Development of Diabetes Diagnosis System with Artificial Neural Network and Open Source Environment

I.K. Mujawar Assistant Professor Vivekanand College, Kolhapur, Maharashtra, India. mujawarirfan@gmail.com B.T. Jadhav Associate Professor and Principal, Yashwantrao Chavan Institute of Science, Satara, Maharashtra, India. btj21875@gmail.com

R.Y.Patil, Member, IEEE Assistant Professor Vivekanand College, Kolhapur, Maharashtra, India rajashrishendre@gmail.com

Abstract— An artificial neural network started to be used in medical processes and problems when it was claimed that these processes and problems might be complex in nature and that would be best handled by nonlinear approaches. An artificial neural network has been used and applied in great numbers for medical and pathobiological processes with huge success. It is used in disease processes to represent its every subspecialty. Disease diagnosis is one of the important and highly attentive biomedical areas where an artificial neural network has been used to a great extent. In the proposed work, patient's data was collected from hospital and used in the system design. The proposed work presents an architecture, design and development of the online system for a diabetes diagnosis. The proposed system is designed with an open source development environment which will be useful for people in the Diabetes Diagnosis process.

Index Terms— Diabetes Mellitus, Artificial Neural Network, Feed-Forward Network, Open Source Libraries, Online Medical System.

I. INTRODUCTION

Diabetes is death threatening and widely spreading disease, commonly known as blood sugar where glucose level goes high than regular limits due to little or defective or no insulin creation or its use. Type 1, Type 2 and Gestational Diabetes are main categories in which Diabetes is classified. In upcoming 20-30 years, the world's 700 million people of different parts will suffer by this chronic disease. Current studies show that 450 million individuals are affected with Diabetes and its complications. Diabetes can be managed and its complications can be avoided by regular observations, medications, diet and exercise [1, 2, 3, 4]. In recent days, more adequate ways like smart applications, online counseling, and telemedicine are adopted by many people conveniently to deal with medical problems.

The artificial neural network started to be used in medical processes and problems when it was claimed that these processes and problems might be complex in nature and that would be best handled by nonlinear approaches. An artificial neural network has been used and applied in great numbers for medical and pathobiological processes with huge success. ANN is used in disease processes to represent its every subspecialty. Disease prediction is one of the important and highly attentive biomedical areas where ANN has been used to a great extent.

V.B. Waghmare, Member, IEEE

Assistant Professor and Head of Computer

Science Department

Vivekanand College,

Kolhapur, Maharashtra, India. vishal.pri12@gmail.com

II. RELATED WORK

Chongjian Wang et al. presented a classification approach for identifying the risk of Type 2 diabetes risk [5]. Researchers proposed an artificial neural network with three layers where the input layer contains 17 neurons which were nonbiochemical parameters, the hidden layer contains 19 neurons and the output layer contains 1 neuron. 8640 data instances were used to carry out the research. Researchers also used multivariate logistic regression (MLR) for T2DM risk identification. Their work demonstrated that the ANN model gives better sensitivity and specificity (86.93% and 79.14% respectively) than the MLR model (60.80% and 75.48% respectively).

An attempt to classify diabetes into prediabetes and Type 2 diabetes was presented by Dijana Sejdinovie et al. in their research work [6]. Dataset used in this work consists of 310 data instances which were collected from healthcare institutions. 2 input parameters were used in the input layer and the output layer consists of 1 neuron. Proposed ANN model trained with 5, 10, 15 and 20 neurons in the hidden layer. Researchers achieved overall 93.7% classification accuracy for the designed model.

Nesreen Samer El_Jerjawi et al. presented a neural network model for diabetes prediction with one input layer with 8 neurons, three hidden layers and one output layer with 1 neuron [7]. Researchers used a dataset containing 1006 data instances and for ANN model design they used Just Neural Network (JNN) tool. Researchers achieved 87.30% accuracy for Diabetes prediction.

Another attempt to diabetes diagnosis was made by Ebenezer Obaloluwa Olaniyi et al. by presenting an ANN [8]. They used Pima Indian Diabetes Dataset and presented a three-layered back-propagation neural network model with 8 neurons in the input layer, 6 neurons in the hidden layer and 2 neurons in the output layer. Researchers tested and compared their model with other models and achieved 82% diabetes recognition accuracy.

Oluwatosin Mayowa Alade el al. presented an ANN-based Expert System for diabetes diagnosis [9]. Researchers designed a four layered neural network model with a back-propagation learning algorithm and Bayesian Regulation algorithm for avoiding overfitting. The proposed model contains 8 neurons in the input layer, two hidden layers contain 10 neurons each and the output layer contains 1 neuron. Researchers used a Matlab tool for designing the proposed model and a web-based application was developed for the implementation of the work done. Javascript was used for writing model training code.

Rajeeb Dey et al. presented an ANN model with 6-10-1 and 6-14-14-1 architecture for diabetes diagnosis [10]. They collected 530 patient's data on diabetic and non-diabetics. A Matlab tool was used to carry out the experiment and researchers achieved 92.5% accuracy.

An artificial neural network model was proposed by Sareh Mortajez et al. for Type 2 diabetes diagnosis [11]. They proposed the MLNN model and used the Genetic Algorithm (GA) for optimizing the weights of neurons. They tested the model with different count of neurons (2, 3, 4 and 5) in the hidden layer to minimize the mean square error. Researchers achieved 84.5% classification accuracy.

Siti Farhanah Bt Jaafar et al. presented a neural network with a backpropagation algorithm [12]. They used the PIDD dataset and proposed two models with 8-15-14-14-1 and 4-25-23-1 architectures were used to carry out experiments. The first architecture had 8 neurons in the input and 15, 14, 14 neurons in the first, second and third hidden layers respectively and 1 neuron in the output layer. It had given better accuracy than the second architecture (4-25-23-1).

Suyash Srivastava et al. proposed an ANN model and used PIDD dataset for the model design [13]. Researchers carried out their proposed work by using Python programming language and achieved 92% diabetes prediction accuracy. T. Jayalakshmi et al. also proposed an ANN model and used PIDD dataset for model design [14]. They proposed a threelayered model where the input layer has 8 neurons, the hidden layer has 8 neurons and the output layer contains 1 neuron. Researchers used the Levenberg Marquardt learning algorithm and PCA and K-nearest neighbor data preprocessing techniques. A Matlab tool was used to carry the experiment. Researchers successfully showed the impact of data preprocessing over the improvement of the proposed model.

Hasan Temurtas et al. presented two models for diabetes diagnosis [15]. The first model was the feed-forward neural network model with the Levenberg-Marquardt learning algorithm and the second was the probabilistic neural network model. The first model has the four-layered (8-50-50-2) architecture and the second model was the three-layered model with a single input layer with 8 inputs, a single hidden layer for radial basis functions and a single output layer with 2 outputs. Researchers achieved 82.37% and 78.13% accuracies for the first and second models respectively.

Gavin Robertson et al. had given a contribution in blood glucose level monitoring by designing a neural network based on Elman recurrent network which was trained with the Levenberg-Marquardt learning algorithm [16]. Automated Insulin Dosage Advisor (AIDA) simulator was used as a source of data and a Matlab tool was used to design the proposed network. The model presented by the researchers was the most accurate in the long term and short term blood glucose level predictions during night time i.e. nocturnal period

Og^{*}uz Karan et al. presented a neural network model for diabetes diagnosis which was implemented on client-server architecture [17]. Researchers presented two-tier architecture (patient's PDA and Server) based on an artificial neural network. With 11 input nodes in the input layer and 2 nodes in the output layer, researchers designed the proposed model.

III. DEVELOPMENT OF ONLINE SYSTEM FOR DIABETES DIAGNOSIS

In recent years, many studies have shown the importance and efficiency of artificial neural computing applications in the prediction and diagnosis of the diseases in the fields of medicine. It is the way used in this system for the Diagnosis of Diabetes where an artificial neural network approach is used. Multilayer feed-forward neural network with а backpropagation algorithm is used where the patient's background history about diabetes, family background information and lab tests information is used as inputs to the proposed network model and output is class 1 or class 0 i.e. diagnosed with diabetes or not respectively. Designing and developing of the proposed system is given in the next few sections.

A. Data collection

In the proposed system design, to carry out experiments required dataset is created by collecting diabetes patient's data from the 'Swad Diabetic Care Centre'. It has 461 data instances which are of both genders of male and female and age from 9 to 79 years. From 461 data instances 107 instances are of non-diabetics and 354 of diabetics [18].

B. Attribute selection

In the diabetes prediction using an artificial neural network, five attributes are considered. These are high-risk attributes used in work presented for a diabetes diagnosis. These attributes are input to the designed neural network and one output attribute is Class 0 i.e. No Diabetes or Class 1 i.e. Diabetes. The only one output attribute is used to classify the test data in Class 0 means data instance with No Diabetes and Class 1 means data instance with Diabetes. Table I shows these attributes with their codes.

C. Data Preprocessing

Data may be inconsistent and incomplete when used in the experiments carried out with machine learning techniques. The data preprocessing is a technique used for the transformation of such data into meaningful and clean data. In the second way of diabetes diagnosis, as an artificial neural network machine learning technique is used, data considered for the same requires the data preprocessing to achieve better results [19, 20]. The following data preprocessing techniques are used to get data in a structured form.

TABLE I. INPUT AND OUTPUT ATTRIBUTES

Attribute No.	Attribute Name	Attribute Code	
A1	Body Mass Index	BMI	

A2	Age	Age
A3	Fasting Plasma Glucose	FPG
A4	Post Prandial Glucose	PPG
A5	Diabetes Pedigree Function	DPF
01	Class 0 or Class 1	OC

1) Finding Data Outlier

Data instances those are far different from other data instances are Data Outliers. In the proposed work, from 461 data instances, 4 data instances were found as data outliers. Input attribute Fasting Plasma Glucose (FPG) has 1 whereas Postprandial Glucose has 3 data outlier instances. These are shown in the Figure 1.



Fig. 1. Data Outliers

2) Filling Missing Values

To make complete data set for the betterment of final results, finding out missing values and take appropriate action with these values is very important. Missing values of the data set can be dealt with by removing the whole record having missing values or filling it with either mean or average or smallest or largest value of an attribute of a data set. In the proposed work, missing values of a dataset are filled by mean of the respective attribute.

3) Data Normalization

Data normalization or Data rescaling is another important step in data preprocessing. In machine learning techniques, data rescaling plays a vital role to improve results. In the data set used, there are attributes of varying ranges or scales. In this experiment, all the input attribute values have taken in the range of 0 and 1. Some values after rescaling have shown in Table II.

TABLE II. NORMALIZED DATASET RECORDS

BMI	AGE	FPG	PPG	DPF
0.4496	0.1367	0.4070	0.7487	0
0.4643	0.2405	0.4166	0.4568	0
0.6024	0.3417	0.6089	0.7106	0.8571
0.7865	0.3417	0.5769	0.8324	0
0.6464	0.3797	0.6089	0.7106	0.5714
0.7029	0.3797	0.3012	0.3299	0
0.6902	0.3924	0.3525	0.3807	0
0.7400	0.4050	0.4070	0.3426	0.5714
0.6569	0.4177	0.2628	0.5736	0.5714
0.7613	0.4177	0.4423	0.6218	0.5714

D. Training

In an artificial neural network model, training is a vital state where the model learns using a learning algorithm. Back Propagation algorithm is one of the learning algorithms. In the proposed work Back Propagation algorithm is used. In the training process, the model adjusts its weights with respect to the input given through the input layer until the output of the model gets to the desired value by using cost function[21]. At the input layer Rectifier Activation function (relu) is used where at the output layer Sigmoid Activation function is used. In hidden layer Rectifier Activation function (relu) is used.

1) Selection of final model

The final selection of the neural network model is done by using the tuning of the hyperparameters of the proposed model. In this process optimization of many of the parameters is done and based on the better optimization results selection of parameters for the final model design is done. In the proposed model design, batch size, number of epochs, weights, optimization algorithm, learning rate, momentum, activation function and number of perceptrons in the hidden layer are the hyperparameters which are optimized and their optimized values are considered in the model design [19, 20].

a) Tuning Batch Size and Epoch

In the final selection of batch size and the number of epochs for the proposed model, a list of batch size with size 5 is taken as [10, 20, 30, 40, 50] and a list of the number of epochs with size 6 is taken as [50, 100, 200, 300, 400, 500] respectively. After performing the test we found, with batch size 40 and 500 epochs model gives the highest accuracy with 93.79%.

Best: 0.937931 (0.079833) with: {'batch_size': 40, 'epochs': 500}

b) Tuning of Optimization Algorithm

In the proposed model we found 'Adam' optimizer gives better accuracy from SGD, Adam, Adamax, Nadam and Adagrad optimizer algorithms.

Best: 0.900000 using {'optimizer': 'Adam'}

c) Tuning of Learning Rate and Momentum

Weights are changed at each neuron at the end of the batch and it is controlled by the Learning Rate of the model. Momentum is one of the hyperparameters which controls the influence of the previous update to the current update of weights. In this parameter tuning, we found, 0.1 and 0.4 is learning rate and momentum respectively giving the best accuracy with 91.72% among the set of learning rates and momentum has taken.

Best: 0.917241 using {'learn_rate': 0.1, 'momentum': 0.4}

d) Tuning of Activation Function

Activation functions control the nonlinearity of neurons. Different activation functions are used at hidden layers. The sigmoid activation function is used in the output layer. After tuning this parameter with softmax, relu, tanh, sigmoid and linear activation function values we found that 'relu' activation function is giving the highest accuracy among all.

Best: 0.927835 using {'activation': 'relu'}

e) Tuning of Number of Neurons in Hidden Layer

The tuning of the number of neurons in hidden layers is also an important parameter to tune. There are few basic rules in deciding the number of neurons in hidden layers as the count should be 2/3 of input layer neurons or twice the input layer neurons or the size should be in between input layer neurons and output layer neurons. But we have taken a list of total neurons in the hidden layer and optimized the model with different count of neurons in the hidden layer. We found that 8 neurons in the hidden layer which is taken into the model give more accuracy than other counts.

Best: 0.889655 using {'neurons': 8}

E. Architecture of proposed neural network model

It is the process of designing an artificial neural network that starts with data collection, data preprocessing and data selection discussed in the previous few points. Along with these steps, determination and establishment of a neural network of a particular type, selection of count of layers, number of neurons of each layer, and activation function are important elements that must be considered in the same process [22].

In the proposed work, multilayer feed-forward neural network architecture is designed with a back-propagation learning algorithm [23, 24, 25]. In this design, one input layer with one hidden layer and one output layer exists so it is multilayer perceptron architecture shown in Figure 2. It consists of five neurons i.e. processing elements in the input layer and one neuron in the output layer which is for showing the result of the Diabetes Diagnosis class. There are eight neurons in the hidden layer of the proposed architecture.

F. System Development

1) Development Environment

Python programming language and Flask micro framework with Spyder Scientific Development Environment is used. Keras, TensorFlow, matplotlib, PyMySQL, pandas, jupyter, scikit-learn, NumPy and spyder packages are used in the development of the proposed module. The graphical user interface of the module is designed with HTML5, Bootstrap, CSS and JavaScript.

*Code Samples for Data Preprocessing*For Finding Data Outlier

datafrg= pnd.read csv("dataset.csv")

datafrg.plot(kind = 'Box', figsize = (20,10))

pl.show()

datafrg = datafrg [datafrg ['FPG']<320]

datafrg = datafrg [datafrg ['PPG']<450]



Fig. 2. Proposed ANN Model

For Filling Missing Values

datafrg.datafrg[datafrg['FPG']==0,'FPG']=datafrg ['FPG'].mean()

datafrg.loc[datafrg ['FPG']==0,'PPG']= datafrg

['PPG'].mean()

datafrg.to_csv('cleaneddata.csv')

For Data Normalization

datafrg = datafrg / datafrg.max()

datafrg.to_csv('normalizeddata.csv')

- 3) Code Samples for Model Design
- For Loading Dataset
- dt = numpy.loadtxt("dataset.csv", delimiter=",")
- For Model Creation

mymodel = keras.models.Sequential()

• For Model Compilation

mymodel.compile(loss='binary_crossentropy',
optimizer='adam', metrics=['accuracy'])

For Model Saving

mymodel.save('mymodeltrained.h5')

IV. RESULT AND DISCUSSION

In the proposed artificial neural network model multilayer feed-forward network is used with five input parameters passed to the input layer, one hidden layer with eight neurons and an output layer with one neuron. The model summary is given in Table III.

Out of 457 data instances of the Data set, 85% of data instances i.e. 388 data instances are used for the training of the proposed model and the remaining 15% data instances i.e. 69 data instances used for testing. While conducting the test of the proposed neural network model, we got 94.20% accuracy as out of 69 data instances, the model drew the correct result for 65 data instances.

TABLE III. MODEL SUMMARY

Number of Total Data Instances of Patient Dataset	457
85% of Total Data Instances of Patient Dataset taken for Training the Model	388
15% of Total Data Instances of Patient Dataset taken for Testing the Model	69
Number of Neurons in Hidden Layer	8
Number of Neuron in Output Layer	1
Number of Epochs	500
Activation Function at Input Layer	Relu
Activation Function at Output Layer	Sigmoid
Batch size	40
Training Accuracy	95.52%
Testing Accuracy	94.20%

Table IV shows 40 data instances with the actual class of diabetes diagnosis i.e. whether the patient is diagnosed with diabetes or not and output in the form of Class 0 or Class 1 generated by the model designed. Highlighted row of Table IV shows incorrect results drawn by the proposed artificial neural network model.

TABLE IV. DATA INSTANCES WITH ORIGINAL DIAGNOSIS CLASS AND SYSTEM'S OUTPUT

FPG	PPG	BMI	AGE	DPF	Actual Class (0/1)	System Output (Class 0
105	205	10.00	10		(0,1)	or 1)
127	295	18.39	10	0	1	1
138	302	26.57	16	0.5	1	1
120	170	24.5	18	0	0	0
130	180	18.99	19	0	1	1
180	320	32.17	27	0	1	1
190	280	26.44	30	0.5	1	1
126	165	25.2	30	0	0	0
110	150	28.23	31	0	0	0
82	226	26.87	33	0.5	1	1
138	245	31.14	33	0.5	1	1
92	125	27.99	37	0.5	0	0
169	308	30.15	38	0	1	1
133	222	23.75	40	0	1	1
102	300	24.36	41	0.5	1	1
95	131	25.1	43	0	0	0
85	142	20.06	43	0.5	0	0
120	169	29.52	43	0	0	1

216	282	27.06	44	0	1	1
93	120	30.4	45	0.125	0	0
87	208	31.5	45	0	1	1
160	270	32.29	46	0.5	1	1
104	237	27.95	48	0.5	1	1
235	350	28.4	49	0.5	1	1
105	137	22.52	50	0.5	1	1
140	194	25.83	50	0	1	1
88	125	27.99	50	0.5	0	0
92	1128	25.44	51	0.5	0	0
82	121	25.61	52	0.5	0	0
127	249	25.83	53	0.5	1	1
260	320	23.84	54	0	1	1
143	338	30.02	54	0	1	1
97	127	25.2	55	0.5	0	0
85	123	21	55	0	0	0
95	448	25	55	0.5	0	0
110	150	22.72	56	0	1	1
175	185	32.42	56	0.5	1	1
102	124	24	56	0.75	0	0
100	124	29.45	56	0.5	0	0
151	320	18.94	57	0.5	1	1
163	206	25.4	57	0.5	1	1

V. CONCLUSION

In the work proposed for Diabetes diagnosis, an artificial neural network is used and it is implemented with open source environment. The experiment which we carried out with this approach, we achieved 95.52 % training accuracy whereas 94.20% testing accuracy. The graphical user interface is designed and developed so that users of this system can use it for Diabetes Diagnosis. This work will be extended for betterment in diagnosis accuracy and diabetes classification module will be added in the proposed system.

REFERENCES

- Cho, N. H., et al. "IDF Diabetes Atlas: global estimates of diabetes prevalence for 2017 and projections for 2045." Diabetes research and clinical practice 138 (2018): 271-281.
- [2] Goldenberg, Ronald, and Zubin Punthakee. "Definition, classification and diagnosis of diabetes, prediabetes and metabolic syndrome." Canadian journal of diabetes 37 (2013): S8-S11.
- [3] Colledge, Nicki R., et al. Davidson s principles and practice of medicine. Churchill Livingstone/Elsevier, 2010.
- [4] American Diabetes Association. "American diabetes association standards of medical care in diabetes-2017." Diabetes Care 40.1 (2017): S1-S135.
- [5] Wang, Chongjian, et al. "Evaluating the risk of type 2 diabetes mellitus using artificial neural network: an effective classification approach." Diabetes research and clinical practice 100.1 (2013): 111-118.
- [6] Sejdinović, Dijana, et al. "Classification of prediabetes and type 2 Diabetes using Artificial Neural Network." CMBEBIH 2017. Springer, Singapore, 2017. 685-689.
- [7] El_Jerjawi, Nesreen Samer, and Samy S. Abu-Naser. "Diabetes prediction using artificial neural network." (2018).
- [8] Olaniyi, Ebenezer Obaloluwa, and Khashman Adnan. "Onset diabetes diagnosis using artificial neural network." Int. J. Sci. Eng. Res 5.10 (2014): 754-759.
- [9] Alade, Oluwatosin Mayowa, et al. "A neural network based expert system for the diagnosis of diabetes mellitus." International Conference on Information Technology Science. Springer, Cham, 2017.
- [10] Dey, Rajeeb, et al. "Application of artificial neural network (ANN) technique for diagnosing diabetes mellitus." 2008 IEEE Region 10 and the Third international Conference on Industrial and Information Systems. IEEE, 2008.
- [11] Sareh, M., and J. Amir. "An Artificial Neural Network Model to Diagnosis of Type II Diabetes." Journal of Research in Medical and Dental Science 7.1 (2019): 66-70.

- [12] Jaafar, Siti Farhanah Bt, and Dannawaty Mohd Ali. "Diabetes mellitus forecast using artificial neural network (ANN)." 2005 Asian Conference on Sensors and the International Conference on New Techniques in Pharmaceutical and Biomedical Research. IEEE, 2005.
- [13] Srivastava, Suyash, et al. "Prediction of Diabetes Using Artificial Neural Network Approach." Engineering Vibration, Communication and Information Processing. Springer, Singapore, 2019. 679-687.
- [14] Jayalakshmi, T., and A. Santhakumaran. "A novel classification method for diagnosis of diabetes mellitus using artificial neural networks." 2010 International Conference on Data Storage and Data Engineering. IEEE, 2010.
- [15] Temurtas, Hasan, Nejat Yumusak, and Feyzullah Temurtas. "A comparative study on diabetes disease diagnosis using neural networks." Expert Systems with applications 36.4 (2009): 8610-8615.
- [16] Robertson, Gavin, et al. "Blood glucose prediction using artificial neural networks trained with the AIDA diabetes simulator: a proof-of-concept pilot study." Journal of Electrical and Computer Engineering 2011 (2011).
- [17] Karan, Oğuz, et al. "Diagnosing diabetes using neural networks on small mobile devices." Expert Systems with Applications 39.1 (2012): 54-60.
- [18] Mujawar, I. K., et al. "Online approach for Diabetes Diagnosis and Classification with Expert System Modules using Fuzzy Logic." 2019 IEEE Pune Section International Conference (PuneCon). IEEE, 2019.

- [19] Kotsiantis, S. B., Dimitris Kanellopoulos, and P. E. Pintelas. "Data preprocessing for supervised leaning." International Journal of Computer Science 1.2 (2006): 111-117.
- [20] Mohd Nawi, Nazri, Walid Hasen Atomia, and Mohammad Zubair Rehman. "The effect of data pre-processing on optimized training of artificial neural networks." (2013).
- [21] Joshi, Sneha, and Megha Borse. "Detection and prediction of diabetes mellitus using Back-propagation neural network." 2016 International Conference on Micro-Electronics and Telecommunication Engineering (ICMETE). IEEE, 2016.
- [22] Steenbergen, N. Predicting glucose concentration in type 1 diabetes patients using artificial neural networks. BS thesis. 2014.
- [23] Olaniyi, Ebenezer Obaloluwa, and Khashman Adnan. "Onset diabetes diagnosis using artificial neural network." Int. J. Sci. Eng. Res 5.10 (2014): 754-759.
- [24] Adeyemo, A. B., and A. E. Akinwonmi. "On the diagnosis of diabetes mellitus using artificial neural network model artificial neural network models." Afr J Comput Ict 4 (2011): 1-8.
- [25] Soltani, Zahed, and Ahmad Jafarian. "A new artificial neural networks approach for diagnosing diabetes disease type II." International Journal of Advanced Computer Science and Applications 7.6 (2016): 89-94.