A New Artificial Neural Networks Approach for Diagnosing Diabetes Disease Type II

Zahed Soltani
Department of Computer engineering
Urmia Branch, Islamic Azad University
Urmia, Iran

Ahmad Jafarian*
Department of Mathematics
Urmia Branch, Islamic Azad University
Urmia, Iran

***Abstract*—Diabetes is one of the major health problems as it causes physical disability and even death in people. Therefore, to diagnose this dangerous disease better, methods with minimum error rate must be used. Different models of artificial neural networks have the capability to diagnose this disease with minimum error. Hence, in this paper we have used probabilistic artificial neural networks for an approach to diagnose diabetes disease type II. We took advantage of Pima Indians Diabetes dataset with 768 samples in our experiments. According to this dataset, PNN is implemented in MATLAB. Furthermore, maximizing accuracy of diagnosing the Diabetes disease type II in training and testing the Pima Indians Diabetes dataset is the performance measure in this paper. Finally, we concluded that training accuracy and testing accuracy of the proposed method is 89.56% and 81.49%, respectively.**

***Keywords*—diabetes type 2; probabilistic artificial neural networks; data mining; mean squares error; Naive Bayes**

I. INTRODUCTION

Diabetes is one of the most common diseases in the world. This disease is divided into two types: type 1 and type 2. In this paper, we focused on diagnosing diabetes type 2 using PPN. The causes of diabetes include genetics, unsuitable diet, lack of physical activity, obesity, etc. Postponing the diagnosis and treatment of type 2 diabetic patients leads to some major issues such as heart attacks, strokes, blindness, and kidney failure, and in some cases it causes Mutilation [1,2,3,4]. Indeed one of the most important problems in the medical world is timely and exact diagnosis of diseases. Generally, diagnosis is a complex task which requires high skill and experience. Timely diagnosis and specialized medical care of patients can reduce issues of patients, as well as treatment costs in other therapeutic courses [5,6]. A lot of diverse solutions have been proposed for different diseases up to now. Artificial neural networks are the most common solution. They are a branch of artificial intelligence and accepted as a novel technology in computer science. Artificial neural network is a technique which tries to simulate behavior of the neurons in humans' brain. This technique has had a wide usage in recent years. Diagnosis, estimation, and prediction are main applications of artificial neural networks. Artificial neural networks with their own data try to determine if a person is patient or not. With diverse models of ANNs it is possible to perform the diagnosis task on different diseases, so we can address the issue of identifying the disease at the first phase using different models of ANNs with reduced human related errors. This way, the patient can prevent irreversible complications and save his health. In addition, in recent years ANNs applied in all fields of medical sciences, and have been accepted by physician for both identifying the diseases and patient treatments [6, 7]. When using ANNs for identifying diseases, the aim is to achieve a high accuracy rate, and quality of diagnosis depends on training and testing a data set [7]. Before starting the task of diagnosis, models of ANNs must be trained according to patient's data sets. After training and testing, the patient's data sets using models of ANNs, the way for achieving high accuracy and minimum error rate is provided [7, 8]. In this paper we aims at achieving maximum accuracy rate in training and testing phases of diagnosing the diabetes disease type 2 using PPN models. Some advantages of PNN include its novelty for identifying diabetes type 2 instead of using traditional approaches with human error, as it reduces the human costs and has made a major contribution to medical science. To this end, a data set named Pima Indians Diabetes with 768 data samples is used, such that each sample includes certain features. By using these 768 data sample, we train the PNN model. Furthermore, we implemented the model of PNN for diagnosing diabetes diseases in MATLAB.

The remainder of this paper is organized as follows. We first summarize some related works in Section 2. We study and discuss the Pima Indians Diabetes data set in section as well as the proposed PNN model for diagnosing diabetes type 2 based on both training and testing Pima Indians Diabetes. Then we investigate and compare the related works on identifying diabetes type 2 using PNN model in training and testing phases in section 4. Finally, in section 5 we conclude this paper and suggest some future works in the area of diabetes type 2 diagnoses.

II. RELATED WORKS

Significant research in the context of diagnosis of diabetes using artificial neural networks and data mining techniques have been done till now. In this section we study some of these researches, and then compare them with the accuracy of training and testing the process of diagnosing diabetes type 2.

Sa'di et.al. [4] used data mining techniques such as Naïve Bayes, J48 and Radial Basis Function Artificial Neural Networks for diagnosing diabetes type 2. They took advantage of a data set with 768 data samples, 230 of them selected for test phase. Naive Bayes algorithm with 76.95% accuracy outperformed J48 and RBF with 76.52% and 74.34% accuracies, respectively.

Authors in [9] use back-propagation multi-layer artificial neural networks for identifying diabetes type2. Back-propagation is a supervised learning algorithm and is based on error correction. Back-propagation compares the computed output value with the real value and tries to modify the weights according to the calculated error, such that after each round, the size of obtained error be less than the value in the previous round. In order to train Back-propagation, we used Sigmoid function. Also, we used Pima Indians Diabetes in our evaluations. This data set contains 768 data sample, 568 of which were used as training set and 268 as testing set. Furthermore, back-propagation multi-layer artificial neural networks consists of 8 neurons in input layer, 6 neurons in hidden layer, and 2 neurons in output layer. Note that input layer neurons are those 8 features which are used in dataset. The achieved diagnosis accuracy after 2000 rounds of dataset training becomes 82%. Hence, Back-propagation algorithm has the highest accuracy compared to BSS, EM, KNN, C4.5 approaches.

Al-Rofiyee, *et al.* [10] use multilayer perceptron (MLP) artificial neural networks for identifying diabetes type 2. There are a lot of problems which are not linearly separable. Diabetes diagnosis is one of these problems, because diagnosing diabetes disease using single layer perceptron is wrong and there is no right answer for it. In order for artificial neural network to learn non-linear, it should designed as multi-layer. Each layer can contain different number of neurons. Therefore, they used multi-layer perceptron for diagnosing diabetes type 2. Multi-layer perceptron networks consist of an input layer, several hidden layers and an output layer. In this paper, MLP model includes one input layer with feature of Pima Indians Diabetes, hidden layer with certain neurons and an output layer which has the responsibility of diagnosis. About 20% of data are used as training set, 60% as testing set and finally 20% are used as application set. Time and number of neurons in hidden layer of MLP model are two important parameters. Finally, highest diagnosis accuracy in training phase using MLP model with maximum time and minimum number of neurons in hidden layer in comparing with same times and neuron numbers was 97.61%.

In [11], authors use a dataset with 250 data samples for diagnosing diabetes disease. Each of these 250 data samples consist of 27 features. These features include blood pressure, creatine, pH urine, and fasting blood sugar. Also, the average age of patients in their dataset is between 25 and 78 years. Multi-layer feed-forward artificial neural networks with back-propagation are used for diagnosis. Three training functions namely BFGS Quasi-Newton, Bayesian Regulation and Levenberg-Marquardt are applied in back-propagation algorithm. Finally, back-propagation with Bayesian Regulation function achieved 88.8% of diagnosing accuracy which performed better than BFGS Quasi-Newton and Levenberg-Marquardt functions. Furthermore, in [2], data mining techniques with Pima Indians Diabetes dataset is used for identifying diabetes. The applied data mining techniques include SVM, KNN, C4.5 and artificial neural networks with input, hidden and output layers. Finally, artificial neural networks have a higher diagnosing accuracy compared to other data mining techniques.

In [12], authors use general regression neural networks and Pima Indians Diabetes for identifying diabetes type 2. GRNN model in this paper is assumed to be a four-layer model; one input layer with 8 features from Pima Indians Diabetes, two layers which have 32 and 16 neurons, respectively. Finally, output layer has one neuron. This neuron determines if a person is patient or not. It is used for classification of Pima Indians Diabetes dataset into healthy and patient classes. The above mentioned dataset with 576 data sample as training set and 192 data set as testing set is used for training and testing processes. The accuracy rate achieved for training and testing phases are 82.99% and 80.21%, respectively. Training phase for diagnosing diabetes type 2 obtained a higher value of accuracy compared to other works studied in this paper.

III. PROPOSED MODEL

In order to identify diabetes and other diseases such as heart diseases [13, 14], Parkinson's disease [15, 16], and lung cancer, having a data set is very important and necessary, since ANNs are trained by these data sets and they can perform the diagnosis task. Therefore, in this paper, we used Pima Indians Diabetes [19] with 768 data sample for diagnosing diabetes type 2. This data set consists of 9 features for each data sample. Table (1) shows these 9 features.

According to Table 1, there are 9 features for each data sample. The first 8 features are inputs, and the last feature is the only output. In order to classify the 768 data samples, 9th feature is used as it is classified into two classes: class zero (healthy) and class 1 (patient).

TABLE I. FEATURES OF PIMA INDIANS DIABETES FOR DIAGNOSING DIABETES DISEASE TYPE 2 [19]

No. Attributes	**Attributes**	**Descriptions and Attribute values**
1	Number of Times Pregnant (NTP)	Numerical values
2	Plasma Glucose Concentration (PGC)	Numerical values
3	Diastolic Blood Pressure (DBP)	Numerical values in (mm Hg)
4	Triceps Skin Fold Thickness (TSFT)	Numerical values in mm
5	2-Hour Serum Insulin (2-HSI)	Numerical values in (mu U/ml)
6	Body Mass Index (BMI)	Numerical values in (weight in kg/(height in m)^2)
7	Diabetes Pedigree Function (DPF)	Numerical value
8	Age	Numerical values
9	Diagnosis of type 2 diabetes disease	Yes=1 No=0

TABLE II. STATISTICAL ANALYSIS FOR MEAN AND STANDARD DEVIATION IN PIMA INDIANS DIABETES DATA SET [19]

No. of Feature	Feature Name	Mean	Standard Deviation
1	Number of times pregnant	3.8	3.4
2	Plasma glucose concentration	120.9	32.0
3	Diastolic blood pressure	69.1	19.4
4	Triceps skin fold thickness	20.5	16.0
5	2-Hour serum i insulin	79.8	115.2
6	Body mass index	32.0	7.9
7	Diabetes pedigree function	0.5	0.3
8	Age	33.2	11.8

The average age of this data set is between 21 and 81 years.

In addition, according to Pima Indians Diabetes data set which has 768 data samples, Table 2 shows the Mean and standard deviation of the data set.

Nowadays, artificial neural networks can be used in all industries. Artificial neural network consists of a set of neurons which are characterized by special arrangement. The main parts of an artificial neural network are neurons and connections between them. Neurons are conjunct processing elements which work together to solve a problem [8, 20, 21]. Learning capability is the main advantage of ANNs, since an ANN will adjust in learning process for information classification, and identifying patterns [18, 20]. ANNs with a high ability to diagnose diseases help in medicine sciences. The reasons ANNs are used in medicine include high accuracy of physicians in their decision makings, increase confidence, creating medicine tools, reduce costs, etc. [22]. ANNs consist of different models such as PNN, MLP, RBF, and GRNN [8, 20, 21]. In this paper, we use PNN model for diagnosing diabetes type 2. PNN model has a parallel structure and is special for information classification. In contrast to other ANNs such as MLP, PNN has a higher speed in training the data, and it finds answers faster than MLP [6, 20, 21]. This model consists of 3 layers: input layer, hidden layer, and output (competitive) layer. The hidden layer is also called radial base layer, as PNN model is a mode of RBF model. Hidden layer units uses Gaussian transmission function, and number of neurons in this layer is same as number of rounds in training data set. This layer computes distance between input vector and training inputs, and provides a vector where its elements determine the distance between the input and training inputs. Hidden layer generates a vector of probabilities as output. Finally, this layer selects probability values from probabilities vector and generates value 1 for it and 0 for other probabilities [6, 20, 21]. Gaussian transmission function which is used in hidden layer calculated as follows [20, 21]:

$$D_{t,r}(P) = \frac{1}{(2\pi\sigma^2)} \exp\left(-\frac{\left\| P - P_{t,r} \right\|^2}{2\sigma^2} \right) \quad (0)$$

Where P (P1, P2,…, Pn) is the input vector which consists n variables. Neurons of hidden layer are divided into t groups, each group includes one class. r is the neuron in group t which is calculated using Gaussian function. σ is a constant that increases training and testing accuracy and is used by neurons of hidden layer. Sum of neurons in a group of t is computed using equation (2) [8, 20, 21].

$$G_t(P) = \sum_{r=1}^{Z_t} W_{tr} D_{t,r}(P), \quad T \in \{1,...,T\}, \quad (0)$$

Where, Z_t is the number of neurons of the pattern in a class. W_{tr} is the coefficients of weights, where $G_t(P)$ generate 1 for maximum of the probability values. Classified pattern vector P belongs to the class which relates to sum unit with maximum output value and is calculated as follows [8, 20, 21].

$$Y_r = \begin{cases} 1 & if \quad S_r \; is \max of \{G_1, G_2\} \\ 0 & else \end{cases} \quad (0)$$

Training in ANNs is a process where ANN learns to identify the pattern in its inputs which has the form of a training data set. In fact, ANN adopts (adjust) its weights in response to the inputs at the training phase, such that the real output of ANN converges to the desired output. Once actual output of ANN becomes the desired output, the process of training the networks terminates, such that ANN leads to the least error rate [5, 6, 8, 20, 21]. After training the ANN using training data until achieving minimum error rate, other data which have no effects in training process feeds to the ANN as testing inputs. Then ANN's response is compared with the desired response and accordingly we get a trained network. If ANN responds in a right way to when it is tested, the training process terminates, otherwise ANN training starts again. Lastly, when the program passes the training phase, and generates the right response on every input, then it is obvious that ANN's weights adjusted in a right way. Therefore, from now on these values are used for diagnosis and prediction tasks [5, 6, 8, 20, 21]. Additionally, Mean Square Error (MSE) is the performance index in the process of training PNN model. MSE reduces amount of computations and memory requirements for computational tasks, and it is trying to minimize the MSE value for training data [20,21]. Hence, number of neurons in hidden layer according to MSE is same as number of rounds at the last desired execution of PNN model in training process. For computing MES equation (4) is used [20, 21].

$$MSE = \frac{1}{Q} \sum_{k=1}^{1} \left(t(k) - a(k) \right)^2 \quad (0)$$

Where a(k) is the real output of the network, t(k) is the expected output and Q is number of rounds. In this paper we used Pima Indians Diabetes data set which consist 768 data samples. 90% of data samples, that is, 691 of them used for training, and 10%, that is, 77 out of 768 data samples used for testing. After training and testing the data using PNN model, if a right output generates, it used as the values for diagnosing diabetes disease type 2. Figure 1 illustrates PNN model scheme for diagnosing diabetes type 2.

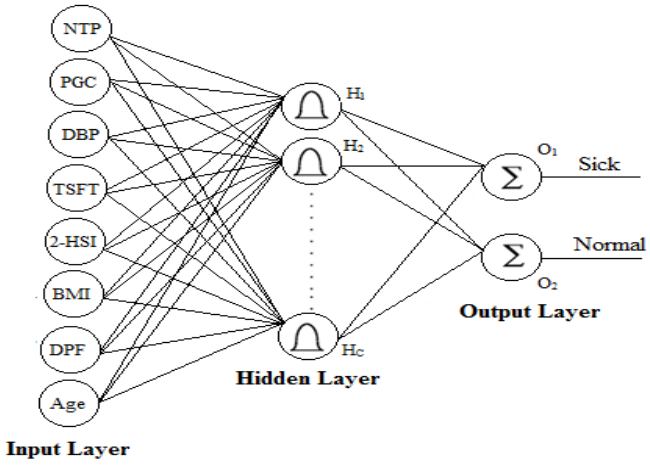

Fig. 1. Proposed PNN model scheme for diagnosing diabetes type 2

As shown in Figure 1, at input layer, the number of neurons is same as 8 features in Pima Indians Diabetes data set which showed in Table 1. As shown in Table 1 each feature's name has been written in every neuron. Number of neurons in hidden layer is denoted by Hc. Hc is same as number of training data sets. In the output layer, the number of neurons equals to 2 defined classes, that is, class 1 (sick) and class 2 (normal). Furthermore, we used MATLAB to implement PNN model for diagnosing diabetes type 2. Finally, according to the proposed model shown in Figure 1, Pima Indians Diabetes dataset, and number of samples for training and testing 768 data samples, achieved outputs of Table 3 and Figure 2 for minimizing mean squares error during the process of data training, and Figure 3 which shows training accuracy and testing accuracy.

TABLE III. MEAN SQUARES ERROR IN EACH ROUND FOR DIFFERENT VALUES OF NEURONS OF HIDDEN LAYER IN PNN MODEL

No. of neurons in hidden layer	MSE
0	0.898927
25	0.522076
50	0.490339
75	0.459006
100	0.42791
125	0.396591
150	0.374956
175	0.314095
200	0.299721

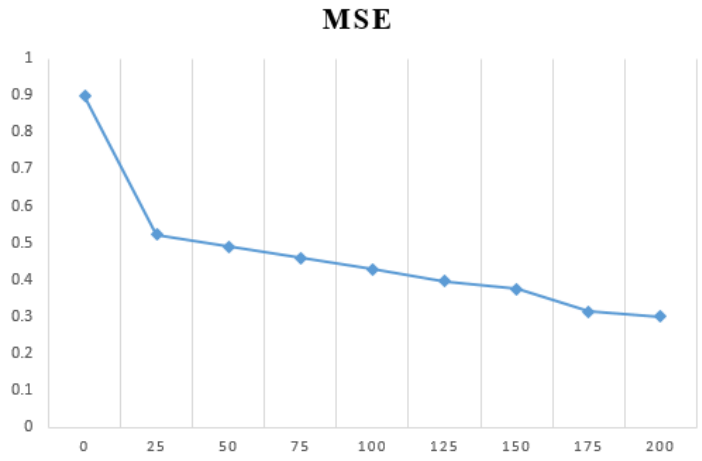

Fig. 2. Decreasing diagram of mean squares error for different values of neurons of hidden layer in PNN model

According to Table 3 and Figure 2, after 200 rounds, achieve its lowest possible value, 0.299721. Number of rounds is same as number of neurons in hidden layer. Number of neurons in hidden layer in each round increases by 25 units, and this continues until 200 neurons, such that mean squares error reaches its lowest possible value.

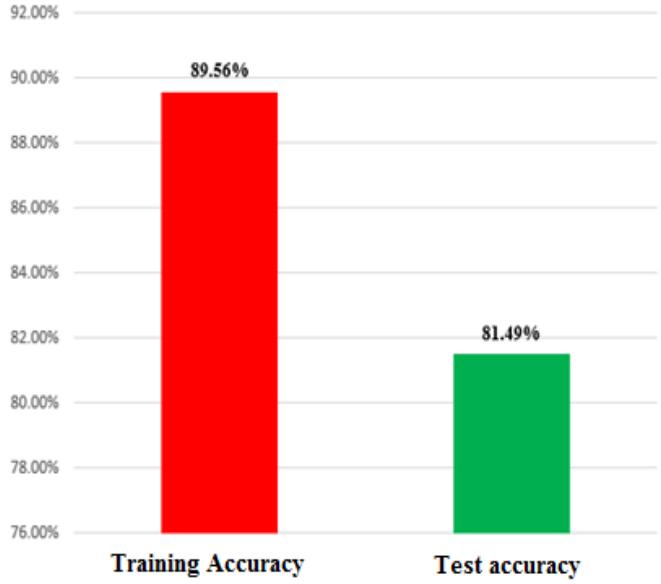

Fig. 3. Accuracy of diabetes type 2 diagnosis in training and testing Pima Indians Diabetes data set.

According to Figure 3 and Pima Indians Diabetes data set which consists 768 data samples, obtained accuracy rate of training phase for 90% of data (691 data samples) is 89.56%. Furthermore, obtained accuracy rate of testing phase for 10% of data (61 data samples) is 81.49%. Therefore, training the data has a higher accuracy rate than testing phase. However, both training and testing measures on Pima Indians Diabetes data sets using PNN model achieved a good accuracy rate.

IV. Discussion and Evaluation

In this section, we discuss and investigate results of diagnosing accuracy in training and testing phases as well as studies presented in section 2. Figures 4 and 5 show these comparisons.

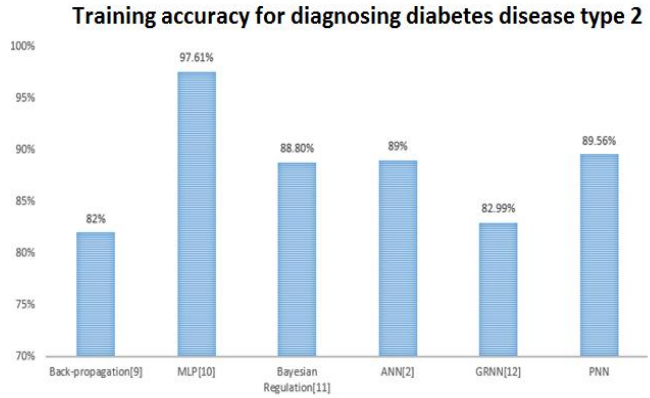

Fig. 4. Comparison of training accuracy of diagnosing diabetes type 2 between proposed and other approaches in training phase

As shown in Figure 4, all traditional studies with Pima Indians Diabetes data set except Bayesian Regulation [11] perform the task of identifying diabetes type 2. It is obvious that our method using PNN model outperforms other models such as back-propagation[9]،Bayesian Regulation[11], ANN[2], and GRNN[12] in terms of accuracy of diagnosing diabetes type 2. MLP [10] is the only model which has a higher accuracy than our PNN model. Back-propagation [9] with 82% accuracy is the worst approach. This value is close to the 82.99% accuracy which belongs to GRNN [12]. Furthermore, MLP [10] with 97.61% accuracy is the best approach in training phase. Also, accuracy of ANN [2] and PNN are too close to each other.

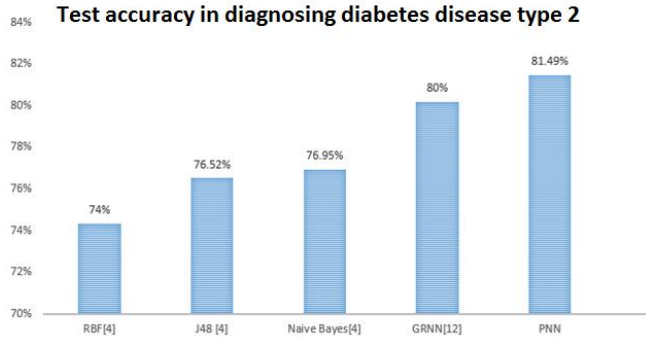

Fig. 5. Comparison of test accuracy of diagnosing diabetes type 2 between proposed and other approaches in testing phase

As shown in Figure 4, all traditional studies with Pima Indians Diabetes data set perform the task of identifying diabetes type 2. It is concluded that PNN model has the highest accuracy for diagnosing diabetes type 2 compared to other models. RBF [4] with 74% accuracy has the worst accuracy rate among the compared models. J48 [4], Naïve Bayses [4] are too close to each other.

V. Conclusion and Future Work

Advances in information and communication technologies leads to the use of artificial intelligence technologies in various fields including medicine science. By using artificial neural networks, we can design and implement the complex medical processes as software. These software systems in turn are effective for different fields of medicine sciences such as diagnosis, treatment and to help Surgeons, physicians and the public. These systems can be implemented in different scales in a parallel and distributed manner. In general, ANNs are parallel processing systems which are used for identifying complex patterns among data. Therefore, in this paper PNNs are applied to identifying diabetes disease type 2. We implemented the PNN model in MATLAB. The Pima Indians Diabetes data set was used for diagnosing diabetes type 2, which consists 768 data samples with 9 features. 90% of these 768 samples are used as training set and 10% used as testing set. The method achieved 89.56% of diagnosis accuracy in training phase, and 81.49% in test phase. Both training and testing measures could identify the diabetes disease type 2 with a good accuracy. As a future work, we will use the combination of fuzzy and artificial neural networks or combination of genetic and artificial neural networks for diagnosing diabetes type 2.

References

[1] M. Khashei, S. Eftekhari, J. Parvizian, “Diagnosing Diabetes Type II Using a Soft Intelligent Binary lassification Model”, *Review of Bioinformatics and Biometrics (RBB),* Vol. 1, Issue 1,(2012), pp.9-23.

[2] M. Durairaj, G. Kalaiselvi, “Prediction Of Diabetes Using Soft Computing Techniques- A Survey”, *international journal of scientific & technology research*, vol. 4, issue 03,(2015), pp.190-192.

[3] Type 2 Diabetes*: The Basics*, (2016), http://www.webmd.com/diabetes/type-2-diabetes-guide/type-2-diabetes.

[4] S. Sa'di, A. Maleki, R. Hashemi, Z. Panbechi, K. Chalabi, “Comparison of Data Mining Algorithms in the Diagnosis of Type Ii Diabetes*”, International Journal on Computational Science & Applications (IJCSA),* Vol.5, No.5,(2015), pp. 1-12.

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

[6] S. Sa'di, R. Hashemi, A. Abdollapour, K. Chalabi, M. A. Salamat, “A Novel Probabilistic Artificial Neural Networks Approach for Diagnosing Heart Disease”, *International Journal in Foundations of Computer Science & Technology (IJFCST)*, Vol.5, No.6, (2015), pp.47-53.

[7] F.S. Gharehchopogh, Z.A. Khalifelu, "Neural Network Application in Diagnosis of Patient: A Case Study*", IEEE, International Conference on Computer Networks and Information Technology(ICCNIT2011), Abbottabad, Pakistan*, (2011), pp. 245-249.

[8] H. Demuth, M. Beale, "Neural Network Toolbox for Use with MATLAB", User’s Guide, Version 4, The MathWorks, Inc. 3 Apple Hill Drive Natick, MA 01760-2098,840 pages, 2002.

[9] E. O. Olaniyi, K. Adnan, “Onset Diabetes Diagnosis Using Artificial Neural Network*”, International Journal of Scientific & Engineering Research*, vol.5, issue 10,(2014), pp. 754-759.

[10] A. Al-Rofiyee, M. Al-Nowiser, N. Al-Mufadi, M. A. AL-Hagery, “using prediction methods in data mining for diabetes diagnosis”, *posters*, (2014).

[11] S. Kumar, A. Kumaravel, “Diabetes Diagnosis using Artificial Neural Network”, *international journal of engineering sciences & research technology*, Vol.2, No.6,(2013), pp. 1642-1644.

[12] K. Kayaer, T. Yildirim, “Medical Diagnosis on Pima Indian Diabetes Using General Regression Neural Networks”, (2003).

[13] R. Raut, S. V. Dudul, "intelligent Diagnosis of Heart Diseases using Neural Network Approach", *International Journal of Computer Applications*, Vol.1, No.2,(2010), pp. 97-102.

[14] A. T. sayad, P. P. halkarnikar, "diagnosis of heart disease using neural network approach", *International Journal of Advances in Science Engineering and Technology*, Vol.2, No.3, (2014), pp.88-92.

[15] S. Bhande, R. Rau, "Parkinson Diagnosis using Neural Network: a Survey", *International Journal of Innovative Research in Science, Engineering and Technology*, Vol. 2, Issue 9, (2013) pp. 4843-4846.

[16] M. Can, "Neural Networks to Diagnose the Parkinson's disease", *southeast europe journal of soft computing*, Vol.2, No.1, (2013), pp.68-75.

[17] J .Kuruvilla, K. Gunavathi, "Lung cancer classification using neural networks for CT images, "*Computer Methods and Programs in Biomedicine*", Vol. 113, No.1, (2014), pp. 202-209.

[18] M .A. Hussain, T .M. Ansari, P .S. Gawas, N .N .Chowdhury, "Lung Cancer Detection Using Artificial Neural Network & Fuzzy Clustering", "*International Journal of Advanced Research in Computer and Communication Engineering"*, Vol. 4, No.3, (2015), pp.360-363.

[19] Pima Indians Diabetes Data Set, https://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes [Last Available: February 2016].

[20] S. N. Sivanandam, S. N Deepa, "Introduction to Neural Networks Using Matlab 6.0", *Tata McGraw-Hill Education*, ISBN: 978-0-07-059112-7, (2006), 656 pages.

[21] Neural network Toolbox, http://www.mathworks.com/products/neural-network/?s_tid=srchtitle [Last Available: February 2016].

[22] S .Moein, "Medical Diagnosis Using Artificial Neural Networks", *Part of the Research Essentials Collection*, DOI: 10.4018/978-1-4666-6146-2, 310 pages, 2014.