2018 21st International Conference of Computer and Information Technology (ICCIT),
21-23 December, 2018

Implementation of a Web Application to Predict Diabetes Disease: An Approach Using Machine Learning Algorithm

Samrat Kumar Dey
Department of CSE
Military Institute of Science and Technology
Dhaka, Bangladesh
samratcsepstu@gmail.com

Ashraf Hossain
Department of CSE
Dhaka International University
Dhaka, Bangladesh
asrafhossain197@gmail.com

Md. Mahbubur Rahman
Department of CSE
Military Institute of Science and Technology
Dhaka, Bangladesh
mahbubcse@yahoo.com

***Abstract*— Diabetes is caused due to the excessive amount of sugar condensed into the blood. Currently, it is considered as one of the lethal diseases in the world. People all around the globe are affected by this severe disease knowingly or unknowingly. Other diseases like heart attack, paralyzed, kidney disease, blindness etc. are also caused by diabetes. Numerous computer-based detection systems were designed and outlined for anticipating and analyzing diabetes. Usual identifying process for diabetic patients needs more time and money. But with the rise of machine learning, we have that ability to develop a solution to this intense issue. Therefore we have developed an architecture which has the capability to predict where the patient has diabetes or not. Our main aim of this exploration is to build a web application based on the higher prediction accuracy of some powerful machine learning algorithm. We have used a benchmark dataset namely Pima Indian which is capable of predicting the onset of diabetes based on diagnostics manner. With an accuracy of 82.35% prediction rate Artificial Neural Network (ANN) shows a significant improvement of accuracy which drives us to develop an Interactive Web Application for Diabetes Prediction.**

Keywords— Diabetes, SVM, ANN, Naïve Bayes, Min Max Scaling

I. INTRODUCTION

Diabetes is the fast-growing disease among the people even among the youngsters. Increase the level of the sugar (glucose) in the blood causes diabetes. Diabetes can be classified into two categories such as type 1 diabetes and type 2 diabetes. Type 1 diabetes is an autoimmune disease. In this case, the body destroys the cells that are essential to produce insulin to absorb the sugar to produce energy. And this kind of diabetes can cause obesity. The obesity is the increase in body mass index (BMI) than the normal level of BMI of an individual [1]. Type 1 diabetes can occur in childhood or adolescence age. Type 2 diabetes usually affects the adults who are obese. In this type, the body resists observing insulin or fails to produce insulin. Type 2 generally occurs in the middle or aged groups. Moreover, there are other causes for diabetes such as bacterial or viral infection, toxic or chemical contents in food, autoimmune reaction, obesity, bad diet, change of lifestyles, eating habit, environment pollution, etc. Diabetes leads to various diseases such as cardiovascular complications, renal issues, retinopathy, foot ulcers, etc. The Machine learning algorithms usually find out the hidden pattern from the large dataset and find out the desire approximate final result. Machine learning is a subfield of AI, And Machine learning algorithms can classify into three categories such as Supervised Learning, Unsupervised Learning and Reinforcement Learning. In our system, we use supervised learning algorithms for testing out accuracy among some sort of popular Machine Learning (ML) algorithms. Supervised Learning algorithms learn the pattern from pre-existing data and try to predict new result based on the previous learning. ML algorithms are to identify existing data like probability-based, function-based, rule-based, tree-based, instance-based, etc. Different Machine Learning algorithms are introduced using various data mining algorithms for assisting medical experts. The effectiveness of the decision support system is recognized by its accuracy. So the main aim of building a decision support system is to predict and diagnose a particular disease with a maximum degree of accuracy [2][3][4]. In this system, we use the pre-existing data sets called Pima Indian to train and evaluate our model which is an open source dataset.

We have systematized the residuum of this research as follows. In Section II, we discuss some relevant work which was performed previously. Our proposed architecture has been demonstrated in Section III. Section IV presents the insight description of our Methodology and different algorithms. Design of application and implementation are described in Section V. Experimental procedure and discussion of result analysis are presented in Section VI. Utterly, we conclude the paper by presenting future work in Section VII.

II. RELAVENT WORK

This section demonstrates some of the previous research works and these are co-respond to proposed work. Author in Reference [5] proposed that the given information is pre-handled to separate all the metadata. KNN is utilized to locate the nearest neighbors of the given dataset. On the other hand, if the desired level is found then algorithms stops execution, if it's not then the classifier of the system is applied. According to Reference [1] Naive Bayes (NB) which is a precise machine learning algorithm is used to utilize arrange Arabic web documents. K-Nearest Neighbor classification has been studied for estimating economic conditions. The economics suffering and insolvency will sustain a deliberate distance can be estimating using utilizes of KNN technique. The amount of insolvent organization has risen after the circumstances of worldwide emergency of economic.

978-1-5386-9242-4/18/$31.00 ©2018 IEEE

Proportions for anticipating economic distress have pulled in an assortment of concern for scholastics and economic and financial institutions. Authors in reference [6] proposed that withdrawing insulin from the human body which is leads carries glucose into the blood cell and Diabetes is a lifelong disease. Diabetes also influences some sort of difficulties like stroke, heart disease, blindness, kidney failure loss of weight, buried vision etc. In reference [7] the Artificial Neural Network is broadly used in the research field because of its scalable ability of higher dimensional data and very complex models. Basically, its representing human neural construction demonstrate in a mathematical way with capable of learning and generalization.

III. Proposed Architecture

Supervised Learning algorithms learns the pattern from pre-existing data and try to predict new result based on the previous learning. ML algorithms are used to identify existing data like probability-based, function-based, rule-based, tree-based, instance-based, etc. Following Fig.1. indicates the ample architecture of our proposed Model. According to our model patients are required to provide their medical data for successful diagnosis of their diabetes test.

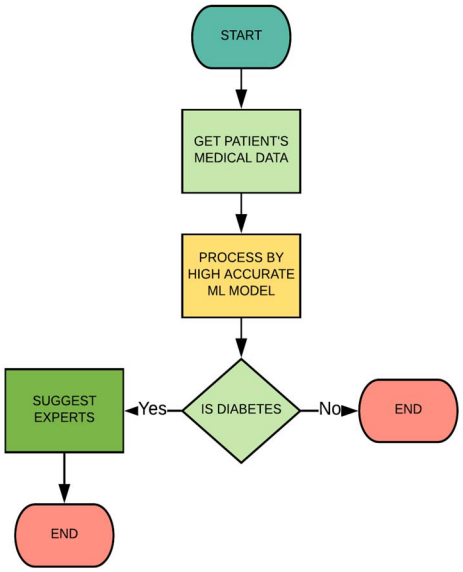

Fig. 1. Flowchart of Diabetes prediction Model

IV. Methodologies

A. *Support Vector Machine (SVM)*

Support vector machine is a supervised learning algorithm mainly used for classification problem. Overfiting derives an ML model to misclassify the data from a given dataset. In that case SVM can prevent overfitting nature from samples data and produce better accuracy [8][9]. SVM possess a linear hyperplane with a margin which divides the dataset into positive and negative samples [8][9].

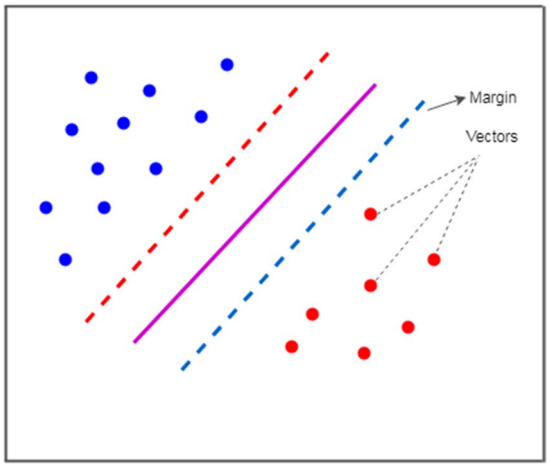

Fig. 2. Graphical representation of SVM working procedure

SVM select the hyper planes with maximum possible distance [9]. The SVM decision boundary in mathematical expression as follows:

$$min\frac{1}{2}\sum_{j=1}^{n}\theta_j^2$$

Where,

$$\theta^T x^{(i)} \geq 1 \quad if\ y^{(i)} = 1$$
$$\theta^T x^{(i)} \leq -1 \quad if\ y^{(i)} = 0$$

B. *K- Nearest Neighbors(KNN)*

K-Nearest Neighbors is a supervised learning algorithm, K- means a number of a vector. The working methodology of KNN is pretty simple, it's predict based on the value of K-parameter. Graphical representation of K Number of Nearest Neighbors are depicted in following Fig. 3.

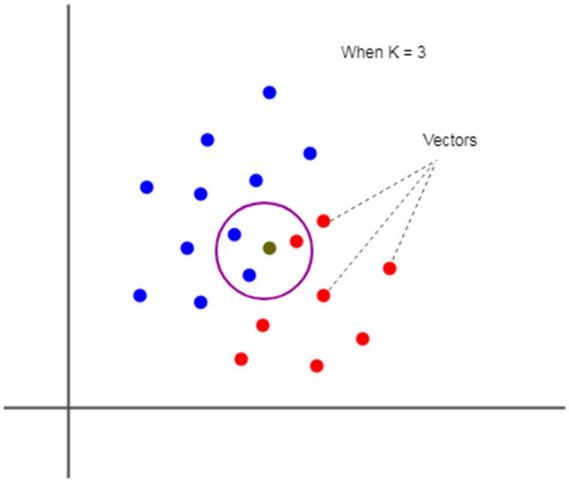

Fig. 3. Graphical representation of KNN working Procedure

From the above figure when K is 3 then it tries to catch up with three nearest neighbor vectors. In next steps KNN finds the highest ratio of neighbor vector and then it finally produces the result for the input vector.

C. *Naive Bayes Algorithm*

The Naive Bayes Algorithm is another Machine Learning algorithm for classification problems. Naive Bayes is an efficient classification algorithm in data mining that can

handle missing values during classification [8] [9]. Naive Bayes Algorithm is pretty fast Machine Learning and efficient model, basically, this model is used for text classification few known examples are spam filtration, sentimental analysis and classifying new articles. The named of Naive is called for it's some sort of features distinct of an event of another feature. And Bayes refers to the statistician and philosopher Thomas Bayes theorem [10]. The NB theorem can be expressed mathematically as follows:

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}$$

- P (A | B): Probability of occurrence of event A given the event B is true.
- P (A) and P (B): Probabilities of the occurrence of event A and B respectively.
- P (B | A): Probability of the occurrence of event B given the event A is true.

Bayes Theorem for Naive Bayes Algorithm state the following relationship

$$P(C_i|x_1,x_2, \ldots\ldots\ldots, x_n) = \frac{P(x_1,x_2,\ldots\ldots\ldots,x_n \mid C_i)}{P(x_1,x_2,\ldots\ldots\ldots,x_n)} \quad for\ 1 \leq i \leq k$$

This relationship can be simplified as follows

$$P(C_i \mid x_1, x_1, \ldots\ldots, x_n) = \left(\prod_{j=1}^{j=n} P(x_j \mid C_i)\right) \cdot \frac{P(C_i)}{P(x_1, x_1, \ldots\ldots, x_n)} \quad for\ 1 \leq i \leq k$$

D. *Artificial Neural Network (ANN)*

Artificial Neural Network (ANN) is considered as the method to represent human neuron in a Mathematical way by reflecting its learning and generalization abilities. ANN model can scale up a highly nonlinear system in which the relationship between the variables is unknown or very complex [11].

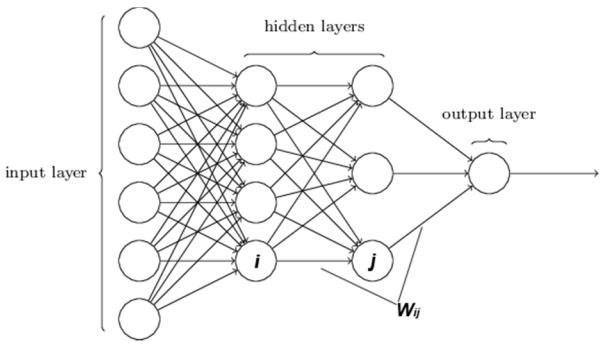

Fig. 4. General Structure of Artificial Neural Network with two hidden layer

A neural network is consist of various neurons followed by some layers. Nodes represent the structure of human neuron as like dendrite and axon where linked between nodes represents as axon with the weighted connection. The total structure of the neural network is shaped by input layer, one or more hidden layers, and the output layer where i^{th} neuron represents a connection with j^{th} neuron of total structure and W_{ij} represents as the strength of the link between neurons. The nodes of the structure of an ANN take inputs (features) then send them to next hidden layer through some sort of weighted link where i^{th} nodes send data to j^{th} nodes for processing and calculating the weighting sum and totaling a bias term (θ_j) [11]. Mathematical representation of above discussion can be expresses as follows:

$$net_j = \sum_{i=1}^{m} x_i * w_{ij} + \theta_j \ (j = 1, 2, \ldots, n)$$

V. Design and Implementation

As we have mentioned earlier our goal is to develop a Web-based application which can predict diabetes based on patients health data. In order to find most suitable ML algorithm that is capable of predicting diabetes more precisely, we have tested some sort of powerful machine learning models like SVM, KNN, Naive Bayes and ANN. For the successful evaluation of these models, we have used some machine learning libraries such as Scikit-learn, numpy, matplotlib, pandas, and tensorflow.js. In order to get rid of overfitting problem we split up the dataset into two subparts: one is for testing and another is for training. Based on different training dataset size we have achieved the accuracy rate for every defined Model. However, preprocessing of Indian Pima Dataset can produce higher prediction accuracy. Therefore, we also calculate the accuracy for our defined model in order to improve the prediction accuracy. Data normalization is an effective way to increase the accuracy of certain machine learning models and some machine learning model do not perform well without data Normalization. In our proposed model we used Min Max Scaler (MMS) as Normalization model. MMS basically scale down the data in range of [0, 1] or [-1, 1]. The Mathematical formula for Min Max Scaling can be represented as follows:

$$x' = \frac{x - min(x)}{max(x) - min(x)}$$

By using Min Max Scaler (MMS) preprocessing method with every defined model we have achieved higher prediction accuracy than previous calculation. For the implementation of our designed architecture we have developed a web application using tensorflow.js based on highest prediction accuracy. Following Fig. 5. shows that how a machine learning function called *SetClassifiers* is used to set specific classifier in Scikit Learn. Accuracy Calculation of defined algorithm is depicted in Fig.6. where a method called *getScores* is used.

```
def SetClassifires():
    svm_clf = svm.SVC()
    knn = neighbors.KNeighborsClassifier(n_neighbors=3)
    gnb = GaussianNB()
    ann = MLPClassifier(solver='lbfgs',
                        alpha=1e-5,
                        hidden_layer_sizes=(5, 2),
                        random_state=1)
```

Fig. 5. Set Classifier function for the selection of Classifier in Scikit Learn

```
def getScores(X_train, y_train, X_test, y_test):
    SVM_Score = svm_clf.fit(X_train, y_train).score(X_test, y_test)
    KNN_Score = knn.fit(X_train, y_train).score(X_test, y_test)
    GNB_Score = gnb.fit(X_train, y_train).score(X_test, y_test)
    ANN_Score = ann.fit(X_train, y_train).score(X_test, y_test)
```

Fig. 6. Score Calculation code snippet of different algorithms

VI. RESULTS

In this section we will discuss regarding our results which we have achieved after experimental design. Following TABLE I. represents an insight description of our Pima Indian Dataset. This dataset is mainly based on the females those were living at Pima Indian heritage. Following 8 features ***(a-h)*** of Pima Indian dataset helps us to predict the diabetes of any Individuals with the help of our proposed methodologies.

a) Numbers of time Pregnant
b) Glucose Test
c) Blood Pressure
d) Triceps skinfold thickness
e) 2-Hour Serum Insulin
f) Body Mass Index
g) Diabetes Pedigree function
h) Age

TABLE I. DIFFERENT ATTRIBUTES OF INDIAN PIMA DATASET [12]

Class	Attribute Number
Pregnancy Count	1
Glucose concentration in plasma	2
Blood pressure (diastolic, mm Hg)	3
Thickness of triceps skin fold (mm)	4
2-Hour serum insulin (μ U/ml)	5
Body mass index	6
Pedigree function of diabetes	7
Years of age	8

Following TABLE II depicts the average accuracy rate varies with training dataset size. We have experimented with different size of training data for SVM, KNN, GNB and ANN. Without Min Max Scaling (MMS) it shows an average accuracy of 76.25% with Gaussian Naïve Bayes Algorithm.

TABLE II. ACCURACY OF DIFFERENT MACHINE LEARNING ALGORITHMS BASED ON PIMA INDIAN DATASET

Training Dataset Size	SVM	KNN	GNB	ANN
368	63	64	76	63
468	63	68	77	63
568	63	68	76	63
668	63	66	76	63
Average	**63**	**66.5**	**76.25**	**63**

In order to improve the detection accuracy we have performed some data preprocessing by using Min Max Scaler Method. Following TABLE III. shows that by using Min Max Scaling Method we have achieved more higher accuracy than previous calculation in TABLE II. According to TABLE III. Artificial Neural Network achieved 82.35% detection accuracy based on the diagnosis nature of Indian Pima Dataset. Therefore we have built a web application using ANN Model which is capable of predicting whether a patients has diabetes or not.

TABLE III. ACCURACY OF ALGORITHMS WITH MIN MAX SCALER METHOD

Training Dataset Size	SVM + MMS	KNN + MMS	GNB + MMS	ANN + MMS
368	78	75.3	79.1	81.8
468	78	76	79.3	82.3
568	78.2	75.1	79.3	82.9
668	78	75.6	79.5	82.4
Average	**78.05**	**75.5**	**79.3**	**82.35**

Following Fig. 7. Represents our developed web application that has been developed based on highest accuracy of ANN model with the support of Min Max Scaling Methods. Here we have used PHP Web programming language as backend development, JavaScript frameworks as frontend and Tensorflow.js for the code implementation of Machine Learning Model. Our proposed architecture usually collects values of dataset from SQL database and train the Artificial Neural Network model during training session. During the period of prediction, users' needs to provide some information as input shown in following Fig. 7. so that our developed web application can predict whether the test result is either positive or not. In order to test the diabetes users needs to provide following information in web application. Some of very necessary information like Blood Pressure, Body Mass Index (BMI), Serum Insulin, Oral Glucose Tolerance Test, etc. are required.

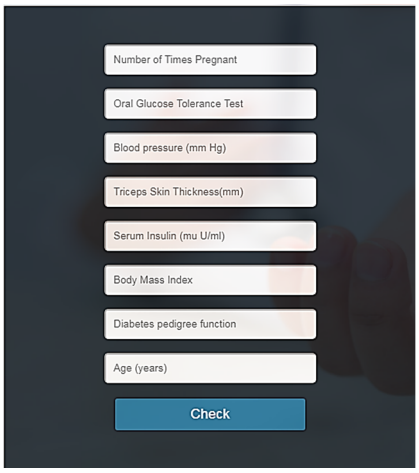

Fig. 7. Our developed web application based on ANN MMS detection Model.

In order to evaluate the detection accuracy of our model we compared our work with some of state of the art research work. Following TABLE IV. enlists some of the research work which driven with different machine learning methodologies for different medical data prediction. From the following it is clear that our proposed model have highest accuracy than other research work. Graphical representation of TABLE II and TABLE III is depicted in following Fig. 8 and Fig.9.

TABLE IV. COMPARISON OF OUR PROPOSED MODEL WITH OTHERS

Model Name	Accuracy
Gaussian Naive Bayes (GNB) [13]	76.52%
General Regression Network (GRNN) [14]	80.21%
Backpropagation Genetic Algorithm (BGA) [15]	74.80%
Fuzzy Min Max (FMM) [16]	69.28%
Our Proposed Model (ANN with MMS)	**82.35%**

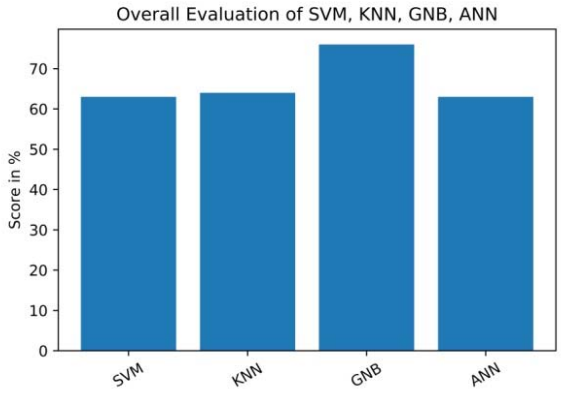

Fig. 8. Graphical representation of Accuracy of different machine learning algorithm

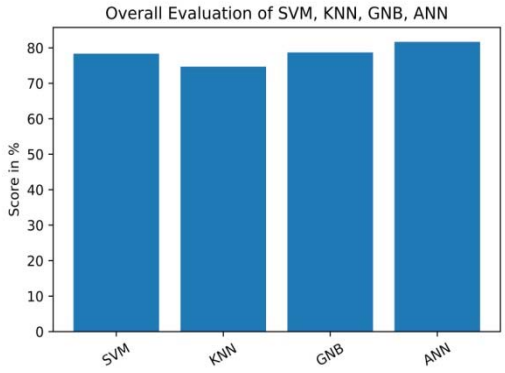

Fig. 9. Graphical representation of Accuracy of different machine learning algorithm with Min Max Scaler Method

VII. Conclusion

In this paper, we proposed a web based application for the successful prediction of Diabetes Diseases. From different machine learning algorithms Artificial Neural Network (ANN) provide us highest accuracy with Min Max Scaling Method on Indian Pima Dataset. As we have proposed and developed an approach for diabetes disease prediction using machine learning algorithm, it has significant potential in the field of medical science for the detection of various medical data accurately. In near future our focus is to use a deep learning model and prepare a Location based Dataset from real medical data for the successful prediction of diabetes disease.

References

[1] M. Elkourdi, A. Bensaid, and T. Rachidi, “Automatic Arabic Document Categorization Based on the Naïve Bayes Algorithm,” Alakhawayn University, 2001.

[2] S. Kumari and A. Singh, "A data mining approach for the diagnosis of diabetes mellitus," 2013 7th International Conference on Intelligent Systems and Control (ISCO), Coimbatore, 2013, pp. 373-375.

[3] R. Motka, V. Parmarl, B. Kumar and A. R. Verma, "Diabetes mellitus forecast using different data mining techniques," 2013 4th International Conference on Computer and Communication Technology (ICCCT), Allahabad, 2013, pp. 99-103.

[4] V. Vijayan, and A. Ravikumar, “Study of Data Mining algorithms for Prediction and Diagnosis of Diabetes Mellitus,” International Journal of Computer Application, Vol 94, pp .12-16, June 2014.

[5] H. Shee, K. W Cheruiyot and S. Kimani, “Application of k-Nearest Neighbour Classification in Medical Data Mining,” International Journal of Information and Communication Technology Research, Volume 4, No. 4, April 2014.

[6] C. Kalaiselvi and G. M. Nasira, "A New Approach for Diagnosis of Diabetes and Prediction of Cancer Using ANFIS," 2014 World Congress on Computing and Communication Technologies, Trichirappalli, 2014, pp. 188-190.

[7] I. Aleksander, and H. Morton, “An introduction to neural computing,” Int Thomson Computer Press, London 1995.

[8] P. Tan, M. Steinbach, V. Kumar, “Introduction to Data Mining,” Pearson Education, Inc,2006.

[9] V. Chandra S.S, and A. Hareendran S, “Artificial Intelligence and machine learning,” PHI learning Private Limited, Delhi 110092, 2014.

[10] R. Jain, “Introduction to Naive Bayes Classification Algorithm in Python and R,” February 2, 2017. [Online]. Available: https://www.hackerearth.com/blog/machine-learning/introduction-naive-bayes-algorithm-codes-python-r/. [Accessed Sept. 22,2018].

[11] F. Amato, A. López, E. M. Peña-Méndez, P. Vaňhara, A. Hampl and J. Havel, “ Artificial neural networks in medical diagnosis,” 2013.

[12] F.S Panchal, A. Ganatra, Y. Kosta, and M. Panchal, “Review on Methods of Selecting Number of Hidden Nodes in Artificial Neural Network,” International Journal of Computer Theory and Engineering, vol. 3, no. 2, pp. 332–337, 2011.

[13] D.k. Chouby, S. Paul, S. Kumar, and S. Kumar, “Classification of Pima indian diabetes dataset using naive bayes with genetic algorithm as an attribute selection,” The International Conference on Communication and Computing Systems, ICCCS-2016.

[14] K. Kayaer, and T. Yildirim, “Medical Diagnosis on Pima Indian Diabetes Using General Regression Neural Networks,” Proceedings of the International Conference on Artificial Neural Networks and Neural Information Processing, pp. 181–184, 2003.

[15] H. Hasan Örkcü, and H. Bal, “Comparing Performances of Backpropagation and Genetic Algorithms in the Data Classification,” Expert Systems with Applications, vol. 38, 2011, pp. 3703-3709.

[16] M. Seera, and C.P. Lim, “A hybrid intelligent system for medical data classification,” Expert Systems with Applications, Vol. 41 pp. 2239–2249.