. [2020] [2] evaluated a 'DELM' based prediction model for improving type2 diabetes prediction accurately and achieved 92.8% accuracy. Tuan Minh LE et al. [2020][3] proposed a machine learning model using two Feature Selection Methods such as Grey Wolf Optimization (GWO) and Adaptive Particle Swarm Optimization (APSO) technique and obtained 96% accuracy for Grey Wolf Optimization – MLP technique and 97% accuracy for Adaptive Particle Swarm Optimization -MLP technique respectively. Deepti Sisodia et al. [2018] [5] implemented three machine learning classification algorithms for early stage detection of diabetes and achieved highest of 76.30% accuracy. Quan Zou et al.[2018][6] predicted diabetes using some machine learning techniques and achieved a maximum of 80.84% accuracy when using Random Forest. Thiyagaranjan C *et al*. [2017] [7] applied a Transudative Extreme Learning Machine (TELM) technique for accurate diagnosis of diabetes mellitus and obtained 96% accuracy. Shantan Sawa et al. [2017][8] established a model for predicting diabetes using rough set clusters and achieved 1.36%, 2.09% & 6.34% of errors for diabetic patients with center values for the Plasma Glucose Concentration (V2), Body Mass Index (V6), & Age (V8) respectively. Dilip Kumar Choubey et al. [2016] [11] implemented a hybrid intelligent system GA-MLPNN for diagnosis of diabetes disease and obtained 79.1304% accuracy. Neha Shukla et al. [2016] [12] proposed a model for diabetes prediction using Random Forest Tree classification techniques and achieved 92.96% accuracy. Muhammad Akmal Sapon et al. [2011][15] has established Scaled Conjugate Gradient Algorithm that produces the best performance in predicting diabetes and obtained with a maximum of 0.88026 Correlation Co-Efficient (R) compared to other Conjugate Gradient methods.

III. METHODOLOGY

Fig. 1 shows the proposed methodology.



978-1-6654-1758-7/21/$31.00 ©2021 IEEE

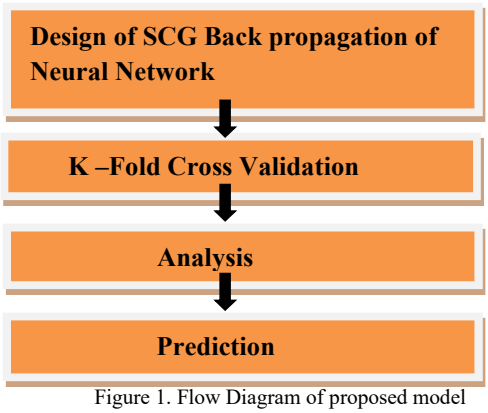


Figure 1. Flow Diagram of proposed model

Proposed Algorithm

Step1: Start
Step2.1: Import PIMA Indian Diabetes Dataset
Step2.2: Import data into the MATLAB in table format which consists of 9 columns and 768 rows
Step2.3: Convert table into the Array
Step3: Apply Scaled Conjugate Gradient Back Propagation Algorithm
Step4: Training dataset
Step5: Calculation of error and accuracy
Step6: Testing dataset
Step7: Calculation of error, accuracy, time elapses and plots Confusion Matrix
Step8: Stop

IV. EXPERIMENTATION

i) Data Set

The PIMA Indian diabetes dataset is developed from female patients for the age group 21 to 81 years. This contains 768 instances, which include 268 diabetics and 500 non-diabetics patients. This system uses 8 input attributes and 1 target attribute. Input attributes are: Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI (Body Mass Index), Diabetes Pedigree Function, Age and 1 output field named 'Existence of Diabetes'. It has two levels 0 and 1. The value of the integer 0 stands for "Non Diabetic Patient" & 1 stands for "Diabetic Patient". The classification is done based on the confusion matrix. Parameters used are: True Positive (TP), True Negative (TN), False positive (FP) and False Negative (FN). PIMA Indian diabetes dataset are explained in following Table I.

TABLE I. PIMA Indian Diabetes Dataset

S. No	Attribute	Statement	Limit
1	Pregnancies	Total count of pregnancies.	[0,17]
2	Glucose	Detects the levels of glucose focus by examining glucose concentration.	[0,199]
3	Blood pressure	Hypertension in mmHg.	[0,122]
4	Skin thickness	Rash and thickening of the skin (triceps).	[0,99]
5	Insulin	Two hours serum insulin in mu U/ml.	[0,846]
6	Body mass index	Weight in kg/ (height in m)^2)	[0,67.1]
7	Diabetes pedigree function	diabetes hereditary effectiveness	[0.078, 2.42]
8	Age	Patient's age	[21,81]
9	Target attribute	Existence of diabetes, yes or no.	[0,1]

ii) Data Normalization

The clinical datasets used in this work are normalized using the following mathematical formula:

$$\text{Normalized (X)} = \frac{\text{Original value in the given set}}{\text{Maximum value in the given range}}$$

The normalized value 'X' lies in the interval [0, 1] Numerical variables such as 'age' is normalized on to the interval [0, 1]. For example, the age of patients range from 21 to 81 years, and thus normalized value of 57 years old patient age is 0.703703704.

iii) K- Fold Cross Validation

Here data sets are randomly partitioned into 'k' number of mutually exclusive subsets or folds. "D1, D2, D3 ... DK" are almost the same size. Training and testing are performed in 'K' times.
The first iteration D_1, is taken as the test set and the remaining D_2, D_3, D_4…DK are served as the training sets.
The second iteration D2, is served as the test set and the remaining D1, D3, D4...DK are served as the training sets.
Similarly the i^{th} iteration Di is served as the test set and the remaining D1, D2 D(i-1), D(i+1), DK are served as the training sets. Each sample is used same number of times for training and once for testing.

$$\text{Accuracy} = \frac{\text{Total number of correct classification from K iteration}}{\text{Total number of tuples in the initial data.}}$$

The classification accuracy is k-fold cross validation of the samples. The results obtained from each fold is averaged and used for comparative analysis.

iv) Scaled Conjugate Gradient Algorithm

It is a feed forward neural network based on supervised learning algorithm. This algorithm was introduced by Martin F. Moller in 1991 and it doesn't contain any of the user dependent parameters. The algorithm keep away from time-consuming line search per learning iteration, leading to better performance than standard back propagation algorithm, conjugate gradient algorithm with line search and 'Broyden Fletcher Goldfarb Shanno' (BFGS) algorithm. It is based on a time-consuming methodology for line search. In between other conjugate algorithms, it requires a larger number of iterations but less computational complexity.

v) Classification Accuracy

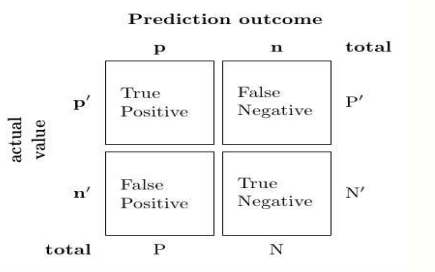


Figure 2. Confusion Matrix

The accuracy of the classification is to predict the performance results. It is measured using the confusion matrix tools. It analyzes how much the classifier can recognize tuples of different classes. The processing outcome is shown in Figure 2.

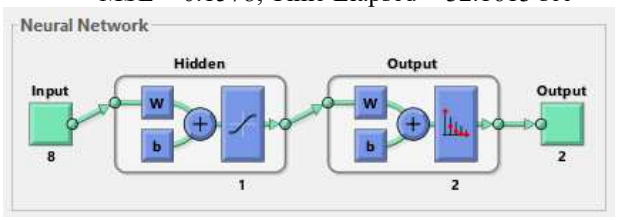
There are four additional terms "True Positive", "True Negative", "False Positive" and "False Negative",

$$\text{Accuracy} = \frac{True\ Positive + True\ Negetive}{TruePositive + True\ Negative + False\ Positive + False\ Negative}$$

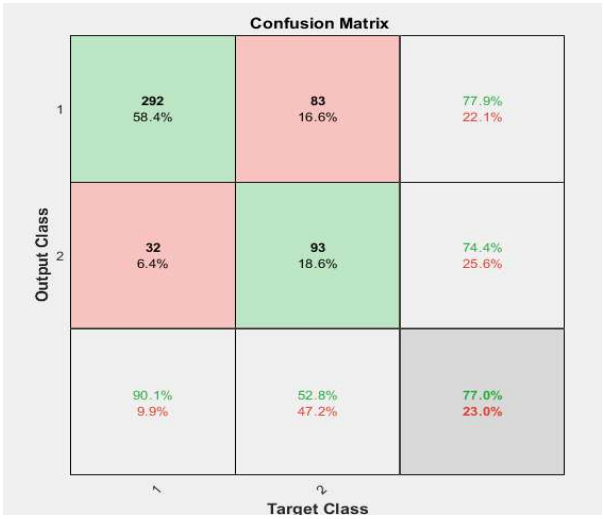
V. EXPERIMENTAL RESULTS & OBSERVATIONS

All experimental results are obtained by using MATLAB coding & implementation.

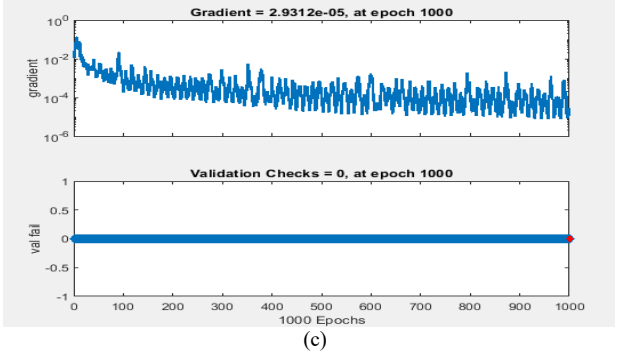
1) Input percentage of training data: 65
Enter learning Rate: 0.6
Number of Input Neurons: 8
Number of Hidden Neurons: 1
Number of Output Neurons: 2
Performance = 0.2412
Percent Errors = 0.2300
Accuracy = 0.7700
Percentage of Accuracy = 77
MSE = 0.1578; Time Elapsed = 32.1613 sec



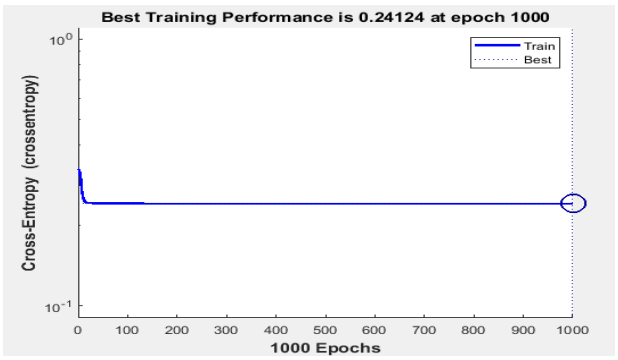
(a)



(b)



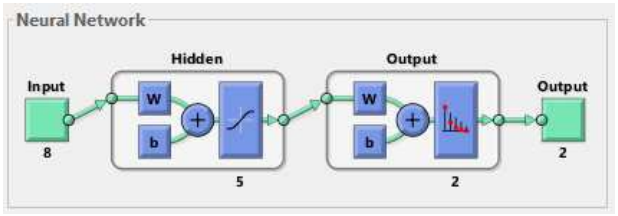
(c)



(d)

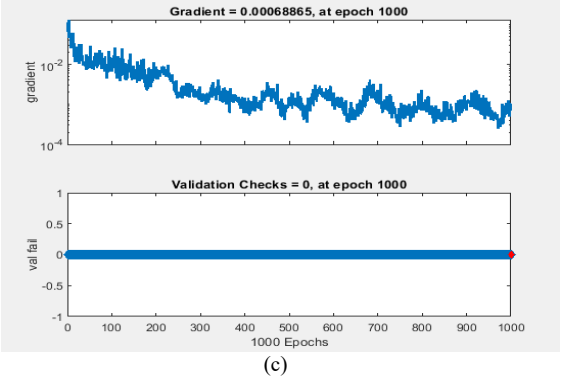
Figure 3. (a) Neural Network model, (b) Confusion matrix, (c) Gradient & Validation Checks and (d) Best training performance.

2) Input percentage of training data: 65
Enter learning Rate: 0.6
Number of Input Neurons: 8
Number of Hidden Neurons: 5
Number of Output Neurons: 2
Performance = 0.1794; Percent Errors = 0.1720
Accuracy = 0.8280
Percentage of Accuracy = 82.8000
MSE = 0.1117; Time Elapsed = 19.8221 sec

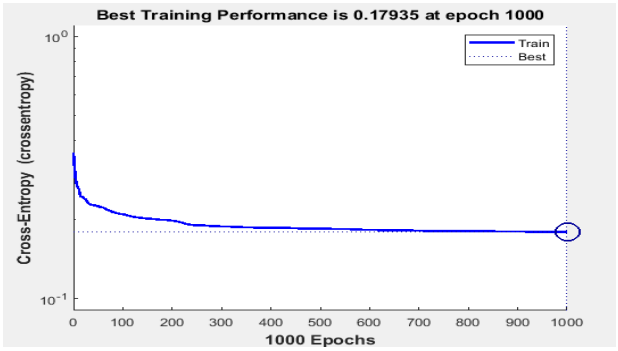


(a)

(b)



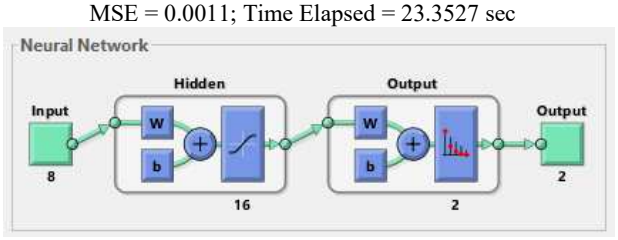
(c)



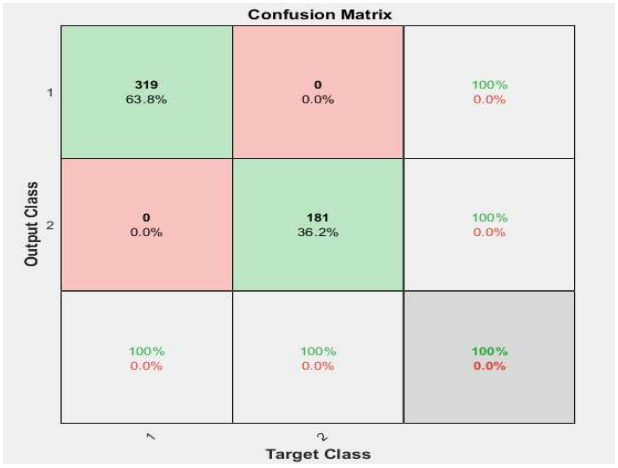
(d)

Figure 4. (a) Neural Network model, (b) Confusion matrix, (c) Gradient & Validation Checks and (d) Best training performance.

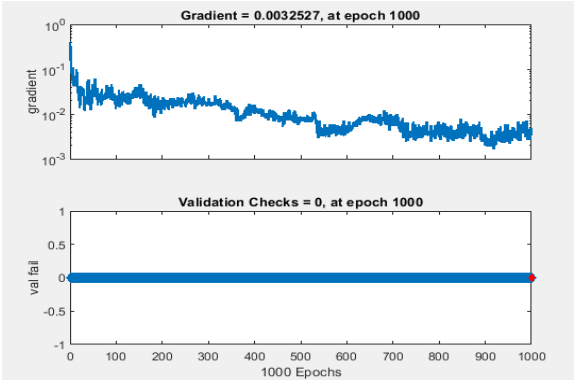
3) Input percentage of training data: 65
Enter learning Rate: 0.6
Number of Input Neurons: 8
Number of Hidden Neurons: 16
Number of Output Neurons: 2
Performance = 0.0043
Percent Errors =0
Accuracy = 1; Percentage of Accuracy = 100
MSE = 0.0011; Time Elapsed = 23.3527 sec



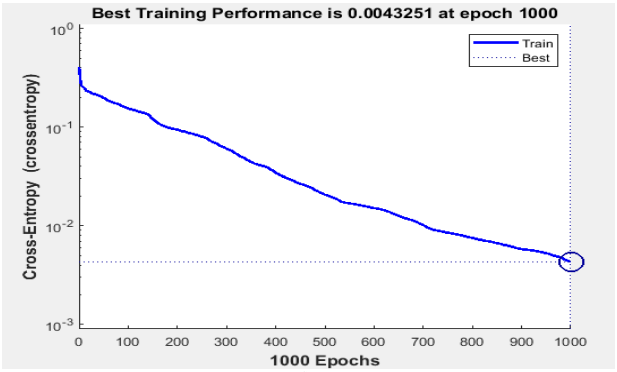
(a)



(b)



(c)



(d)

Figure 5. (a) Neural Network model, (b) Confusion matrix, (c) Gradient & Validation Checks and (d) Best training performance.

From above experimental result it is shown that when input percentage of training data is 65, learning rate is 0.6, number of input neurons is 8, number of output neurons is 2, then, for different numbers of hidden neurons, percentage of accuracy, time elapsed and M.S.E (Mean Squared Error) are changed. It is explained in TABLE II.

TABLE II. Numbers of hidden neurons vs. percentage of accuracy, time elapse in second and mean squared error

Number of Hidden Neurons	Percentage of Accuracy (%)	Time Elapse in Second	Mean Squared Error
1	77	32.1613	0.1578
5	82.8	19.8221	0.1117
10	93.2	29.4503	0.0546
14	96.6	17.6462	0.0250
15	98.6	17.2020	0.0106
16	100	23.3527	0.0011

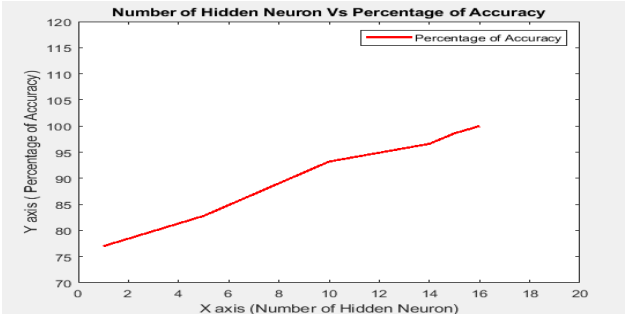


Figure 6. Numbers of Hidden Neurons Vs Percentage of Accuracy

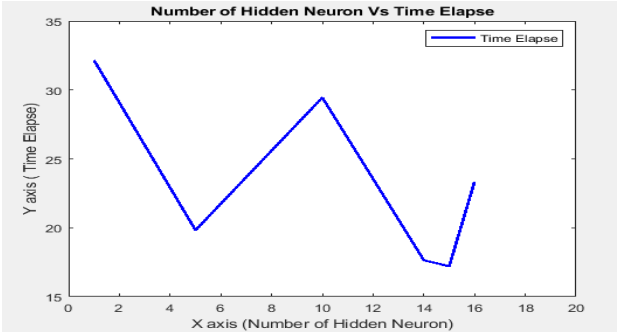


Figure 7. Number of hidden neurons vs. Time Elapse

Numbers of hidden neurons vs. percentage of accuracy are shown in Figure 6. In figure 6, 'X' axis represents the number of hidden neurons and 'Y' axis represents percentage of accuracy. When the number of hidden neurons is increased from 1 to 16, then percentage of

accuracy increases from 77% to 100%. In Figure 7, 'X' axis represents the number of hidden neurons and 'Y' axis represents time elapse in seconds. When the number of hidden neurons is increased from 1 to 16, then time elapse (in seconds) varies from 32.1613sec to 23.3527 sec.

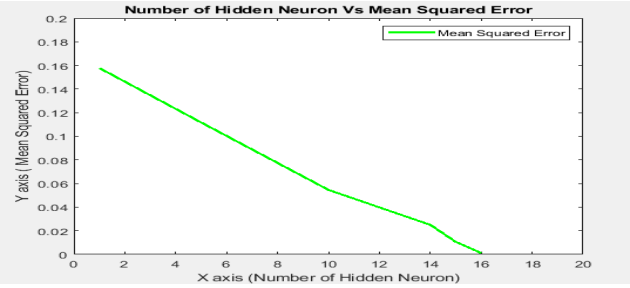


Figure 8. Numbers of Hidden Neurons Vs Mean Squared Error

Numbers of hidden neuron vs. M.S.E (mean squared error) is shown in Figure 8. In Figure 8, 'X' axis stands for the number of hidden neurons and 'Y' axis stands for mean squared error. When the number of hidden neurons is increased from 1 to 16, then the mean squared error decreases from 0.1578 to 0.0011, as the percentage of accuracy increased. During our experimentation, we found that if the numbers of neurons in hidden layer are 16 or more, then the accuracy of the result is 100%.

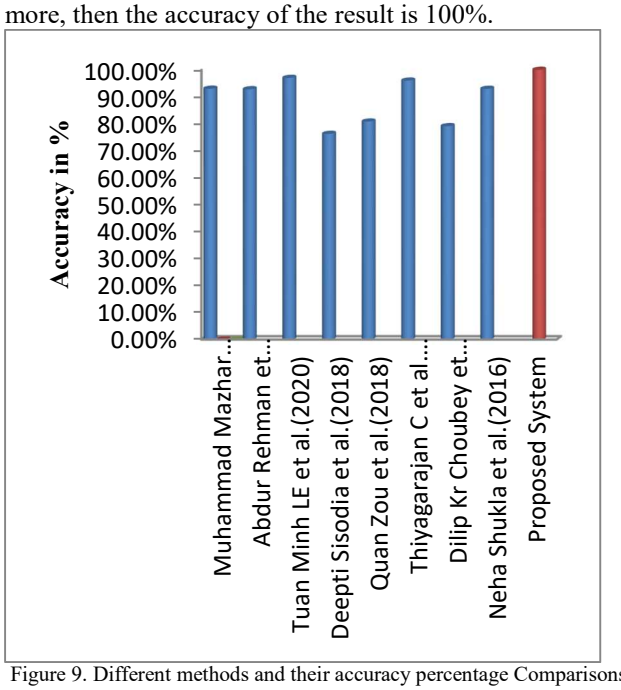


Figure 9. Different methods and their accuracy percentage Comparisons

V. CONCLUSION

The system with accurate diagnosis has been developed using the scaled conjugate gradient back propagation neural network. From Figure 9, (histogram) it can be observed that suggested process has the highest maximum accuracy rate compared to various other methods. Experimental results prove that the percentage of minimum prediction accuracy is 77% and the maximum prediction accuracy is 100% for the taking of different hidden neurons. Thus the experimental results give good and encouraging results to predict diabetes disease with the best possible improved accuracy.

REFERENCES

[1] M.M. Bukhari, B.F. Alkhamees, S.Hussain, A. Gumaei, A. Assiri, and S.S.Ullah, "An improved artificial neural network model for effective diabetes prediction," *Wiley, Hindawi*, 2021.

[2] A.Rehman, A.Athar, M.A.Khan, S.Abbas, A. Fatima, and A.Saeed, "Modelling, simulation, and optimization of diabetes type II prediction using deep extreme learning machine," *Journal of Ambient Intelligence and Smart Environments*, vol.12, pp.125-138, 2020.

[3] T.M.Le, T.M.Vo, T.N.Pham, and S.V.T.Dao, "A novel Wrapper–Based feature selection for early diabetes prediction enhanced with a meta heuristic," *IEEE Access*, vol. 9, pp.7869-7884, 2020.

[4] D.E.Gbenga, D.J.Hemanth, H.Chiroma, S.M.Abdulhamid, and A.J.Taiwo, "Non-nested generalization (NNGE) algorithm for efficient and early detection of diabetes," *Information Technology and Intelligent Transportation Systems, Academia*, pp. 233-241, 2019.

[5] D.Sisodia, and D.S.Sisodia, "Prediction of diabetes using classification algorithms," *International Conference on Computational Intelligence and Data Science, Procedia computer science*, vol.132, pp.1578-1585, 2018.

[6] Q.Zou, K.Qu, Y.Luo, D.Yin, Y.Ju, and H.Tang, "Predicting diabetes mellitus with machine learning techniques," *Frontiers in genetics*, vol. 9, 2018.

[7] C.Thiyagarajan, K.A.Kumar, and A.Bharathi, "Diabetes mellitus diagnosis based on transductive extreme learning machine," *International Journal of Computer Science and Information Security (IJCSIS)*, vol.15, no. 6, 2017.

[8] S.Sawa, H.Balaji, N.Ch.SN.Iyengar, and R.D.Caytiles,"Predicting diabetics accuracy using rough set clusters," *International Journal of Grid and Distributed Computing*, vol.10, no. 9, pp.47-56, 2017.

[9] V.Balpande, and R.Wajgi, "Review on prediction of diabetes using data mining technique," *International Journal of Research and Scientific Innovation (IJRSI)*, vol.4, 2017.

[10] M.Z.Hasan, M.S.Uddin, M.S.Islam, and M.U.Salma, "Learning to classify diabetes disease using data mining techniques," *International Journal of Computer Science and Information Security*, vol.15, no.1, 2017.

[11] D.K.Choubey, and S.Paul, "GA MLP NN: a hybrid intelligent system for diabetes disease diagnosis," *International Journal of Intelligent Systems and Applications*, vol. 8, no.1, pp.49-59, 2016.

[12] N.Shukla, and M.Arora, "Random forest v/s scaled conjugate gradient to predict diabetes mellitus", *International Journal of Computational Intelligence Research*,vol.12, no. 2,pp. 117-123,2016.

[13] N.Chandgude, and S.Pawar, "Diagnosis of diabetes using fuzzy inference system," *International Conference on Computing Communication Control and Automation (ICCUBEA), IEEE*, pp. 1-6, 2016.

[14] P.Radha, and B.Srinvasan, "Hybrid prediction model for the risk of cardiovascular disease in type-2 diabetic patients," *International Journal of Advance Research in Computer Science and Management Studies*, vol. 2, no.10, pp.52-63, 2014.

[15] M.A.Sapon, K.Ismail, S.Zainudin, and C.S.Ping, "Diabetes prediction with supervised learning algorithms of artificial neural network," *International Conference on Software and Computer Applications, Kathmandu, Nepal*, vol.9, 2011.

[16] https://www.kaggle.com/uciml/pima-indians- diabetes- database.