COMPARISON AND EVALUATION DATA MINING TECHNIQUES IN THE DIAGNOSIS OF HEART DISEASE

Serdar AYDIN, Meysam Ahanpanjeh and Sogol Mohabbatiyan

Social Sciences of Institute, Atatürk Üniversitesi Kampüsü, Erzurum, Turkey

ABSTRACT

Heart disease is one of the biggest health problems in the world because of high mortality and morbidity caused by the disease. The use of data mining on medical data brought valuable and effective life achievements and can enhance medical knowledge to make necessary decisions. Data mining plays an important role in the field of medical science to solve health problems and diagnose ailments in critical conditions and in normal conditions. For this reason, in this paper, data mining techniques are used to diagnose heart disease from a dataset that includes 200 samples from different patients. Techniques used to diagnose heart disease include Bagging, AdaBoostM1, Random Forest, Naive Bayes, RBF Network, IBK, and NNge that all the techniques used to diagnose heart disease use Weka tool. Then these techniques are compared to determine which is more accurate in the diagnosis of heart disease that according to the results, it was found that the RBF Network with the accuracy of 88.2% is the most accurate classification in the diagnosis of heart disease.

KEYWORDS

Heart Disease, Data Mining, Bagging, AdaBoostM1, Random Forest, Naive Bayes, RBF Network, IBK,NNge

1.INTRODUCTION

Heart disease is the most leading cause of death in the world. Indeed, the mortality of the disease in developing countries is higher than developed countries. Heart is the most important organ in the body that the least disorder in this organ causes disruption in the life of every human being, the heart with the unbelievable order always serves the rest of the body [1, 2]. Risk factors for heart disease include older age, male gender, family history of premature heart disease, high blood pressure (hypertension), Increase blood fat, particularly cholesterol, diabetes, smoking, obesity, sedentary lifestyle (not doing physical activity), impaired clotting Blood and etc. As well as factors such as chest pain, shortness of breath, palpitations and etc. are signs of heart disease [1, 2, 3].

In general, the diagnosis is a complex task that requires experience and high skills, however, early detection and medical care of heart patients can greatly prevent sudden death in these patients and reduce the high costs of surgery and other treatment courses [1, 2, 3]. Up to now, various solutions have been proposed to solve this problem. One of the most important solutions is data mining techniques. Data mining techniques play an important role for diagnosis of ailments and the purpose of using data mining is to enhance precision in medical diagnosis. Having extract of important information from a large mass of data and correlation of these data imply the advantages of the use of data mining [3, 4]. New knowledge about human tasks and changes

DOI:10.5121/ijcsa.2016.6101

vocabulary is constantly being discovered.There is a steady stream of new events in the world and theredesign of systems of artificial intelligence to match the new knowledge,and in practice this is a very hard workand to the extent it is impossible, butthe machine learning procedures, such as data mining may be able to trace manyof these new knowledge [4, 5]. As a result, the main goal of data mining is the discovery ofhidden knowledge of data that lies in the enormous banks ofinformation, andtoachieve to this immense knowledge should first make available an accrete environment of data that is called data analysis site, then search targetdata, then be transformation on them, then explore the discovery knowledge that called data mining with tools are used in the data mining and finally, in the last step of knowledge discovery, results bepresented to the user is quite understandable [2, 4, 5, 6].

The most important applications of data mining include commercial and financial affairs, medical affairs, medical environmental,analysis relating to DNA, exploredissonanceand spurious documents ,telecommunications, sports and entertainments, libraryandinformation update[5, 6, 7]. Also some of the techniques have been applied under thecommon techniques of data mining include inquiry tools, statistical techniques, illustrated, continuous analytical processing, case-based learning, decisiontree, dependency rules, neural networks, genetic algorithms, and etc. That is these techniques have been good solutions in the field of medicine to diagnose critical and normaldiseases. The application of these techniques is in discoveringand disclosing new information and relationships embedded in the large andcomplex data sets with new patterns of learning and inference relations [5,6,7].The data used in this paper has beencollecting from a dataset called Long Beach VA that contains 200 samples of data from heart patients that in this paper with the use of data mining techniques such asBagging, AdaBoostM1, Random Forest,Naïve Bayes ,RBFNetwork, IBK, and NNge to diagnoseheart disease with the use of these 200 sample data. And all of thesetechniques also diagnose heart disease with the Weka tools.

This paper is organized as follows:In the Section 2, we will review the previous work in the field of diagnosis of heart disease with data mining techniques; In the Section 3,we will discuss the Long Beach VA data collection and data mining techniques; In Section 4, we will discuss the results of data mining techniques used to diagnose heart disease, we then compare and evaluate the results of these techniques to determine which technique is more accurate for the diagnosis of heart disease; In the Section 5,we draw conclusions and discuss future work.

2.PREVIOUS WORK

Significant research in the field of diagnosis of heart disease is performed with data mining techniques will be discussed in this section.

Rajkumar and his colleagues [8] diagnosed heart disease with the use of data mining techniques such as Decision List, Naive Bayes, and K-Nearest Neighbors (KNN). They used a dataset with 3000 samples of data, each sample consist of 14 features of characteristics of heart disease. 70% of data is used for education and 30% for testing. Performance analysis of these techniques is based on accuracy and time taken to build the model. The accuracy and the time spent in Naive Bayes, respectively are 52.33% and 609.The accuracy and time spent in Decision List, respectively are 52% and 719, as well as the accuracy and time spent in KNN respectively are 45.67% and 1000,so accuracy for the time taken for Naive Bayes is higher than other techniques.
Also in another way, researchers [9] diagnosed coronary heart disease using data mining techniques based on a dataset, which includes 303 samples with 16 features of demographic (features of population/group statistics), 14 features ofpatient's physical examination, and 7

features of signals ECG (electrocardiogram signal). In this paperthey have used Support Vector Machine (SVM) using Sequential Minimal Optimization (SMO) techniques, Naive Bayes techniques, and a proposed technique that uses the SMO and Naive Bayes techniques in different ways.The indicators are Accuracy, Sensitivity, and Specificity. The rate of mentioned parameters of the proposed technique is far more than other techniques.

Probable predicting was performed the diabetic patients to heart disease using Naïve Bayes classification model [10]. Dataset includes 500 samples of diabetic patients that each sample contains 9 features, such as blood pressure, blood glucose, patient's body weight, family history, age, cholesterol and etc. Classes are defined by high cholesteroland low cholesterol in Dataset in form of YES (high cholesterol) and NO (lower cholesterol). Prediction accuracy was 74% inNaive Bayes method.

In Reference [11] diagnosis of coronary artery heart disease was carried out with the use of classification techniques such as SMO, Naïve Bayes, AdaBoost and C4.5 that is one of the techniques of the decision tree. Features used to diagnose arelaboratory and echo features and demographical features. SMO techniques allocatedmaximum amount ofAccuracy and Sensitivity than other techniques. Also Naïve Bayes in terms of Specificity was the best.

In another study [12] conducted a diagnosis of heart disease with KNN. Dataset used included 303 samples with 13 features of heart patients. In KNN, the K value was assigned from1 to 13, thehighest accuracy was 7that the equivalent of 97.4% has been obtained.

Researchers in [2] have used 40 samples of data to diagnose heart disease. Multilayer Perceptron (MLP) model was data mining technique that used for diagnosis. In this way to MLP, the output layer has two neurons that one neuron for heart patients and another neuron for healthy people were used. Accurately diagnose heart disease in testing level was 85%. Dataset heart disease included 1000 samples of data have been used to diagnosis of coronary heart disease [13]. Data mining methods used for Decision Trees, SVM, and MLP,and SVM has been the highest accuracy for diagnosis and it was equal to 92.1%.

3.A COLLECTION OF DATA USED AND DATA MINING TECHNIQUES

Data mining is a new subject with vast and diverse applications that is introduced as one of the best science of the world, which leads to achange in the age of technology and in all areas of application [6].

Today, with the development of database systems and high volume of data stored in these systems, the need for a tool to be able to process data stored and resulting from this process was in the hands of users.So, when the volume of data is high, although users are experienced, they may not determine the useful patterns in the high mass of data or if they are able to do this, both operations cost in terms of material and human resources is very high [14, 15].On the other hand users usually are raised a hypothesis, and then based on the observed reports they prove or disprove (reject) the hypothesis, whereas nowadays there is a need for procedures to discover

knowledge with minimum user intervention and automatically express patterns and logical relationships.Data mining is one of the most important methods by which useful patterns in data are known with a minimum intervention of users and provide information to users and analysts, and on the basis of these information critical decisions are adopted in organizations [14, 15, 16].

In data mining, a part of the science of statistics named data exploration analysis is used where the discovery of hidden and unknown information in the massive volume of data is emphasized. In addition, in data mining database theories, artificial intelligence, machine learning, and statistics are combinedto application is made. Thus, the more volume of data andthe more complex relations between them, the more difficult access to hidden information contained in the data, and the role of data mining as a method of discovery, is more clear [15, 16].With regard to the application of data mining techniques in medicine, it can be identified dangerous and chronic disease and the result of this work causethe best management and reduce the cost of therapy in the treatment process[17].That's why in this paperwith the use of data mining techniques such as Bagging, AdaBoostM1, Random Forest, Naive Bayes, RBF Network, IBK, and NNge we examine heart disease diagnosis. The data set used for data mining techniques have been collected from Long Beach VA Hospital [18], including 200 samples. Each sample includes 14 features. In this paper, we used to from data of Table (1)and it shows the 14 features.

No. of Feature	**Feature**	**Descriptions and Feature values**
1	Age	Numerical values
2	Sex	Male=1 Female=0
3	Chest pain type	Typical angina=1 Atypical angina=2 Non-angina pain=3 Asymptomatic=4
4	Resting blood pressure	Numerical values in mm hg
5	Serum cholesterol	Numerical values in mm/dl
6	Fasting blood sugar	Fasting blood sugar > 120 mg/dl (True=1; False=0)
7	Resting electrographic results	Normal=0 Having ST-T wave abnormality=1 Left ventricular hypertrophy=2
8	Maximum heart rate achieved	Numerical values
9	Exercise induced angina	Yes=1 No=0
10	ST depression induced by exercise relative to rest	Numerical values
11	Slope of the peak exercise ST segment	Up sloping=1 Flat=2 Down sloping=3
12	Number of	Value = 0-3

	major vessels colored by fluoroscopy	
13	Defect type	Normal=3 Fixed defect=6 Reversible defect=7
14	Diagnosis of heart disease	sick=1 Normal=0

Table 1: Long Beach VA features for the diagnosis of heart disease

According to Table 1, which shows the characteristics of the Long Beach VA data collection, this datasets contains 14 features for each sample. 13 features of Table 1 are the input, and the fourteenth feature is output. The value of the fourteenth feature is0 (zero) and 1.Zero means healthy(Normal) and 1 meansPatient (Sick).We specify the fourteenth feature astwo healthy classes (Normal) and Patient (Sick) and the purpose of the classification is based on these two classes.

3.1.Bagging

In this way, the frequent use of data that their distribution is uniform, the sampling is performed by permutation. Because sampling is done by permutation some cases may appear more than once in the same case study. After teaching classification, one class will be assigned with the highest number of votes. In other words, the approach in theBagging technique is a didacticdataset and it is the representative of the society under examination, and a variety of research states can be realized in the community of simulated datasets. So using re-sampling by using different data sets, required diversity would be attained, and when a new sample enters into each of the classifications, the majority agreed is used to diagnose the class[19, 20].

3.2.AdaBoostM1

The word AdaBoostM1 is derived from Adaptive Boosting, and it isa way for using the multiple learning algorithmsand for integration based on the output of all of them. In fact AdaBoostM1 is a meta-algorithm that is used to improve performance and solve the problem of unbalanced categories along with other learning algorithm. In this technique, the classification of each new stage is set for the benefit of wrong examples of classification in the previous steps. AdaBoostM1 is sensitive to noisy and scatter data, but it is superior than most of the learning algorithmsfor over-fitting problem. Base classification used here is just better than random classification and thus with more repetitions technical performance improves. This technique, even classification ofhigh random error with a negative index,improvesthe overall performance [21, 22].

3.3.Random Forest

Random Forest consists of decision trees. Every decision tree is formedby subset of training datawhich randomly selected. The decision tree is a method for displaying a series of laws that are leading to a category or value.The difference between the methods of decision tree is that how the distance to be measured. Decision trees that are used to predict the cluster variables called classification trees because theyare located the samples in clusters or classes. Every decision tree

in Random Forest provides results for classification and final results of Random Forest, is that most of the trees have announced. To build Random Forest, it can be preserved a number of decision trees that want to exist in the forests. One of the advantages ofRandom Forest is that it requires insignificantpreprocessing. Also there is no need to choose the required variables at the beginning and Random Forest model itself chooses the useful variables [23, 24, 25].

3.4.Naïve Bayes

This algorithm is derived from the Bayesian theory that is based on the probability of occurrence or non-occurrence of a phenomenon for classification.Based on the intrinsic properties of the possibility (especially Share probability), Naive Bayes will provide good results with received initial training. In Naive Bayes, the method of learning is a kind of learning byobserver and controller. In fact, this technique can be formed models for classification and prediction of several purposes. Also Naive Bayes is used to solve problems, recognition and classification of data among the different categories of data, and as soon as possible this technique can form the model that lead to issues such as classification with 2 or more classes. However, with design issues and assumptions about the Naive Bayes, thismethod is more suitable for classification of more issues in the real world [9, 11, 17].

3.5.RBF Network

Artificial neural networks are the most versatile and the most practical methods for modeling large and complex issues that include hundreds of variables. Artificial neural networks can be used to classificationissues that are outputs of a class, or regression issues that are outputs of a numeric value. Each artificial neural network includes an input layer and each node in this layer is equivalent to one of the prediction variables. Artificial neural networks are used for complex data analysis that don't simply do by other algorithms. RBF Network, which is one of the Artificial Neural Networks, is derived from Radial Basis Function. RBF artificial neural network has several sequential layers that are made up of three layers: input, hidden and output. Hidden layer neurons in Networks RBF have Gaussian non-linear function. Hidden layer neuronsare multidimensional units, and dimension of theseneurons is equaled by the number of RBF input. RBF training is conducted in both supervised and unsupervised way. This is a learning process, that is,at first with one of the clustering methods, hidden layer of Gaussian function parameters are set and then the link between the hidden layer and output layer is regulated with a learning algorithm and withsupervision on the standard propagation of error (Back propagation) or Levenberg-Marquardt method [17, 26, 27].

3.6.IBK

IBK is a learning technique in Lazy category.Lazy categories store training samples and until the time of classification do not do any real work. IBK uses KNN technique.Indeed, IBK is a classifier and a close neighbor with K classification and uses the noted distance criteria. The number of the closest distance to k = 1 pre assumption, can properly be defined in the object editor. Predictions of a pre-owned neighbor can be weighted based on their distance to the test sample[28, 29].

3.7.NNge

NNge is derived from Non Nested generalized exemplars. This method isused the nearest neighbor method to determine the level of each samples that is not covered by the entry techniques of decision table. Decision Table is based on categories of similar properties, it is evaluated the properties of sub categories with searchingthe first and the best and can be used the cross validation to evaluate [28, 29].

4.RESULTS AND DISCUSSION

In this section, we study the diagnosis of heart disease, according to results of data mining techniques such as Bagging, AdaBoostM1, Random Forest, Naive Bayes, RBF Network, IBK, and NNge that each one explained before, then we compare them to decide which is more accurate in the diagnosis of heart disease. As stated, in the paper 200 sample data of Long Beach VA is used and by using the Weka, 80% of data that is 160 data samples for training and 20% of data that is 40 data samples are chosen for testing.The results in this paper have been taken from test data that is 40 data samples. Since classification of data includes two classes, the normal class and sick class (14th feature in Table 1). For this reason we show the accuracy of these criteria: Precision, Recall, and F-Measure to diagnose heart disease by using 40 experimental data samples from 200 data samples fora normal class. Each of these criteria has been obtained fromEquation 1, 2 and 3 [30].

$$\text{Precision} = \frac{TP}{TP + FP} \tag{1}$$

$$\text{Re}call = \frac{TP}{TP + FN} \tag{2}$$

$$\text{F - Measure} = \frac{2 * \text{Precision} * \text{Recall}}{(\text{Precision} + \text{Recall})} \tag{3}$$

According to equation 1, 2 and 3, TP (True positives) is equivalent to the number of samples that correctly have been identified positive(True detected positives). As well as FP (False positives) is equivalent to the number of samples that wrongly have been identified positive (False detected positives), and FN (False negative) is equivalent to the number of samples that wrongly have been identified negative (False detected negative) [30]. According to this explanation, the results of any use of data mining techniques are based on Table (2). This table shows the achieved accuracy based on healthy class (Normal) of 14^{th} feature in Table (1).

Techniques	TP Rate	FP Rate	Precision	Recall	F-Measure
Bagging	0.111	0.161	0.167	0.111	0.133
AdaBoostM1	0.222	0.194	0.25	0.222	0.235
Random Forest	0.333	0.161	0.375	0.333	0.353
Naive Bayes	0.444	0.258	0.333	0.444	0.381
RBF Network	0.333	0.032	0.75	0.333	0.462
IBK	0.444	0.194	0.4	0.444	0.421
NNge	0.333	0.226	0.3	0.333	0.316

Table 2: the results of data mining techniques based on Normal class

According to the results of Table 2, Figures 1, 2 and 3, respectively, shows the comparison of the accuracy of these criteria: Precision, Recall and F-Measure.

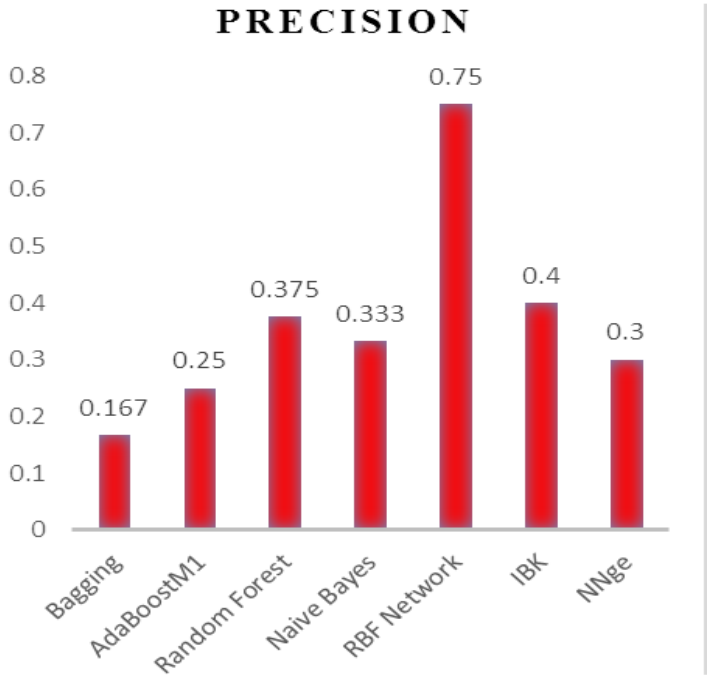

Figure 1:comparison of the data mining techniques based on the Precision criterion for diagnosis of heart disease

According to Figure 1, the highest Precision accuracy in Normal class is 0.75 that belong to RBF Network technique.

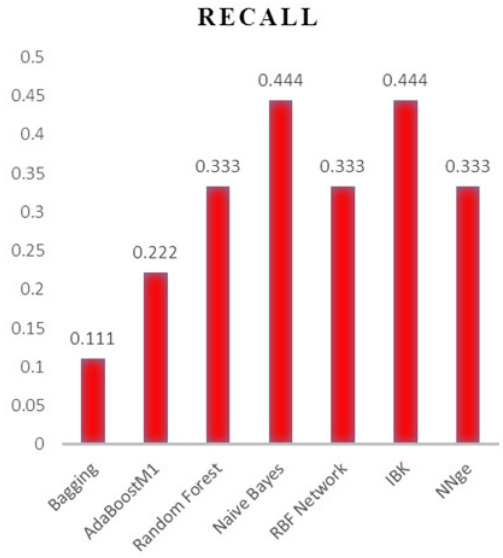

Figure 2:comparison of the data mining techniques based on the Recall criterion for diagnosis of heart disease

According to Figure 2, the highest Recall accuracy in Normal class is 0.444 that Naive Bayes and IBK techniques are equal to this amount (0.444). The Random Forest, RBF Network, and NNge have much in common with the amount to be 0.333.

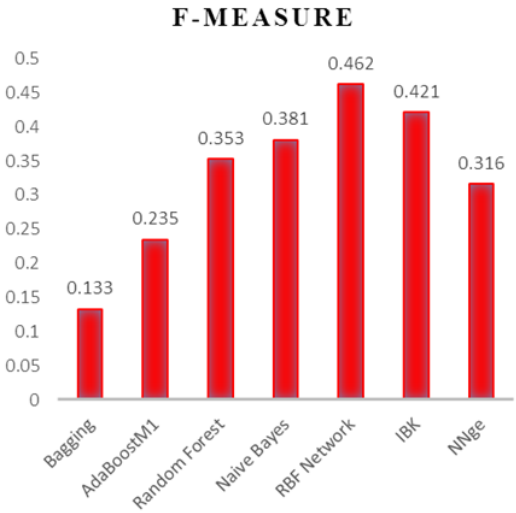

Figure 3:comparison of the data mining techniques based on the F-Measure criterion for diagnosis of heart disease

According to Figure 3, the highest F-Measure accuracy in Normal class is 0.462 that belong to RBF Network technique and after RBF Network, IBK technique that is equal to 0.421 and is close to the RBF Network, has the most value compared to other techniques.

Each data mining techniques has a level of error in the diagnosis of heart disease and by using four criteria MAE, RMSE, RAE, and RRSE in the Table (3) any technical errors in the diagnosis of heart disease is shown

Techniques	Mean Absolute Error (MAE)	Root Mean Squared Error (RMSE)	Relative Absolute Error (RAE)	Root Relative Squared Error (RRSE)
Bagging	0.3789	0.4564	%94.0283	%99.0412
AdaBoostM1	0.3683	0.4606	%91.1297	%95.0478
Random Forest	0.32	0.4561	%87.8644	%98.9706
Naive Bayes	0.3515	0.4582	%96.5071	%99.6432
RBF Network	0.293	0.3782	%804437	%90.3551
IBK	0.2778	0.5212	%76.2712	%99.5309
NNge	0.325	0.5701	%89.2327	%95.2133

Table 3: the results of errors obtained from data mining techniques to diagnosis of heart disease

According to the results of Table 3, Figures 4, 5, 6 and 7, respectively, shows the comparisonof the error of these criteria MAE, RMSE, RAE, and RRSE for diagnosis of heart disease,to decide which technique of these criteria has the lowest error.

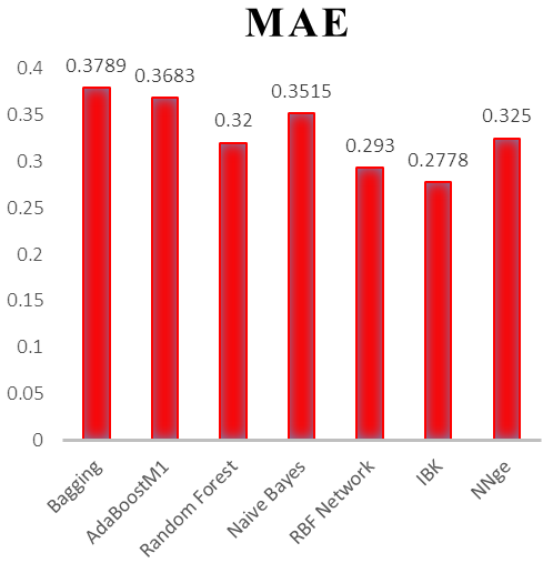

Figure 4: comparison of the data mining techniques based on the MAE criterion for diagnosis of heart disease

According to Figure 4, IBK has the lowest error in MAE criterion compared to other techniques.Also, RBF Network that itslevel of error is close to the IBK, andafter IBK has less error than other techniques in diagnosis of heart disease.

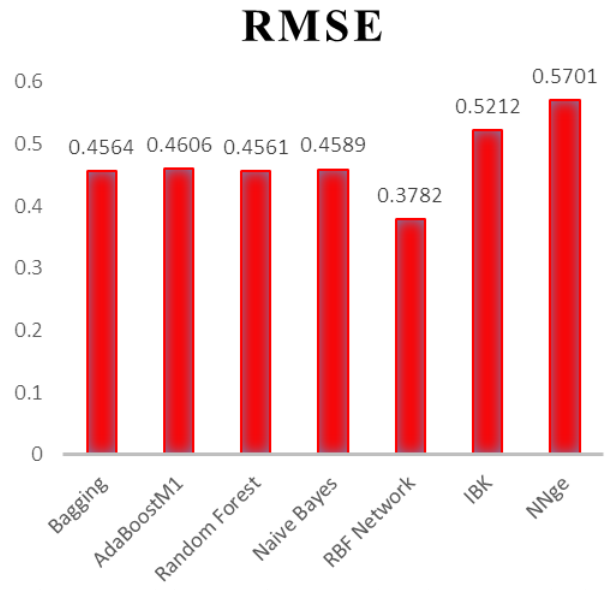

Figure 5: comparison of the data mining techniques based on the RMSE criterion for diagnosis of heart disease.

According to Figure 5, RBF Network has the lowest error in RMSE criterion compared to other techniques and the level of error in Bagging, AdaBoostM1, Random Forest, and Naive Bayes techniques is almost close together.

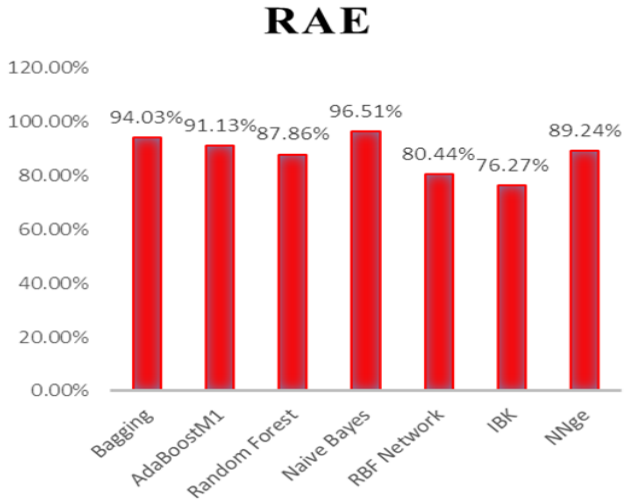

Figure 6: comparison of the data mining techniques based on the RAE criterion for diagnosis of heart disease.

According to Figure 6, IBK has the lowest error in RAE criterion compared to other techniques. Also, RBF Network after IBK has less error than other techniques in diagnosis of heart disease.

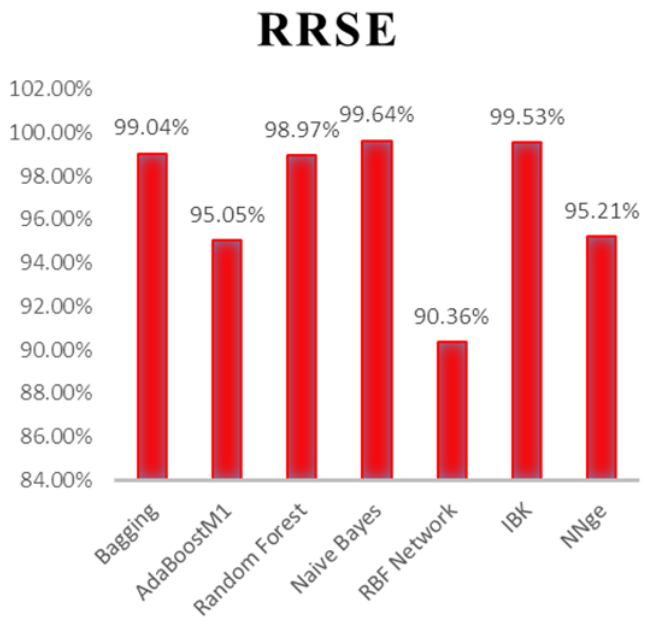

Figure 7: comparison of the data mining techniques based on the RRSE criterion for diagnosis of heart disease.

According to Figure 7, RBF Network has the lowest error in RRSE criterion compared to other techniques. And after RBF Network, AdaBoostM1 and NNge techniques that their level of error is closer together, have had the lowest error to diagnosis of heart disease.

As noted, 20% of data that is 40 data samples have been tested and data classification accuracy rate for diagnosis of heart disease is performed by testing data according to equation (4).

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{4}$$

According to Equation 4, each of the factors, TP, FP and FN in Eq. 1, 2 and 3 are explained and TN (True negatives) is equivalent to the number of samples that correctly have been identified negative (truly detected negative) [30]. According to this description, Table 4 shows data classification accuracy rate.

Techniques	**Correctly Classified Instances (%)**	**Incorrectly Classified Instances (%)**
Bagging	67.5 (27 data)	32.5 (13 data)
AdaBoostM1	67.5 (27 data)	32.5 (13 data)
Random	72.5 (29 data)	27.5 (11 data)

Forest		
Naive Bayes	77.5 (31 data)	22.5 (9 data)
RBF Network	82.5 (33 data)	17.5 (7 data)
IBK	72.5 (29 data)	27.5 (11 data)
NNge	70 (28 data)	30 (12 data)

Table 4: data classification accuracy of heart disease by evaluating experimental data set using with 40 samples

According to Table 4, the comparison of data classification accuracy for diagnosis of heart disease is shown in Figure 8.

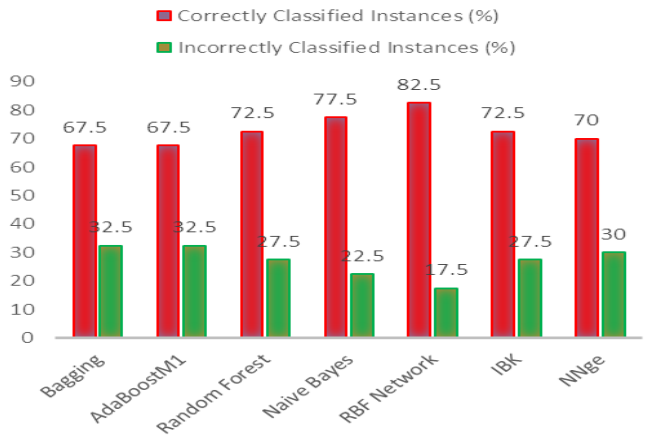

Figure 8:the comparison of data classification accuracy by testing 40 samples

As the results in Table 4 and Figure 8 are clear, RBF Network technique has the most accuracy for the diagnosis of heart disease and the value of this technique is equal to 82.5%, This means that RBF Network has been able to classify correctly the 33 data samples withthe evaluation of test data set from 40 data samplesand has not been able to classify correctly the 7 samples from the 40 data samples. Alsothe classification accuracy of Bagging and AdaBoostM1 techniques was equal to 67.5%, which means that both techniques were evaluated on a test sample set of 40 data samples, and 27 samples were correctly classified. According to Figure 8, Bagging and AdaBoostM1 techniqueshave the lowest correct classification accuracy in the diagnosis of heart disease.Also the accuracy of Random Forest and IBK was equal, it means that both techniques with the classification accuracy of 72.5% were able to correctly classify 29 data samples. Classification accuracy of NNge was 70% because NNge have correctly classified 28 data samples. Also Naive Bayes after RBF Network has the most (maximum) classification accuracy among other techniques, andits value was 77.5%, which means that Naive Bayeshave correctly classified 31 samples.

5.CONCLUSIONS AND FUTURE WORKS

Heart disease is the leading cause of death of many people in the world today and data mining have had an important role in the diagnosis of heart disease and other diseases. For this reason, in

this paper, data mining techniques such as Bagging, AdaBoostM1, Random Forest, Naive Bayes, RBF Network, IBK, and NNge were used to diagnosisof heart disease in order thatthe role of data mining in diagnosis of this disease have been shown. For diagnosis of heart disease a data set of 200 samples called Long Beach VA was used. Of these 200 samples, 20% of samples ,that is 40 data samples, were selected for the test, then each of these techniques by testingthese 40 data samples were compared with each other. It became clear that RBF Network in diagnosis of heart disease has had the most classification accuracy and it was equal to 88.2%. Also the RBF Network in two criteria of Precision and F-Measure has had the highest accuracy and lowest error in RMSE and RRSE compared to other data mining techniques. Things that can be done in the future are another critical illness used in data miningin order that the role of data mining in medical science has been developed.

REFERENCE

[1] B. Zebardast, A. Ghaffari, M. Masdari, "A New Generalized Regression Artificial Neural Networks Approach for Diagnosing Heart Disease", International Journal of Innovation and Applied Studies, Vol. 4, No. 4, pp. 679-689, 2013.

[2] 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(ICCNIT 2011), Abbottabad, Pakistan, pp. 245-249, 2011.

[3] 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, pp.73-78, 2014.

[4] Z.A. KHALIFELU, F.S. GHAREHCHOPOGH, "A Survey of Data Mining Techniques in Software Cost Estimation",AWERProcedia Information Technology & Computer Science Journal, Vol: 1, pp. 331-342, 2012.

[5] Z.A. KHALIFELU, F.S. GHAREHCHOPOGH, "Comparison and Evaluation Data Mining Techniques with Algorithmic Models in Software Cost Estimation", Elsevier Press, Procedia-Technology Journal, ISSN: 2212-0173, Vol: 1, pp. 65-71, 2012.

[6] H. Jiawei, K. Micheline, "Data Mining: Concepts and Techniques", vol. 2, Morgan Kaufmann Publishers, 2006

[7] R. S. Michalski, I. Bratko, M. Kubat, "Machine Learning and Data Mining: Methods and Applications", Wiley, New York, 1998.

[8] A. Rajkumar, G.S. Reena, "Diagnosis Of Heart Disease Using Data mining Algorithm", Global Journal of Computer Science and Technology" Vol. 10, Issue 10, Ver. 1.0, pp. 38-43, September 2010.

[9] R. Alizadehsani, J. Habibi, M.J. Hosseini, R. Boghrati, A. Ghandeharioun, B. Bahadorian, Z. AlizadehSani, "Diagnosis of Coronary Artery Disease Using Data Mining Techniques Based on Symptoms and ECG Features", European Journal of Scientific Research, Vol.82, No.4, pp.542-553, 2012.

[10] G. Parthiban, A. Rajesh, S.K.Srivatsa, "Diagnosis of Heart Disease for Diabetic Patients using Naive Bayes Method", International Journal of Computer Applications,Vol. 24, No.3, pp. 7–11, June 2011.

[11] R. Alizadehsani, J. Habibi, Z. Alizadeh Sani, H. Mashayekhi, R. Boghrati, A. Ghandeharioun, B. Bahadorian, "Diagnosis of Coronary Artery Disease Using Data mining based on Lab Data and Echo Features", Journal of Medical and Bioengineering, 2012.

[12] M. Shouman, T. Turner, R. Stocker, "Applying k-Nearest Neighbour in Diagnosing Heart Disease Patients", International Journal of Information and Education Technology, Vol. 2, No. 3, pp. 220-223, June 2012.

[13] Y. Xing, J. Wang, Z. Zhao, Y. Gao, "Combination Data Mining Methods with New Medical Data to Predicting Outcome of Coronary Heart Disease", International Conference on Convergence Information Technology(ICCIT 2007), IEEE, Gyeongju, South Korea, pp.21-23, Nov. 2007.

14] F.S. GHAREHCHOPOGH, ”Approach and Review of User Oriented Interactive Data Mining”, 4th International Conference on Application of Information and Communication Technologies (AICT2010), Digital Object Identifier: 10.1109/ICAICT.2010.5611792, IEEE, Tashkent, Uzbekistan, pp.1-4, 12-14 October 2010.

[15] Q. Luo, “Advancing Knowledge Discovery and Data Mining”, 1st International Workshop on Knowledge Discovery and Data Mining (WKDD'08), Adelaide, South Australia, pp. 3-5, 2008.

[16] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, “From Data Mining to Knowledge Discovery in Databases”, AI Magazine, pp. 37-54, 1996.

[17] 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, October 2015.

[18] V.A. Medical Center, Long Beach Clinic Foundation, “Available: https://archive.ics.uci.edu/ml/datasets/Heart+Disease, [Last Accessed 5 November 2015].

[19] M. Galar, A. Fernández, E. Barrenechea, H. Bustince, F. Herrera, “A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches”, IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART C: APPLICATIONS AND REVIEWS, Vol.42, No. 4, pp. 463-484, 2011.

[20] M. C. Tu, D. Shin, D. Shin, “Effective Diagnosis of Heart Disease through Bagging Approach”, IEEE, 2nd International Conference on Biomedical Engineering and Informatics, Tianjin, China, pp.1-4, 2009.

[21] E. AhmedSharaf , M. A. Moustafa, M. Harb, A. Emara, “ADABOOST ENSEMBLE WITH SIMPLE GENETIC ALGORITHM FOR STUDENT PREDICTION MODEL,” International Journal of Computer Science & Information Technology (IJCSIT), Vol. 5(2), pp. 73-85, 2013.

[22] M. Billah, “Symptom Analysis of Parkinson Disease using SVM-SMO and Ada-Boost Classifiers”, BRAC University, Dhaka, Bangladesh, page 46, 2014.

[23] A. S. Abdullah, R. R. Rajalaxmi, “A Data mining Model for predicting the Coronary Heart Disease using Random Forest Classifier”, International Conference on Recent Trends in Computational Methods, Communication and Controls (ICON3C 2012), ICON3C(3), pp.22-25, April 2012.

[24] K. Kalaiselvi, K. Sangeetha, S. Mogana, “Efficient Disease Classifier Using Data Mining Techniques: Refinement of Random Forest Termination Criteria”, IOSR Journal of Computer Engineering (IOSR-JCE), Vol. 14, No. 5, pp.104-111, 2013.

[25] E. E. Tripoliti, D.I. Fotiadis, G. Manis, “Automated Diagnosis of Diseases Based on Classification: Dynamic Determination of the Number of Trees in Random Forests Algorithm” IEEE TRANSACTIONS ON INFORMATION TECHNOLOGY IN BIOMEDICINE, VOL. 16, NO. 4, JULY 2012.

[26] 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.

[27] B. Zebardast, I. Maleki, “A New Radial Basis Function Artificial Neural Network based Recognition for Kurdish Manuscript”, International Journal of Applied Evolutionary Computation, Vol. 4(4), pp.72-87, 2013.

[28] D.S.V.G.K. Kaladhar, B. K. Pottumuthu, P. V. N. Rao, V. Vadlamudi, A. K. Chaitanya, R. H. Reddy, “The Elements of Statistical Learning in Colon Cancer Datasets: Data Mining, Inference and Prediction”, Algorithms Research, Vol. 2, No.1, pp.8-17, 2013.

[29] G. K. M. Nookala, B. K. Pottumuthu, N. Orsu, S. B. Mudunuri, “Performance Analysis and Evaluation of Different Data Mining Algorithms used for Cancer Classification”, International Journal of Advanced Research in Artificial Intelligence , Vol. 2, No.5, pp. 49-55, 2013.

[30] M. Farahmandian, Y. Lotfi, I. Maleki, “Data Mining Algorithms Application in Diabetes Diseases Diagnosis: A Case Study”, MAGNT Research Report, Vol.3, No. 1, pp. 989-997, 2015.