Intelligent Predictive System Using Classification Techniques for Heart Disease Diagnosis

Basheer Mohamad Al-Maqaleh,
Faculty of Computer Science and Information Systems
Thamar University, Thamar, Republic of Yemen
Email: basheer.almaqaleh.dm@gmail.com

Ahmed Mohamad Gasem Abdullah
Faculty of Computer Science and Information Systems
Thamar University, Thamar, Republic of Yemen
Email: gasemahmed25@gmail.com

Abstract— Heart disease continues to claim an alarming number of lives across the globe. The healthcare industry collects huge amounts of healthcare data, which are not mined to discover valuable information for efficient decision-making. The healthcare sector is still information rich but knowledge poo r. However, there is a lack of efficient analysis tools to discover hidden patterns and trends in medical dataset. In the healthcare industry, the data mining techniques are mainly used for classifying and predicting the diseases from medical datasets. In this paper, an intelligent predictive system using classification techniques for heart disease diagnosis, namely, J48 decision tree, Naïve Bayes and Multi-Layer Perceptron Neural Network (MLPNN) are proposed. The main objective of this paper is to study these classification techniques to predict the heart disease and find the best technique of prediction. The obtained results are evaluated by the common performance metrics like Accuracy, TP-rate, Precision, F-Measure and ROC graph.

Keywords- Data Mining; Diagnosis; Heart Disease; Decision Support System; Classification Techniques; Knowledge Discovery.

I. Introduction

The heart disease is the leading cause of death worldwide and the diagnosis is a complicated task that requires both experience and knowledge [1]. Medical datasets contain a wealth of hidden information that can be important in making decision [2]. Data mining is a knowledge discovery process to analyze data and extract useful information for effective decision-making [3, 4]. Data mining can be a useful tool in the health sector and healthcare [5]. In this field, predicting the outcome of a disease is one of the most interesting and challenging tasks in which to develop data mining applications [6, 7].

Classification and prediction are principle goals of data mining [3,8]. Classification in data mining is the process of finding a set of models that describe and distinguish data classes or concepts. Prediction in data mining involves attributes in the dataset to find an unknown class of other attributes. The purpose of prediction in data mining is to discover trends in patient data in order to improve their health. Heart disease prediction using classification data mining techniques is one of the most important and challenging task [6].

In this paper, an intelligent predictive system for diagnosis of heart disease using classification techniques is proposed. In the proposed system, three classifiers like J48 decision tree, Naïve Bayes and Multilayer Perceptron Neural Network (MLPNN) are used to predict the diagnosis of heart disease. The diagnosis of disease is a vital and crucial task in medicine [8]. Efficient and accurate implementation of the proposed system needs a comparative study of various classifiers used and to find the best classifier of prediction.

The remainder of this paper is organized as follows. Section II presents some related work on heart disease prediction system. In Section III, the proposed intelligent predictive system is presented. Section IV describes the features of the dataset used in this work. Computational results and comparative study are presented in Section V and Section VI respectively. Conclusion and future directions are given in Section VII.

II. RELATED WORK

Several studies related to heart disease diagnosis using data mining techniques are proposed in [9, 10, 11, 12, 13, 14]. These studies have applied different techniques to the given problem and achieved different probabilities for different methods.

A decision support in heart disease prediction system is developed in [4]. This system extracts hidden knowledge from a historical heart disease database using both Naive Bayes classifier and Jelinek-mercer smoothing technique. Jelinek-mercer smoothing technique is the more effective than Naive Bayes to predict patients with heart disease. An Intelligent Heart Disease Prediction System (IHDPS) using data mining techniques, namely, decision trees, Naive Bayes, and neural network in proposed in [15]. Each technique has own strength to get appropriate results. In [16] they applied and compared data mining techniques to predict the rise of heart disease using five different algorithms such as C5.0, Naive Bayes, Support Vector Machine (SVM), MLPNN and Logistic Regression with accuracy measures as: 93.02%, 86.05%, 88.37%, 85.22% and 80.23% using 13 medical attributes. An effective data mining association rules for heart disease prediction system is introduced in [17]. The principle of this study is, hence to extract hidden patterns by applying data mining techniques, noteworthy to heart diseases, from a data collected together by a hospital.

Heart disease using data mining algorithm on neural network and genetic algorithm is proposed in [18]. In their approach, the neural network is trained with selected patterns for the diagnosis to heart disease and genetic algorithm has been applied for optimizing the neural network. Comparative study of predicting heart disease by means of data mining is presented in [19]. This comparative study analyzes the existing prediction systems and discussing the various disputes on the existing systems. Prediction of heart disease using classification algorithms is presented in [20]. Classification techniques such as, J48, Naïve Bayes, REPTREE, CART and Bayes Net are used for predicting heart attacks. The research result shows prediction accuracy of 99%. A technique for diagnosis of heart disease using heart rate dataset is suggested in [11]. In this technique neural network and swarm intelligence optimization are used.

In the current work, three classification techniques: J48 decision tree, Naïve Bayes and MLPNN are used to predict the heart disease using WEKA 3.6.11 tool and the best technique of prediction is found.

III. PROPOSED SYSTEM

The main objective of the proposed system is to design an intelligent predictive system using classification techniques, namely, J48 decision tree, Naive Bayes and MLPNN for heart disease diagnosis. The proposed system identifies the features from the medical data using the classifier models. The attributes that one more relevant to heart disease diagnosis can be observed. Also, the proposed system predicts possible heart attacks from the patient dataset using these classification techniques and determines which model gives the highest percentage of correct predictions for the diagnosis.

The data collected from the Ibb Hospital, Ibb city, Yemen are used. The proposed system converts the unused data into a dataset for modeling using different data mining methods. In this stage, after consulting with the domain expert, a few preprocessing steps were implemented on the dataset to make the data more suitable for the mining step. In the next step, appropriate classification techniques for developing a predictive model are selected. These techniques are J48 decision tree, Naive Bayes and MLPNN.

Decision tree is a popular classifier that is simple and easy to implement. The results obtained from decision tree are easier to read and interpret. It can deal with information like nominal, numeric and text. Pruning is the technique in machine learning that reduces the size of decision trees by removing branches of the tree that provide little power to classify instances [3]. J48 decision tree model is the implementation of C4.5 decision tree in WEKA data mining tool. It supports continuous and discrete feature. It can also manage features with missing values [7].

Naïve Bayes is a statistical classifier, which assumes no dependency between attributes. It attempts to maximize the posterior probability in determing the class. Naïve Bayes classifiers are regularly gained from information. It provides new ways of exploring and understanding data [3,7].

An Artificial Neural Network (ANN) is an information processing system that has certain performance characters in common with biological neural networks. The system includes a large number of tiny processors to handle data processing. MLPNN is a finite acyclic graph. The nodes are neurons with logistic activation. It is one of the most commonly used ANN classifier algorithms [7,21]. In addition, WEKA package software has been used for prediction due to its simplicity, proficiency in discovering, analysis and predicting patterns [6, 26].

IV. DESCRIPTION OF THE DATASET

The Transthoracic Echocardiography report dataset is compiled from data collected from Ibb hospital, Ibb, Yemen. A total of 542 records with 11 medical attributes (factors) were obtained from this heart disease database. Table 1 lists the attributes. The attribute “Class” was identified as the predictable attribute with Class 1 "Normal" for patients with no heart disease and Class 2 "Abnormal" for patients with heart disease, for patients may confirm diagnosis of heart failure and for patients may be at risk of life threatening irregular heartbeats. The attribute “Patient ID” was used as the key; the rest are input attributes. It is assumed that problems such as missing data, inconsistent data, and duplicate data have all been resolved.

Predictable attribute (Class)

1. Normal**:** If Eject >= 53 and <= 75%
   Class1 = (Normal: no heart disease), else
2. Abnormal: If Eject < 53 %
   Class2 = (Abnormality: patient has heart disease, may confirm diagnosis of heart failure if Eject>= 40% and < 53%, may be at risk of life threatening irregular heartbeats if Eject < 40).

TABLE 1: HEART DISEASE DATASET ATTRIBUTE DESCRIPTION

No.	Attribute	Description	Values	Range
1	Age	Age in Year	Numeric	>=0
2	Sex	Female or Male	Nominal	F= Female M= male
3	Atrium	Left Atrium	Numeric	18-40 mm
4	Aorta	Aorta Artery	Numeric	20-37 mm
5	L-vent	Left Ventricle	Numeric	38-56 mm
6	L-ventf	Left Ventricle Function	Numeric	22-40 mm
7	R-ven	Right Ventricle	Numeric	< 25 mm
8	Sept	Septum	Numeric	6-11mm
9	L-vpo	Left Ventricle Posterior Wall	Numeric	16-25mm
10	Eject	Ejection Fraction	Numeric	53-75%

V. COMPUTATIONAL RESULTS

These data mining classification techniques were implemented using WEKA 3.6.11 tool. WEKA is a data mining tool which was developed in New Zeland by the University of Waikato that's implements data mining algorithms using JAVA language. WEKA is collection of machine learning algorithms and their application to the data mining problems [26]. Initially dataset had 11 medical attributes and 542 records. Preprocessing steps were applied on dataset to preprocess the dataset. Consequently, the three classifiers such as J48 decision tree, Naïve Bayes and MLPNN were used for diagnosis of patients with heart disease. Also, these classifiers were fed with reduced dataset with 3 attributes from a total of 11 attributes that were available. These 3 attributes that are highly relevant in predicting heart disease from Transthoracic Echocardiography dataset were selected. All the experiments were done on a full training dataset and 10-fold cross validation was used for randomly sampling the training and test sets.

The performances of the three classifiers were evaluated using the standard metrics of Accuracy, TP-rate, Precision and F-Measure which were calculated using confusion matrix [3,7,13]. The confusion matrix is a useful tool for analyzing how good the classifier can identify records of different classes and ROC graph was also used to compare the performance of the classifiers [3]. The performance of the proposed system is demonstrated below-

A. Experiments

1) Experiment one

The first experiment was designed to estimate the performance of a J48 unpruned tree classifier in classifying and predicting heart disease. In this experiment, two ways were considered. On the first way, the proposed classifier was run on a full training set containing 542 instances with 11 attributes. It took 0.11second to build the model and the model generated a tree with a size of 5 and 3 leaves. On the second way, it was run on a full training set containing 542 instances with 3 selected attributes. It took 0.02 second to build the classifier and generated tree with a size of 3 and 2 leaves. This tree is simple and faster than the experiment conducted on full training set with all attributes. Table 2 shows the confusion matrix for experiment one. Also, the performance measures used are shown in Table 3.

TABLE 2. CONFUSION MATRIX FOR EXPERIMENT ONE

Classifier	Confusion Matrix		
	Normal (Predicted)	*Abnormal (Predicted)*	*Actual*
J48 unpruned with all attributes	512	3	***Normal***
	2	25	***Abnormal***
J48 unpruned with 3 selected attributes	515	0	***Normal***
	0	27	***Abnormal***

TABLE 3. DETAILED PERFORMANCE MEASURES FOR EXPERIMENT ONE

Classifier	Accuracy	TP Rate	Precision	F-Measure	ROC Area
J48 unpruned with all attributes	99.41%	0.994	0.994	0.994	0.995
J48 unpruned with 3 selected attributes	100%	1	1	1	1

The results show that performance of J48 unpruned classifier was better on the reduced attributes, the classification accuracy increased from 99.41 % to 100%.

2) Experiment two

The performance of a J48 pruned decision tree classifier in predicting heart diseases is evaluated in this experiment. Table 4 and Table 5 summarize the results. Also, the results showed the effect of attribute selection on classification accuracy and other performance measures used in this experiment.

TABLE 4. CONFUSION MATRIX FOR EXPERIMENT TWO

Classifier	Confusion Matrix		
	Normal (Predicted)	*Abnormal (Predicted)*	*Actual*
J48 pruned with all attributes	513	2	***Normal***
	1	26	***Abnormal***
J48 pruned with 3 selected attributes	515	0	***Normal***
	0	27	***Abnormal***

TABLE 5. DETAILED PERFORMANCE MEASURES FOR EXPERIMENT TWO

Classifier	Accuracy	TP Rate	Precision	F-Measure	ROC Area
J48 pruned with all attributes	99.61%	0.996	0.996	0.996	0.997
J48 pruned with 3 selected attributes	100%	1	1	1	1

3) Experiment three

The third experiment was used to evaluate the performance of Naive Bayes classifier. As mentioned in previous two experiments, two ways were considered and the computational results are shown in Table 6 and Table 7 below.

TABLE 6. CONFUSION MATRIX FOR EXPERIMENT THREE

Classifier	Confusion Matrix		
	Normal (Predicted)	*Abnormal (Predicted)*	*Actual*
Naive Bayes with all attributes	505	10	***Normal***
	0	27	***Abnormal***
Naive Bayes with 3 selected attributes	510	5	***Normal***
	0	27	***Abnormal***

TABLE 7. DETAILED PERFORMANCE MEASURES FOR EXPERIMENT THREE

Classifier	Accuracy	TP Rate	Precision	F-Measure	ROC Area
Naive Bayes with all attributes	98.16%	0.982	0.987	0.983	0.998
Naive Bayes with 3 selected attributes	99.08%	0.991	0.992	0.991	0.999

The performance of Naïve Bayes classifier was better on the selected attributes in all performance measures used as shown in Table 7.

4) Experiment four

In this experiment, the performance of MLPNN in predicting heart disease is evaluated. The results of this experiment are given below in Table 8 and detailed performance measures used are shown in Table 9.

TABLE 8. CONFUSION MATRIX FOR EXPERIMENT FOUR

Classifier	Confusion Matrix		
MLPNN with all attribute	*Normal (Predicted)*	*Abnormal (Predicted)*	*Actual*
	513	2	*Normal*
	3	24	*Abnormal*
MLPNN with 3 selected attributes	515	0	*Normal*
	0	27	*Abnormal*

TABLE 9. DETAILED PERFORMANCE MEASURES FOR EXPERIMENT FOUR

Classifier	Accuracy	TP Rate	Precision	F-Measure	ROC Area
MLPNN with all attribute	99.08%	0.991	0.991	0.991	0.999
MLPNN with 3 selected attributes	100%	1	1	1	1

The obtained results showed that MLPNN classifier performance better on the reduced attributes as accuracy classification increased to 100% from 99.08 % and the execution time decreased to 1.14 second from 3.09 second. MLPNN takes long execution time because it contains huge number of parametres (speed of learning, number of hidden layers, input neurons, number of iteration, updating weights, …) to set up.

VI COMPARATIVE STUDY

The performance of data mining classification techniques used in the proposed system is compared through Accuracy, TP-rate, Precision, F-Measure, ROC area and execution time. It is difficult to compare the accuracy of techniques and determine the best one because their performance is data-dependent. So, the comparative study in this work considered a specific dataset. Table 10 summarizes the comparative performance of various classifiers. Table 10 shows comparative performance based on different measures applied, which proves that all classification techniques with 3 selected attributes outperformed all classification techniques with all attributes.

TABLE 10. SUMMARIZING PERFORMANCE OF VARIOUS CLASSIFIERS IN THE PROPOSED SYSTEM.

Classifier	Accuracy	TP Rate	Precision	F-Measure	ROC Area	Execution time
J48 unpruned with all attributes	99.41%	0.994	0.994	0.994	0.995	0.11
J48 pruned with all attributes	99.61%	0.996	0.996	0.996	0.997	0.13
J48 unpruned with 3 selected attributes	100 %	1	1	1	1	0.02
J48 pruned with 3 selected attributes	100 %	1	1	1	1	0
Naive Bayes with all attributes	98.16 %	0.982	1	1	0.998	0.03
Naive Bayes with 3 selected attributes	99.08 %	0.991	1	1	0.999	0
MLPNN with all attributes	99.08 %	0.991	0.991	0.991	0.999	3.09
MLPNN with 3 selected attributes	100 %	1	1	1	1	1.14

The confusion matrix denotes samples classified as Normal, others as Abnormal and misclassified. Evaluation of the confusion matrices show that J48 unpruned, J48 pruned, and MLPNN for 3 selected attributes show 100% accuracy. The techniques strongly suggest that data mining classification algorithms are able to predict a class for diagnosis. The confusion matrix validates the effectiveness of the classifier.

The experimental results have shown that, in general, J48 decision tree classifier outperformed Naïve Bayes classifier and MLPNN. J48 classifier outperformed the other classifiers by achieving the highest Accuracy, TP rate, Precision, F-Measure, ROC area, and execution time values in the domain of predicting heart disease cases. One possible explanation for superiority of J48 classifier over MLPNN and Naïve Bayes classifier is the nature of the dataset used in this study. Decision tree classifiers tend to perform better on simple datasets and this leads to a conclusion that the classification problem presented by the dataset is a simpler one. Also, J48 pruned decision tree reduces the complexity of the final classifier, and hence improves prediction accuracy by reduction of overfitting.

Several systems related to heart disease diagnosis using data mining techniques are proposed. These systems have applied different techniques to the given problem and achieved different classification accuracies. Table 11 depicts the different systems with their techniques used in heart disease diagnosis along with their accuracies. In addition, Table 11 shows that the proposed system outperformed the other related systems by achieving the highest accuracy.

TABLE 11. DIFFERENT DATA MINING TOOLS USED IN HEART DISEASE PREDICTION SYSTEM WITH ACCURACY.

Systems	Techniques Used	Data Mining Tools	Accuracy
A.Taneja (2013) [22]	J48, Naive Bayes	WEKA 3.6.4	95.56%,92.42%
C. S. Dangare et al (2012) [23]	Neural Network	WEKA 3.6.6	100%
N. Bhatla et al (2012) [24]	Naive Bayes, Decision Trees, Neural Networks	WEKA 3.6.6	90.74%, 99.62%, 100%
		TANAGRA	52.33%, 52%, 45.67%
		WEKA 3.6.0	86.53%, 89%, 85.53%
		.NET platform	96.5%, 99.2%, 88.3%
A. Shetty et al (2016) [25]	MLPNN	WEKA and MATLAB	84%
	Genetic Algorithm		89%
M. M. Kirmani (2017) [21]	Naive Bayes MLPNN	WEKA	99.52%
		.NET platform	96.56%
	Fuzzy Logic	TANAGRA	83.85%
The proposed system	J48 unpruned with all attributes	C# and WEKA 3.6.11	99.41 %
	J48 pruned with all attributes		99.61%
	J48 unpruned with 3 selected attributes		100 %
	J48 pruned with 3 selected attributes		100 %
	Naive Bayes with all attributes		98.16 %
	Naive Bayes with 3 selected attributes		99.08 %
	MLPNN with all attributes		99.08 %
	MLPNN with 3 selected attributes		100 %

V II. CONCLUSION AND FUTURE WORK

Medical diagnosis is considered as a significant task that needs to be carried out precisely and efficiently. In this paper, an intelligent predictive system using classification techniques for heart disease diagnosis is proposed. In the proposed system various classification techniques namely, J48 decision tree, Naïve Bayes and MLPNN in WEKA tool are used to predict the heart disease and to compare the best classifier of prediction. The outcome of predictive classification techniques on the same dataset reveals that all classification techniques used with 3 selected attributes outperformed the same techniques with all attributes. The most effective classifier to predict patients with heart disease appears to be a J48 classifier implemented on 3 selected attributes with a classification accuracy of 100%. The predictive accuracy determined by classification techniques used suggests that parameters used are reliable indicators to predict the presence of heart disease. WEKA will definitely help the medical practitioners to predict the heart disease and will help the researchers to select the appropriate classifier.

The proposed work can be further enhanced and expanded for the automation of heart disease prediction by using hybrid classification techniques. These techniques will be compared for the optimum accuracy.

REFERENCES

[1] B. Venkatalakshmi, and M. V. Shivsankar, " Heart Disease Diagnosis Using Predictive Data mining," International Conference on Innovations in Engineering and Technology (ICIET'14), vol.3, no.3, 2014.
[2] S. B. Shinde, and A. Priyadarshi, " Diagnosis of Heart Disease Using Data Mining Technique," International Journal of Science and Research (IJSR),vol. 4, no. 2, pp.2301-2303, 2015.
[3] M. Bramer, Principles of Data Mining , Springer- Verlag London-Limited, 2007.
[4] R. R. Patil "Heart Disease Prediction System using Naïve Bayes and Jelinek–Mercer Smoothing," IJARCCE, vol. 3, no. 5, pp. 6787-6789, 2014.
[5] R. Tamilarasi, and R. Porkodi,"A Study and Analysis of Disease Prediction Techniques in Data Mining for Healthcare," International Journal of Emerging Research in Management &Technology, vol.4, no.3, pp. 76-82, 2015.
[6] D. Jain, and V. Singh, " Utilization of Data Mining Classification Approach for Disease Prediction: A Survey," International Journal Education and Management Engineering, vol. 6, pp. 45-52, 2016.
[7] K. Manimekalaai, "A Proficient Heart Disease Prediction Method Using Different Data Mining Tools," International Journal of Engineering Science and Computing, vol. 6, no.3, pp. 2676-2680, 2016.
[8] A. Ahmed, and S. Abdul Hannan," Data Mining Techniques to Find Out Heart Diseases: An Overview," International Journal of Innovative Technology and Exploring Engineering, vol.1, no. 4, 2012.
[9] N. Guru, A. Dahiya, and N. Rajpal, "Decision Support System for Heart Disease Diagnosis Using Neural Network," Delhi Business Review, vol. 8, no. 1, 2007.
[10] A. Rajkumar, and G. S. Reena, "Diagnosis of Heart Disease Using Data Mining Algorithm," Global Journal of Computer Science and Technology, vol.10, Issue 10, pp. 38-43, 2010.
[11] P. Gupta, and P. Bajaj, " Heart Disease Diagnosis System Based on Data Mining and Neural Network," JESRT, vol. 3, no. 6, pp.172-176, 2014.

[12] D. Kundra, and N. Kaur "Review on Prediction System for Heart Diagnosis using Data Mining Techniques," International Journal of Latest Research in Engineering and Technology (IJLRET), vol.1, Issue 5, pp. 9-14, 2015.
[13] D. Chandna, "Diagnosis of Heart Disease Using Data Mining Algorithm," International Journal of Computer Science and Information Technologies, vol.5, no.2, pp.1678-1680, 2014.
[14] B. Kaur, and W. Singh ," Review on Heart Disease Prediction System using Data Mining Techniques," IJRITCC, vol. 2, no.10, pp. 3003 – 3008, 2014.
[15] S. Palaniappan, and R. Awang, "Intelligent Heart Disease Prediction System using Data Mining Techniques," International Journal of Computer Science and Network Security, vol. 8, no. 8, pp. 343-350, 2008.
[16] M. Abdar et al., " Comparing Performance of Data Mining Algorithms in Prediction Heart Diseases, " International Journal of Electrical and Computer Enginnering (IJECE), vol. 5, no. 6, 2015.
[17] R. Thanigaivel, and K. R. Kumar, "Boosted Apriori: An Effective Data Mining Association Rules for Heart Disease Prediction System ," Middle-East Journal of scientific Research, vol. 24, pp. 192-200, 2016.
[18] J. Banupria, and S. Kiruthika, " Heart disease using Data Mining Algorithm on Neural Network and Genetic Algorithm," International Journal of Advanced Research in Computer Science and Software Engineering, vol. 6, Issue 8, pp.40-42, 2016.
[19] C. Sowmuiya, and P. Sumitra, "Comparative Study of Predicting Heart Disease by Means of Data Mining," International Journal of Engineering and Computer Science, vol. 5 , Issue 12, pp. 19580-19582, 2016.
[20] H. D. Mosethe, and M. A. Masethe "Prediction of Heart Disease using Classification Algorithms," WCECS, vol 11, San Francisco, USA, 2014.
[21] M. M. Kirmani, " Heart Disease Prediction using Multilayer Perceptron Algorithm," International Journal of Advanced Research in Computer Science, vol. 5, no. 8, pp. 1169-1172, 2017.
[22] A. Taneja, "Heart Disease Prediction System using Data Mining Techniques," Oriental Journal of Computer Science and Technology, vol. 6, no. 4, pp. 457-466, 2013.
[23] C. S. Dangare, and S. S. Apte, " Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques, " International Journal of Computer Applications, vol. 47, no. 10, pp. 44-48, 2012.
[24] N. Bhatla, and K. Jyoti, "An Analysis of Heart Disease Prediction using Different Data Mining Techniques," International Journal of Engineering Research & Technology (IJERT), vol. 1, Issue 8, 2012.
[25] A. Shetty, and C. Naik, "Different Data Mining Approaches for Predicting Heart Disease," International Journal of Innovative Research in Science, Engineering and Technology. vol. 5, Special Issue 9, 2016.
[26] WEKA:// http://www.cs.Waikato.ac.nz/mc/weka/index.html.