Volume 2, Issue 12, December 2012 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

# Diabetes Diagnosis by Using Computational Intelligence Algorithms

Najmeh Hosseinpour
Department of Computer Engineering, Dezfoul Branch, Islamic Azad University, Dezfoul, Iran

Saeed Setayeshi
Faculty of Nuclar Engineering and Physics, Amirkabir University, Tehran, Iran

Karim Ansari-asl
Faculty of Technical and Engineering, Shahid Chamran University
Ahwaz, Iran

Mohammad Mosleh
Department of Computer Engineering, Dezfoul Branch
Islamic Azad University
Dezfoul, Iran

***Abstract— Diabetes mellitus is a chronic disease and one of the most public health challenges in worldwide. Most of discoveries indicate that the best way to overcome diabetes is to prevent the risks of diabetes before becoming a diabetic. With this idea, we would like to find a way to estimate diabetes risk. Data mining techniques could be used as an alternative way in discovering knowledge from the patient medical records and they have shown remarkable success in the area of applying Computer Aided Diagnostic (CAD) systems. In this paper, we have applied several intelligence classifiers such as Bayesian, Functional, Rule-base, Decision Trees and Ensemble for diagnosing diabetes mellitus. Experimental results on Pima Indian Diabetes (PID) dataset show that Bagging ensemble classifier with Logistic core has better performance in comparison with other presented classifiers***

**Keywords**— ***Diabetes mellitus, Machine learning, Classifier, Pima Indian Diabetes (PID)***

## I. INTRODUCTION

Diabetes mellitus is a chronic debases and is countered as one of the most challenging in the World Health Organization (WHO). International Diabetes Federation (IDF) estimates that more than 285 million persons across from the world have affected by diabetes. It is expected until 20 next years, this amount get to 438 million persons. Diabetes mellitus can be divided in two types: Type I (formally known as insulin-dependent diabetes) and Type II (formally known as non insulin-dependent diabetes). In the Type I, person pancreas cannot produce insulin whereas in the Type II, person pancreas can discharge insulin but its absorption amount with body is very low. Type II diabetes mellitus, which will be discussed in this paper, is the most important diabetes type why includes more than 90-95 percentage of diabetes. Amongst problems of the persons who affected by diabetes mellitus can be pointed to stricture of vessels, mental disorders, kidney disease and etc. The main problem related to this destructive and dangerous disease is lack of early detection or weakness in diagnosis which causes patients when understand it, that is perhaps little late for control or treatment. Whereas diabetes mellitus diagnosis performs by using different factors or features, therefore computational intelligent techniques can be accounted as alternative way for acquisition of knowledge from these data. Up to now, several papers have been presented for this purpose [1-9]. In this paper, we have applied intelligence classifiers such as Bayesian, Functional, Rule-based, Decision Trees and Ensembles classifiers for diabetes mellitus diagnosis. The obtained results of experiments show that Decision Trees and Ensemble Classifiers have better accuracy from than other mentioned methods [1-9].

The paper structure is as following. In the section 2, introduction to machine learning will be discussed. Experimental results and evaluations will be presented in the section 3. Finally, paper will be finished by conclusion in section 4.

## II. INTRODUCTION TO MACHINE LEARNING

In general, machine learning techniques are divided in two main categories: unsupervised and supervised methods [10]. In the unsupervised methods, there is no target variable and therefore the method searches correlations and structures among all samples. *Clustering* technique is one of the most important unsupervised methods. Clustering process tries to divide a set of data to several clusters so data placed in a cluster are similar to each other and different from other clusters. Unlike unsupervised method, in the supervised methods, there is a predefined target variable. For this purpose, at first, a learning machine model is constructed by helping set of patterns that their target variables are exist (training set). This machine can recognize target variable a new pattern that there is not in the training set. Amongst the most common supervised machine learning techniques, we can point to *Classification* technique. Up to now, several classifiers have been presented which Bayesian law, Artificial Neural Networks (ANN), Support Vector Machine (SVM), Decision Trees and Ensemble based classifiers are among the most important of them.

**Bayesian Classifier:** A Bayesian network or belief network is a probabilistic graphical model (a type of statistical model) that represents a set of random variables and their conditional dependencies via a directed acyclic graph (DAG). For example, a Bayesian network could represent the probabilistic relationships between diseases and symptoms. Given symptoms, the network can be used to compute the probabilities of the presence of various diseases. Formally, Bayesian networks are directed acyclic graphs whose nodes represent random variables in the Bayesian sense: they may be observable quantities, latent variables, unknown parameters or hypotheses. Edges represent conditional dependencies; nodes which are not connected represent variables which are conditionally independent of each other. Each node is associated with a probability function that takes as input a particular set of values for the node's parent variables and gives the probability of the variable represented by the node [10].

**Artificial Neural Network (ANN)**: An Artificial Neural Network (ANN) is a mathematical or computational model that is taken by the structure and/or functional aspects of biological neural networks. A neural network consists of an interconnected group of artificial neurons, and it processes information using a connectionist approach to computation. In most cases an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase. Modern neural networks are non-linear statistical data modeling tools. They are usually used to model complex relationships between inputs and outputs or to find patterns in data. ANNs that include more than one layer of neurons are called multi-layer neural networks. These networks are divided in two classes feed forward neural network and feedback neural networks [11].

The feed forward neural network was the first and arguably simplest type of artificial neural network devised. In this network, the information moves in only one direction, forward, from the input nodes, through the hidden nodes (if any) and to the output nodes. There are no cycles or loops in the network. Multilayer Perceptron (MLP) and Radial Basis Function (RBF) are famous types of these networks.

A recurrent neural network is a class of neural network where connections between units form a directed cycle. This creates an internal state of the network which allows it to exhibit dynamic temporal behavior. Unlike feed forward neural networks, RNNs can use their internal memory to process arbitrary sequences of inputs. Hopfield neural network and Time Delay Neural Network (TDNN) are popular types of these networks.

**Support Vector Machines (SVMs):** Support Vector Machines (SVMs) are supervised learning models with associated learning algorithms that analysis data and recognize patterns, used for classification and regrission. The basic SVM takes a set of input data and predicts, for each given input, which of two possible classes forms the output, making it a non-probabilistic binary linear classifier. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separated categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on. In addition to performing linear classification, SVMs can efficiently perform non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces [12].

**Decision Trees:** Decision Tree is a hierarchical data structure implementing the divide-and-conquer strategy. A decision tree is composed of internal decision nodes and terminal leaves. Each decision node $m$ implements a test function $f_m(x)$ with discrete outcomes labeling the branches is taken depending on the outcome. This process, starts at the root and is repeated recursively until a leaf node is hit, at which point the value written in the leaf constitutes the output. Each $f_m(x)$ defines a discriminant in the $d$-dimension input space dividing it into smaller regions which are further subdivided as we take a path from the root down. $f_m(.)$ is a simple function and when written down as a tree, a complex function is broken down into a series of simple decisions. Different decisions tree methods assume different methods for $f_m(.)$, and the model class defines the shape of the discriminant and the shape of regions. Each leaf node has an output label, which in the case of classification is the class code and in regression is a numeric value. A leaf node defines a localized region in the input space where instances falling in this region have the same output. The boundaries of the regions are defined by the discriminants that are coded in the internal nodes on the path from the root to the leaf node[13].

**Ensemble Classifier:** An ensemble classifier consists of a set of single classifiers so that the final output dependent on each of its constructor classifiers. The main idea of ensemble classifiers is that, we don't train a single classifier rather we train a set of classifiers and therefore we will have the combination of several classifiers prediction. We can express motivation of this work as following:

- Variance reduction: This means that the results are less dependent on the features of a single training set.
- Bias reduction: This means that the combination of several classifiers might have shown more the concept of class in comparison with a single classifier.

Up to now, several ensemble classifiers such as Bagging, Boosting and Random Forest have been presented [14].

## III. EXPERIMENTAL RESULTS

In this section, the obtained results from applying intelligence classifiers for diabetes mellitus diagnosis on PID dataset have been presented [15]. The PID dataset is one of the standard machine learning datasets of UCI repository which contains information of 768 samples. All patients in this dataset are Pima Indian women at least 21 years old and living near Phoenix, Arizona, USA. Each sample has 8 features as follows:

- *Number of times pregnant*
- *Plasma glucose concentration a 2 h in an oral glucose tolerance test*
- *Diastolic blood pressure (mm Hg)*
- *Triceps skin fold thickness (mm)*
- *2-hour serum insulin (mu U/ml)*
- *Body mass index (kg/m^2)*
- *Diabetes pedigree function*

 

- *Age (years)*

The binary target variable takes the values '0' or '1'. While '1' means a positive test for diabetes , '0' is a negative test. There are 268 cases in class '1' and 500 cases in class '0.' The PID statistical has been shown in Table 1.

**Table 1.** PIMA dataset Statistical

| Features | Mean | Standard Deviation | Min/Max |
|---|---|---|---|
| *Number of times pregnant* | 3.8 | 3.4 | 0 / 17 |
| *Plasma glucose concentration a 2 h in an oral glucose tolerance test* | 120.9 | 32 | 0 / 199 |
| *Diastolic blood pressure (mm Hg)* | 69.1 | 19.4 | 0 / 122 |
| *Triceps skin fold thickness (mm)* | 20.4 | 16 | 0 / 99 |
| *2-hour serum insulin (mu U/ml)* | 79.8 | 115.2 | 0 / 846 |
| *Body mass index (kg/m^2)* | 32 | 7.9 | 0 / 67.1 |
| *Diabetes pedigree function* | 0.5 | 0.3 | 0.078 / 2.42 |
| *Age (years)* | 33.2 | 11.8 | 21 / 81 |

In order to evaluate intelligence classifiers, at first, it is necessary to introduce confusion matrix and then express evaluation criterions as follows:

| | | Actual Class | |
|---|---|---|---|
| | | **Positive** | **Negative** |
| **Predicted Class** | **Positive** | *True Positive* | *False Positive* |
| | **Negative** | *False Negative* | *True Negative* |

**Fig1.** Simple confusion matrix

**True Positive (TP):** This parameter indicates the number of diabetes people that the system has properly detected them as diabetes.
**False Positive (FP):** This parameter indicates the number of healthy people that the system has wrongly detected them as diabetes.
**False Negative (FN):** This parameter indicates the number of healthy people diabetes that the system has wrongly detected them as diabetes
**True Negative (TN):** This parameter indicates the number of healthy people that the system has properly detected them as healthy
Therefore, measures such as *True Positive rate*, *False Negative rate*, *Precision*, *Recall*, *F-measure* and *Classification rate* are defined as followings:

$$\text{False Negative(FN) rate} = \frac{FN}{FN+TN} \times 100(\%) \quad (1)$$

$$\text{Precision} = \frac{TP}{TP+FP} \times 100(\%) \quad (2)$$

$$\text{Recall} = \frac{TP}{TP+FN} \times 100(\%) \quad (3)$$

$$\text{F-Measure} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (5)$$

$$\text{Classification rate} = \frac{TN+TP}{TP+FP+TN+FN} \times 100(\%) \quad (6)$$

As you know, precision is the fraction of retrieved instances that are relevant, while recall is the fraction of relevant instances that are retrieved. A measure that combines precision and recall is the harmonic mean of precision and recall, the traditional F-measure [16].

In order to compute mentioned criterions, we have used from WEKA software. WEKA is a popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand [17]. This software is able to perform pre-processing, feature selection and classification operations well. In the following, the obtained results from applying different classifiers of WEKA on diabetes dataset PIMA have been presented. In order to evaluate WEKA classifiers, measures False Negative rate, Precision, Recall, F-measure and Classification rate with 10-Fold cross validations (CV) have been computed.

In K-Fold cross validation, data set is divided into K partitions, classifier is run K times, using a different partition as test set each time, with the other K-1 as training set.

The results of evaluating Classifiers of Bayesian, Functional, Rule based, Decision Trees and Ensemble are shown in Table 2-6, respectively.

**Table2. The obtained result of applying Bayesian Classifiers on PIMA dataset with 10F CV**

| Measures | Bayesian Classifiers | | | |
|---|---|---|---|---|
| | **BayNet** | **Naïve Bays** | **Naïve Bayesian Multinomial** | **Bayesian Logistic Regression** |
| FN rate | 39.2 | 38.4 | 100 | 0.59 |
| Precision | 79.5 | 80.3 | 65.1 | 73.8 |
| Recall | 81.6 | 84.2 | 100 | 89.2 |
| F-measure | 80.6 | 82.2 | 78.9 | 80.8 |
| Classification rate | **74.34** | **76.3** | **65.1** | **72.39** |
| Time(ms) | 190 | 20 | 20 | 50 |

**Table3. The obtained result of applying Functional Classifiers on PIMA dataset with 10F CV**

| Measures | Functional Classifiers | | | |
|---|---|---|---|---|
| | **LibSVM** | **MLP** | **RBF Network** | **Logistic** |
| FN rate | 51.1 | 39.2 | 45.9 | 42.9 |
| Precision | 77.1 | 79.8 | 77.9 | 79.3 |
| Recall | 92 | 83.2 | 86.8 | 88 |
| F-measure | 83.9 | 81.5 | 82.1 | 83.4 |
| Classification rate | **76.95** | **75.39** | **75.39** | **77.21** |
| Time(ms) | 610 | 3440 | 200 | 30 |

**Table4. The obtained result of applying Rule-based Classifiers on PIMA dataset with 10F CV**

| Measures | Rule based Classifiers | | | |
|---|---|---|---|---|
| | **Conjunctive Rule** | **Decision Table** | **NNge** | **PART** |
| FN rate | 46.6 | 47 | 43.3 | 41.8 |
| Precision | 75.5 | 76.3 | 78.1 | 79 |
| Recall | 77 | 81 | 82.6 | 84.4 |
| F-measure | 76.2 | 78.6 | 80.3 | 81.6 |
| Classification rate | **68.75** | **71.22** | **73.56** | **75.26** |
| Time(ms) | 30 | 140 | 310 | 90 |

**Table 5. The obtained result of applying Decision Trees Classifiers on PIMA dataset with 10F CV**

| Measures | Decision Tree Classifiers | | | |
|---|---|---|---|---|
| | **J.48** | **Decision Stump** | **RepTree** | **LMT** |
| FN rate | 40.3 | 45.2 | 42.2 | 44 |
| Precision | 79 | 77.7 | 78.9 | 79 |
| Recall | 81.4 | 76.9 | 84.4 | 89 |
| F-measure | 80.2 | 78.7 | 81.5 | 83.7 |
| Classification | **73.82** | **71.85** | **75.13** | **77.47** |

| | | | | |
|---|---|---|---|---|
| rate | | | | |
| Time(ms) | 50 | 20 | 30 | 5360 |

**Table 6. The obtained result of applying Ensemble Classifiers on PIMA dataset with 10F CV**

| | Ensemble Classifiers | | | |
|---|---|---|---|---|
| **Measures** | **Bagging (with Logistic core)** | **AdaBoost (with Logistic core)** | **Random Forest** | **Multiclass Classifier (with Logistic core)** |
| FN rate | 42.5 | 42.9 | 47.8 | 42.9 |
| Precision | 79.5 | 79.3 | 76.9 | 79.3 |
| Recall | 88.2 | 88 | 85 | 88 |
| F-measure | 83.6 | 83.4 | 80.7 | 83.4 |
| Classification rate | **77.47** | **77.21** | **73.56** | **77.21** |
| Time(ms) | 33 | 61 | 310 | 30 |

The average F-measure of Bayesian, Functional, Rule-based, Decision Trees and Ensemble classifiers have been presented in Fig 1.

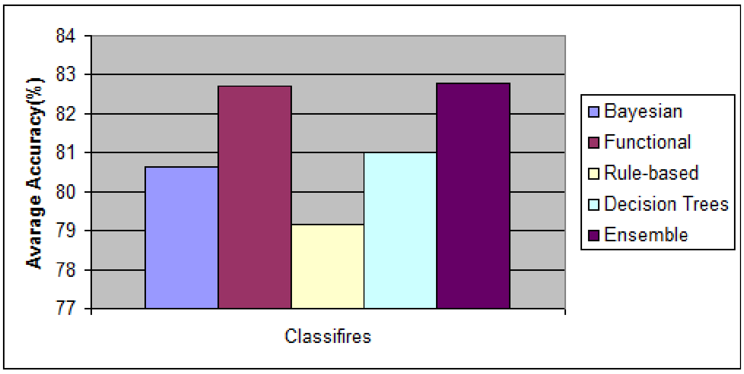


**Fig 1.** Showing average F-measure of different classifiers

As can be seen from Fig 1, the average F-measure of Functional and Ensemble classifiers are better than other classifiers about 83%. Also, the worst F-measure is related to Rule-based classifiers with 79.17%.

In the Fig 2, we have illustrated average classification rate different classifiers. As can be observed, the average classification rate of Functional and Ensemble classifiers are so better than other classifiers.

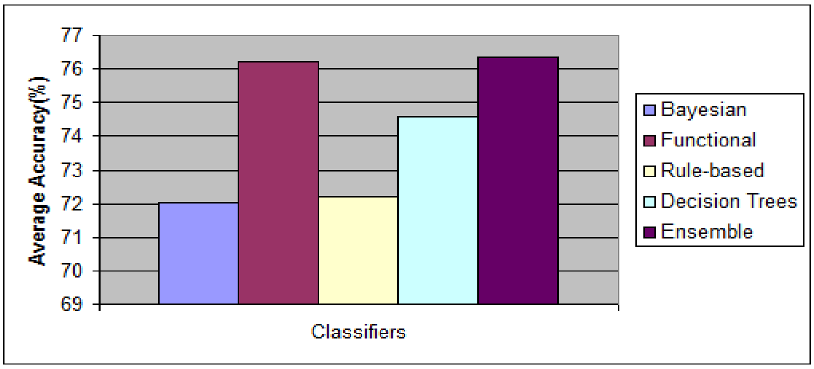


**Fig 2.** Showing average classification rate of different classifiers

Also, the best classification rates of different classifiers have been shown in Fig 3. As can be seen, the best accuracy for 10-F CV case is related to LMT and Bagging classifiers with 77.47%.

 

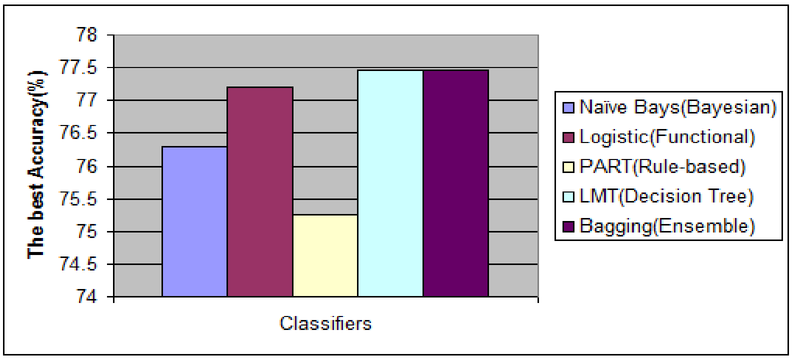


**Fig 3**. Showing the best accuracy of classifiers

Finally, the average time complexities of classifiers have been presented in Fig 4. The best and the worst average time complexities of classifiers are related to Bayesian and Decision Trees classifiers.

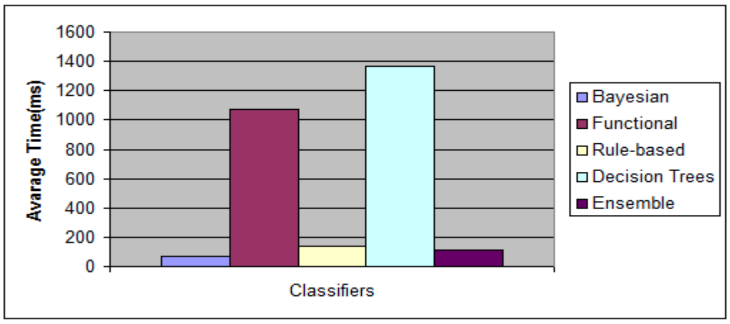


**Fig 4.** Presenting time complexity of different classifiers

Table 7 shows the results of evaluation the best WEKA classifier, Bagging with logistic core, with other presented methods from classification rate view point.

**Table 7.** Obtained classification rate with different classifiers for 10-F CV

| Method | Classification rate (%) |
|---|---|
| A. Suarez and F. Lutsko,1999[2] | 74.80 |
| C. Olaru and et al,2003[3] | 74.43 |
| J. Abonyi and et al,2003[4] | 73.05 |
| K.Kayaer and Yildirim,2003[5] | 77.08 |
| S. Sahan and et al,2005[18] | 75.87 |
| Y.C. Hu, 2007[1] | 74.81 |
| Tsipouras and et al,2008[8] | 75.91 |
| A. Sharif and et al,2008[7] | 74.63 |
| L. Peng and et al,2009[9] | 76.56 |
| **Bagging ensemble classifiers** | **77.47** |

## IV. CONCLUSION

In this paper, we have applied several intelligent classifiers such as Bayesian, Functional, Rule-base, Decision Trees and Ensemble for diabetes mellitus diagnosis on PID dataset. The obtained results show that, the best average F-measure and classification rate are connected to Functional and Ensemble classifiers; the best classification rate is related to LMT (Decision Trees) and Bagging (Ensemble) classifiers as well as the best average time complexity is liked to Bayesian classifiers. Therefore, we concluded that, Bagging with logistic core has the best performance. Also, in order to verify our conclusion, we have compared its classification rate with other presented methods.

## REFERENCES


1. Y.C. Hu, "Fuzzy integral-based perceptron for two-class pattern classification problems," *Information Sciences Elsevier*, vol. 177, no. 7, 2007, pp. 1673-1686.
2. A. Suarez and J.F. Lutsko, "Globally optimal fuzzy decision trees for classification and regression," *IEEE Transactions on Pattern Analysis and Machine Intelligence*,, vol. 21, no. 12, 1999, pp. 1297-1311.


 

3. C. Olaru and L. Wehenkel, "A complete fuzzy decision tree technique," *Fuzzy Sets and Systems*, vol. 138, no. 2, 2003, pp. 221-254.
4. J. Abonyi, et al., "Data-driven generation of compact, accurate, and linguistically sound fuzzy classifiers based on a decision-tree initialization," *International journal of approximate reasoning*, vol. 32, no. 1, 2003, pp. 1-21.
5. K. Kayaer and T. Yildirim, "Medical diagnosis on Pima Indian diabetes using general regression neural networks," *Proc. Proceedings of the international conference on artificial neural networks and neural information processing (ICANN/ICONIP)*, 2003, pp. 181-184.
6. S. ¦ahan, et al., "The medical applications of attribute weighted artificial immune system (AWAIS): diagnosis of heart and diabetes diseases," *Artificial Immune Systems*, 2005, pp. 456-468.
7. A. Sharifi, et al., "Hierarchical Takagi-Sugeno type fuzzy system for diabetes mellitus forecasting," *Proc. International Conference on Machine Learning and Cybernetics*, 2008, pp. 1265-1270.
8. M.G. Tsipouras, et al., "Automated creation of transparent fuzzy models based on decision trees-application to diabetes diagnosis," *Proc. 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2008, pp. 3799-3802.
9. L. Peng, et al., "Data gravitation based classification," *Information Sciences Elsevier*, vol. 179, no. 6, 2009, pp. 809-819.
10. E. Alpaydin, *Introduction to Machine Learning*, The MIT Press, 2010.
11. N.K. Bose and P. Liang, *Neural network fundamentals with graphs, algorithms, and applications*, McGraw-Hill, 1996.
12. K. Soman and V.A. Loganathan. R, *"Machine Learning with SVM and other Kernel Methods"*, Prentice Hall of India, 2009.
13. D.T. Larose and J. Wiley, *Data mining methods and models*, Wiley Online Library, 2006.
14. T. Dietterich, "Ensemble methods in machine learning," *Multiple classifier systems*, vol. 1857/2000, 2000, pp. 1-15.
15. C.L. Blake, & Merz, C. J., " UCI Repository of Machine Learning Databases," 1996; Available from http://www.ics.uci.edu./~mlearn/MLReporsitory.html.
16. D.M.W. Powers, "Evaluation: From precision, recall and f-factor to roc, informedness, markedness & correlation," *Journal of Machine Learning Technologies*, vol. 2, no. 1, 2011, pp. 37-63.
17. R.R. Bouckaert, et al., "WEKA---Experiences with a Java Open-Source Project," *The Journal of Machine Learning Research*, vol. 11, 2010, pp. 2533-2541.
18. S. Sahan, et al., "The medical applications of attribute weighted artificial immune system (AWAIS): diagnosis of heart and diabetes diseases," *Artificial Immune Systems*, 2005, pp. 456-468.

 