An Efficient Rule-based Classification of Diabetes Using ID3, C4.5 & CART Ensembles

[1]Saba Bashir, [2]Usman Qamar, [3]Farhan Hassan Khan, [4]M.Younus Javed
College of Electrical and Mechanical Engineering
National University of Sciences & Technology (NUST), Islamabad, Pakistan
{[1]saba.bashir, [2]usmanq, [3]farhan.hassan, [4]myjaved}@ceme.nust.edu.pk

Abstract – Conventional techniques for clinical decision support systems are based on a single classifier or simple combination of these classifiers used for disease diagnosis and prediction. Recently much attention has been paid on improving the performance of disease prediction by using ensemble-based methods. In this paper, we use multiple ensemble classification techniques for diabetes datasets. Three types of decision trees ID3, C4.5 and CART are used as the base classifiers. The ensemble techniques used are Majority Voting, Adaboost, Bayesian Boosting, Stacking and Bagging. Two benchmark diabetes datasets are used from UCI and BioStat repositories respectively. Experimental results and evaluation show that Bagging ensemble technique shows better performance as compared to single as well as other ensemble techniques.

Keywords – Diabetes, Bagging, Boosting, Adaboost, Bayesian boosting, Stacking, Ensemble Classifiers, Decision trees

I. INTRODUCTION

Computational intelligence is becoming more important in medical applications for some decades now. It improves the disease diagnosis accuracy and reduces its accompanied cost. Plenty of medical information is also available now in order to conduct research and perform experimentation in this area. Large number of datasets, belonging to 'life sciences', are also available online at UCI data repository, for researchers to experiment upon. Almost 28% of the total datasets are medical datasets. Therefore, there is an immense need to constantly improve the performance and disease predication accuracies [1].

Data mining plays a vital role for disease diagnosis and prediction in medical domain. There are various data mining algorithms that are available for deeper and complete understanding of medical data providing solutions to complex problems [2]. In healthcare, data mining can be used to provide analysis of medical centers to provide better resources, early detection and prevention of diseases; and cost savings from unwanted and expensive medical tests [3]. Several data mining techniques are used by different researchers for diagnosis and treatment of different diseases such as diabetes [4], stroke [5], heart disease [6] and cancer [7].

Diabetes is one of the most rapidly increasing diseases worldwide which occurs mostly due to obesity and lack of exercise. Insulin is most important hormone in human body and if it is not properly produced then large amount of sugar is driven out from body and results in all forms of diabetes [8]. The International Federation Association has researched that there is an alarming rise in diabetes patients by year 2030 [2]. Several ensemble and other classification techniques are proposed in the field of data mining for diabetes diagnosis. Ensemble classifiers are considered as more accurate in performance and prediction accuracies as compared to single classifiers. They provide more flexible structure and choose among different alternatives to provide best solution, high predictive performance and greater accuracy [9].

This research paper follows the taxonomy of data mining ensemble techniques and focuses on generating knowledge to make it useful for decision making. More specifically, we have used benchmark diabetes datasets for disease diagnosis and decision making. We have used three types of decision trees as base classifiers which differ on splitting criterion which are ID3, C4.5 and CART. These decision trees are then combined using different ensemble techniques. The proposed research focuses on five ensembles i.e. Majority Voting, Adaboost, Bayesian Boosting, Stacking and Bagging. The main goal in this study is to identify the best ensemble framework for decision trees that helps identifying the diabetes patients efficiently and most importantly, with high accuracy.

The rest of the paper is organized as follows: Section 2 presents related work. Section 3 is in relation to proposed methodology. Section 4 describes results, evaluation and discussion. Finally, conclusion and future work is given in section 5.

II. LITERATURE REVIEW

Medical data mining discovers hidden patterns from datasets efficiently and accurately. These patterns can then be utilized for disease diagnosis and treatment. Following research papers focus on using different data mining techniques for medical datasets.

[10] proposed a framework for diagnosis of diabetes in female patients. It is based on an ensemble of neural network and support vector machine. Pima Indian dataset is used from UCI data repository. Experimental results show good classification and prediction accuracy. It can be considered a good tool for identifying diabetes patients. [11] The proposed algorithm uses fuzzy systems and neural networks for learning membership functions. Steps involved in genetic algorithms are selection, crossover, mutation, and fitness and population calculation. It is concluded from

978-1-4799-7505-1/14 $31.00 © 2014 IEEE
DOI 10.1109/FIT.2014.50

results that new population of diabetes is based on old population and can be used to obtain high chromosomal accuracy. [4] presented a model for identifying diabetes patients from large medical datasets. This approach clusters diabetes patients in healthcare into different subpopulations. Data discretization and transformation techniques are applied to improve the quality of data. Experimental results show that the proposed technique can help decision makers to determine health policies and treatments for diabetes patients.

[12] proposed a hybrid model focused on combining agglomerative hierarchical clustering and decision trees. Pima Indian diabetes dataset is used to evaluate the performance of proposed model. Experimental results show that hybrid model achieved higher accuracy as compared to individual models. The combination of hierarchical clustering with decision trees is recommended for classification. [13] used genetic programming model for diabetes classification. New features are generated from existing diabetes features using genetic algorithm. A three stage model is proposed namely; feature selection, new feature generation and then features are evaluated using SVM and k-NN classifiers. Experimental results indicate that proposed model has high performance as compared to state of art techniques. [14] presented an amalgam model for identifying diabetes patients. The proposed model uses kNN classification technique along with k-means in the presence of several pre-processing steps. Experimental results show high accuracy with different k values.

[15] proposed a framework for diabetes diagnosing for Pima Indian diabetes dataset. It uses SVM classifier to predict the diabetic patients. Feature selection is performed using F-score and k-mean clustering methods to obtain optimal set of features. High accuracy of proposed technique recommended it for disease diagnosis. [16] created a soft switcher in Bayesian framework by combining elements from both averaging and switching techniques. Empirical datasets are used to evaluate the performance of the proposed technique.

An overview of the recent literature shows that extensive research has been conducted for diabetes diagnosis using ensembles. However, decision trees have not been explored adequately for the purpose. This research designed ensemble based methods to evaluate the performance of decision tree ensembles for diabetes datasets.

III. Proposed Methodology

The proposed methodology involves systematic implementation of various decision trees and ensemble techniques in order to diagnose diabetes. Three decision trees with varying splitting criterion i.e. information gain, gain ratio and gini index, are used as base classifiers. These individual classifiers are combined using Majority voting, Bagging, Adaboost, Bayesian Boosting and Stacking ensemble methods. Each ensemble technique generates different performance which is evaluated using parameters like accuracy, sensitivity, specificity, and f-measure. The experimentation is performed using RapidMiner5.

A. Base Classifiers

Decision trees provide white box structure for each provided dataset and can be combined with any other data mining techniques [17]. Each internal node of decision tree is test on an attribute, branches of tree represent possible outcomes from each test and leaf nodes represent output class labels. At start, all data is at root node and then rules are generated as it is traversed down from root to leaf nodes for classification and prediction. Each of the three techniques are describes as follows:

1) Decision tree based on Information Gain (ID3)

It works on the principle of greedy search which employs in top-down manner. Entropy is a measure that is used to divide the instances into subsets. It calculates the homogeneity for a given dataset. If it is completely homogeneous, entropy will be zero. Otherwise it is equally divided to have entropy value one. The information gain is based on decrease in entropy. An attribute which has maximum entropy will return highest information gain value. Therefore, a highest information gain attribute will be selected as the splitting attribute. A decision tree will then be constructed based on finding the most homogeneous branch [18]. Following formulas are used where probability of class i is denoted by p_i and m is the number of classes for target attribute.

$$\text{Entropy} = \sum_{i=1}^{m} -p_i \log_2 p_i$$

$$\text{Info\_ gain} = \text{entropy(parent)} - [\text{avg.entropy(children)}]$$

2) Decision tree based on Gain ratio(C4.5)

It is also commonly known as C4.5. It is a fraction between information gain and its splitting information. It is generally used to reduce the effect of biasness that may occur due to large number of values for a given attribute. A decision tree based on gain ratio outperforms information gain in terms of both accuracy measure and handling complex problems. The decision tree is constructed by considering number and size of branches for a given attribute. Gain ratio reduces the bias effect of information gain and allows uniformity and breadth of values for particular attribute. The attribute with highest gain ratio is selected as splitting attribute in order to construct a decision tree [19]. Following formula is used to calculate the gain ratio for attributes where 'Splitting info' is intrinsic information (column sums of frequency table) for particular attribute.

$$\text{Gain\_ratio} = \frac{\text{Information gain}}{\text{Splitting info}}$$

3) Decision tree based on Gini index (CART)

It is also known as CART (classification and regression tree). It measures the level of impurity for given data and constructs a binary tree where each internal node outputs exactly two classes for a given attribute. Gini index is calculated for each attribute and then the attribute with lowest gini index is selected as the splitting attribute. The tree is constructed by recursively selecting the attribute with lowest gini index [18]. If the probability of the i^{th} class is p_i for k target classes for a given attribute then Gini index is calculated as follows where p_i is probability of class i.

$$\text{Gini\_index} = 1 - \sum_{i=1}^{k} p_i^2$$

B. Ensemble Classifiers

A classifier is characterized as a model which is learned from training dataset. Data mining uses different supervised learning algorithms where each classifier performs predictions based on its learning. The basic idea behind ensemble classifiers is to weigh several individual classifiers and then combine them to obtain the result which outperforms every individual classifier. The classification performance and prediction accuracy of ensemble classifier is higher than single classifiers [20]. The proposed research focused on following ensemble classifiers for combining ID3, C4.5 and CART decision tree algorithms.

1) Majority Voting

Majority voting is used to classify the unlabeled instances based on the highest number of votes (high frequency vote). This technique is also termed as plurality voting (PV). The majority voting ensemble classifier is shown in Fig 1.

Three types of decision trees are used to train the classifiers and then each of them is applied to predict the class of unlabeled diabetes instances. The result of these decision tree classifiers is combined using majority voting scheme. Mathematically, it can be written as [21]:

$$\text{class}(x) = \arg\max_{ci \in \text{dom}(y)} \left(\sum_{k} g(y_k(x), c_i)\right)$$

where the classification of k^{th} classifier is represented by $y_k(x)$ and g(y,c) is an indicator function which is defined as

$$g(y,c) = \begin{cases} 1 & y = c \\ 0 & y \neq c \end{cases}$$

In case of probabilistic classifier, a crisp classification can be obtained by following formula

$$y_k(x) = \arg\max_{ci \in \text{dom}(y)} \hat{P}_{M_k}(y = c_i \mid x)$$

where a classifier is denoted by M_k and probability of class c for an instance x is represented by

$$\hat{P}_{M_k}(y = c_i \mid x)$$

2) AdaBoost

AdaBoost is a popular ensemble algorithm introduced by [22] and it performs boosting by iterative processing. It focuses on the instances that are difficult to classify using other classification techniques. The level of focus depends on the weight that is associated with instances during each iteration. At start, all instances are assigned equal weight. In each iteration, the weight of misclassified instances is increased whereas weight are reduced for the instances which are correctly classified. Moreover, each classified has an associated weight which measures the overall accuracy of that individual classifier. The classifiers are then combined considering their weight and prediction class.

Three decision trees were trained and then they are combined using AdaBoost ensemble to classify unlabeled diabetes instances. Boosting ensemble classifier is shown in Fig 2. Mathematically, the AdaBoost ensemble method can be written as [21]:

$$H(x) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t . M_t(x)\right)$$

where M_t denotes the classification based on voting of all classifiers for a particular instance and α_t is weight.

3) Stacking

Stacking is an ensemble technique which has achieved greater generalization accuracy. Stacking technique is based on single classifiers and it determines which classifier is reliable and which classifier is unreliable. The idea behind stacking approach is to construct a meta-dataset by taking some instances from original dataset. The predictions made by each individual classifier are used as input attributes instead of original dataset. The original training set contains target attribute without making any changes. Base classifiers first classify an unknown instance then these classifications made by the classifiers are combined into final prediction [21]. A stacking based ensemble classifier is shown in Fig 3. The three base classifiers are trained and then they are combined using stacking to classify unlabeled diabetes instances.

4) Bayesian Boosting

Bayesian boosting ensemble technique is used to make an ensemble classifier that eventually performs boolean classification for target attributes. It is an iterative process where weights are assigned and updated in every iteration and then sampling is carried out. The base classifier such as decision tree is applied multiple times sequentially, and then these classifiers are combined into a single coherent classifier. The parameter iteration defines the maximum number of trained classifiers [23]. A boosting ensemble framework is shown in Fig 2.

5) Bagging

Also termed as Bootstrap Aggregation. It is most common and well known ensemble method for data classification and prediction. An improved composite classifier is created in order to improve the classification accuracy as compared to individual classifiers. The outputs of individual classifiers are combined to generate a final

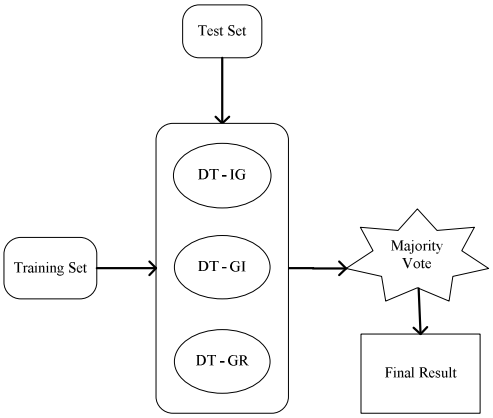

Fig 1. Majority Voting Ensemble Framework

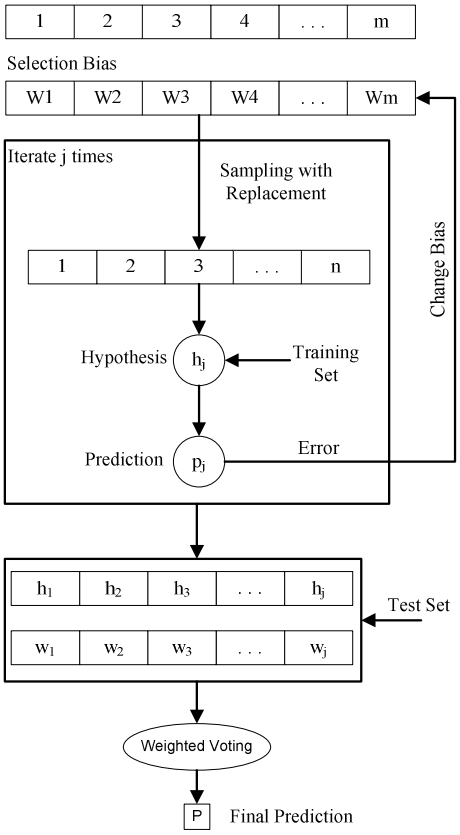

Fig 2. Boosting ensemble classifier framework

output class. All individual classifiers are trained from training set by taking sample of instances with replacement. The size of each sample and training set is equal. [24] proved that bagging can perform better from individual classifiers build from original training data because bagging can eliminate instability of single inducers.

A voting approach is used to combine the results of individual classifiers. Each classifier selects each instance with equal probability which is unlike boosting where the probability of selecting an instance depends on its weight. Bagging ensemble classifier is shown in Fig 4.

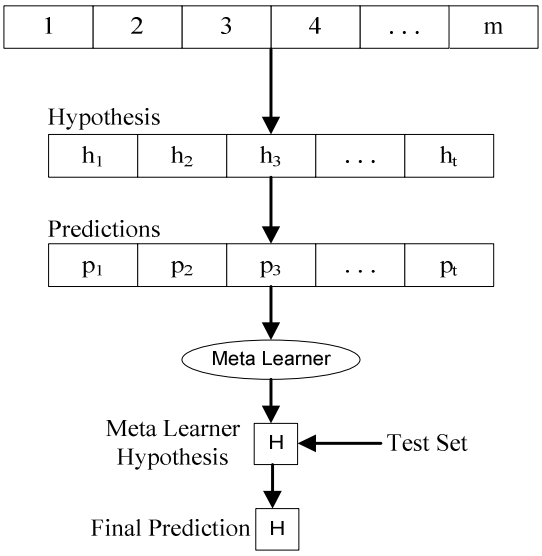

Fig 3. Stacking ensemble classifier framework

IV. Results, Evaluation and Discussion

The experimentation is conducted on two diabetes datasets which are easily available from BioStat and UCI online repositories. The results are presented in the form of confusion matrices in table 1 and 2. The 'TH' and 'TS' columns show the actual class labels as true healthy and true sick, whereas 'PH' and 'PS' rows present the number of individuals predicted by the ensemble as predicted healthy and predicted sick. The datasets are named as Pima Indian Diabetes Dataset (PIDD) and BioStat Diabetes Dataset (BDD). First column presents the ensemble schemes i.e Majority Voting (MV), AdaBoost (AB), Bayesian Boosting (BB), Stacking (S) and Bagging (B). The analysis of tables indicates that both datasets contain different characteristics therefore, different results are obtained.

Accuracy (Acc), sensitivity (Sen), specificity (Spec) and f-measure (F-M) are used for the performance evaluation of these ensemble techniques. The formulas of these performance measures are shown below

$$\text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}}$$

$$\text{F} - \text{Measure} = 2 * \frac{\text{Sensitivity} * \text{Specificity}}{\text{Sensitivity} + \text{Specificity}}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$$

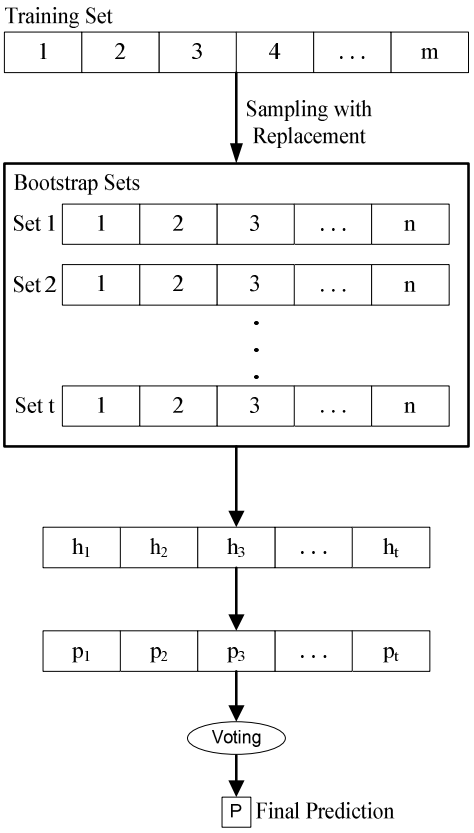

Fig 4. Bagging ensemble classifier framework

where TP, TN, FP, FN are true positives, true negatives, false positives and false negatives respectively.

RapidMiner5 is used for the model building, learning and testing. Stratified sampling is used to handle the class variance problem as there are more healthy persons and fewer sick individuals. Moreover, 10-fold cross validation is applied which uses 90% data as training set and 10% as testing set. The data is divided into ten mutually exclusive sets and each time nine sets are used for training and remaining one set is used for testing. This is repeated 10 times so that each time the training and test sets are different. The results are then averaged over the 10 iterations. Fig 5 and Fig 6 present the comparison of ensemble approaches. Table 3 shows accuracy comparison of different data mining techniques with published research.

The analysis of results clearly shows that Bagging approach is the winner for decision tree ensembles. It has the best performance results as compared to the other similar approaches. It has achieved the highest accuracy levels for both the diabetes datasets whereas sensitivity, specificity and f-measure values are also very competitive. The best bagging results achieved for accuracy, sensitivity, specificity and f-measure are 91.56%, 95.63%, 68.33% and 79.71% respectively.

TABLE I: Comparison of Ensembles for PIDD

PIDD		TH	TS	Acc	Sen	Spec	F-M
MV	PH	447	146	74.09%	89.40%	45.52%	60.33%
	PS	53	122				
AB	PH	421	119	74.22%	84.20%	55.60%	66.97%
	PS	79	149				
BB	PH	413	119	73.18%	82.60%	55.60%	66.46%
	PS	87	149				
S	PH	380	124	68.23%	76.00%	53.73%	62.95%
	PS	120	144				
B	PH	407	103	74.48%	81.40%	61.57%	70.11%
	PS	93	165				

TABLE II: Comparison of Ensembles for BDD

BDD		TH	TS	Acc	Sen	Spec	F-M
MV	PH	322	22	89.33%	93.88%	63.33%	75.64%
	PS	21	38				
AB	PH	328	32	88.34%	95.63%	46.67%	62.72%
	PS	15	28				
BB	PH	317	17	89.33%	92.42%	71.67%	80.73%
	PS	26	43				
S	PH	322	38	85.36%	93.88%	36.67%	52.74%
	PS	21	22				
B	PH	328	19	91.56%	95.63%	68.33%	79.71%
	PS	15	41				

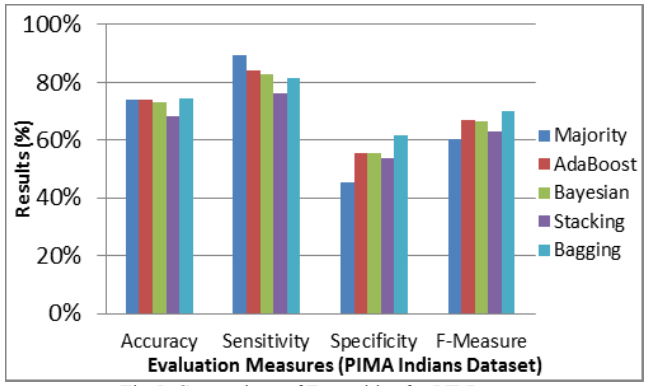

Fig 5. Comparison of Ensembles for PIDD

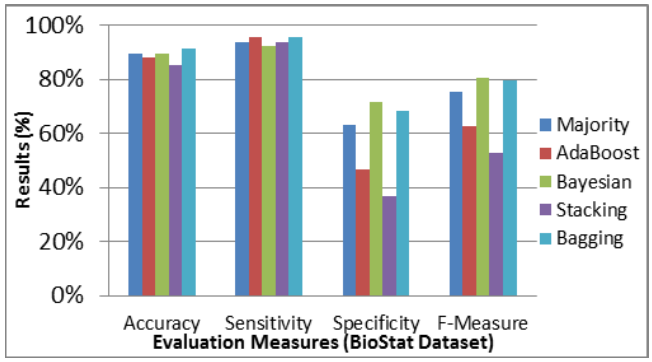

Fig 6. Comparison of Ensembles for BDD

Table III. Comparison of different data mining techniques

Reference	Year	Technique	Accuracy
[25]	2012	C4.5	91%
[26]	2012	J48	41.26%
		Naïve Bayes	47.61%
		CART	44.44%
		C4.5	68%
[27]	2014	C4.5	86%
		SVM	74.8%
		K-NN	78%
		PNN	67%
		BLR	75%
[28]	2012	Bagging	77.4%
		AdaBoost	77.2%
		Random forest	73.5%
		Multiclass	77.2%
[29]	2014	Boosting	81%
Proposed Technique		Majority voting	89.33%
		Adaboost	88.34%
		Bayesian	89.33%
		Stacking	85.36%
		Bagging	91.56%

V. Conclusion and Future Work

Decision trees are one of the most successful data mining techniques that are used for classification and prediction. The proposed research focuses on improving the disease diagnosis performance and accuracy for diabetes datasets using decision trees. This research systematically performs five ensemble classification techniques to improve individual tree performance. Experiments are conducted on Biostat and Pima Indian diabetes datasets. Evaluation of results indicates that Bagging ensemble outperforms other techniques for both the diabetes datasets. In future, similar ensemble techniques can be applied on other disease datasets such as breast cancer, heart disease and liver disease. Moreover, heterogeneous individual classifiers can be used as base classifiers such as Naïve Bayes, SVM and neural networks etc. Neural network and SVM classifiers can also be used to make effective ensemble models.

References

[1]. Karatsiolis, S., Schizas, C.N.: Region based Support Vector Machine Algorithm for Medical Diagnosis on Pima Indian Diabetes DataSet. In: Proceedings of the 2012 IEEE 12th International Conference on Bioinformatics & Bioengineering (BIBE), Cyprus,(2012)

[2]. Lauritzen, J.N., Arsand, E., Vuurden, K.V., Bellika, J.G., Hejlesen, O.K., Hartvigsen, G.: Towards a mobile solution for predicting illness in type 1 diabetes mellitus In: 2011 2nd International Conference on Wireless Communication, Vehicular Technology, (2011)

[3]. Ruben, D.C.J., Data Mining in Healthcare: Current Applications and Issues. (2009)

[4]. Porter, T., Green, B.: Identifying Diabetic Patients: A Data Mining Approach. Americas Conference on Information Systems, (2009)

[5]. Panzarasa, S., Quaglini, S. et al.: Data mining techniques for analyzing stroke care processes. In: Proceedings of the 13th World Congress on Medical Informatics (2010).

[6]. Das, R., Turkoglu, I., Sengur, A.: Effective diagnosis of heart disease through neural networks ensembles. Expert Systems with Applications, Elsevier, (2009).

[7]. Naik, J., Patel, S.: Tumor Detection and Classification using Decision Tree in Brain MRI. International journal of engineering development and research (2013)

[8]. Vijiyarani, S., Sudha, S.: Disease Prediction in Data Mining Technique – A Survey. International Journal of Computer Applications & Information Technology, (2013)

[9]. Rokach, L.: Ensemble-based classifiers. AI Review (2010)

[10]. Zolfaghari, R.: Diagnosis of Diabetes in Female Population of Pima Indian Heritage with Ensemble of BP Neural Network and SVM. IJCEM Vol. 15 (2012)

[11]. Sapna, S., Tamilarasi, a.: Data mining – Fuzzy Neural Genetic Algorithm in predicting diabetes.In: Research Journal on Computer Engineering, (2008).

[12]. Ibrahim, N.H., Mustapha, A., Rosli, R., Helmee, N.H.: A Hybrid Model of Hierarchical Clustering and Decision Tree for Rule-based Classification of Diabetic Patients. In: International Journal of Engineering and Technology (2013)

[13]. Aslam, M.W., Zhu, Z., Nandi, A.K.: Feature generation using genetic programming with comparative partner selection for diabetes classification.In: Expert Systems with Applications. Elsevier (2013)

[14]. NirmalaDevi, M., Appavu, S., Swathi, U.V.: An amalgam KNN to predict diabetes mellitus. In: International conference on Emerging Trends in Computing, Communication and Nanotechnology (ICE-CCN), (2013)

[15]. Gandhi, K.K., Prajapati, N.B.: Diabetes prediction using feature selection and classification. International Journal of Advance Engineering and Research Development (2014)

[16]. Stahl, F., Johansson, R., Renard, E.: Ensemble Glucose Prediction in Insulin-Dependent Diabetes. Data driven modeling for diabetes. Springer (2014)

[17]. Shouman, M., Turner, T., Stocker, R.: Using Decision Tree for Diagnosing Heart Disease Patients. Proceedings of the 9-th Australasian Data Mining Conference (AusDM'11), Australia (2011)

[18]. Bramer, M.: Principles of data mining, Springer.(2007)

[19]. Han, j. and M. Kamber.: Data Mining Concepts and Techniques, Morgan Kaufmann Publishers.(2006)

[20]. Polikar, R.: Ensemble based systems in decision making. IEEE Circuits Syst Mag. (2006)

[21]. Rokach, L.: Ensemble-based classifiers. AI Review 33:1–39 (2010)

[22]. Freund Y, Schapire RE.: Experiments with a new boosting algorithm. In: Machine learning: proceedings of the thirteenth international conference, (1996)

[23]. Data mining and Rapid Miner. http://www.slideshare.net/dataminingtools/rapidminer-data-mining-and-rapid-miner (Last Accessed: 25-6-2014)

[24]. Breiman L Bagging predictors. Mach Learn 24(2):123–140

[25]. Rajesh, K., Sangeetha, V.: Application of Data Mining Methods and Techniques for Diabetes Diagnosis.International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 3, September 2012

[26]. Ashwinkumar.U.M, Anandakumar.K.R: Predicting Early Detection of Cardiac and Diabetes Symptoms using Data Mining Techniques. 2012 2nd International Conference on Computer Design and Engineering (ICCDE 2012)

[27]. Radha, P., Srinivasan, B.: Predicting Diabetes by cosequencing the various Data Mining Classification Techniques IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 1 Issue 6, August 2014

[28]. Hosseinpour, N., Setayeshi, S., Ansari-asl, K., Mosleh, M.: Diabetes Diagnosis by Using Computational Intelligence Algorithms. International Journal of Advanced Research in Computer Science and Software Engineering 2 (12),December - 2012, pp. 71-77

[29]. Ali, R., Siddiqi, M.H.,Idris, M., Kang, B.H.: Prediction of Diabetes Mellitus Based on Boosting. Ensemble Modeling. 8th International conference on Ubiquitous computing & Ambient intelligence. (2014)