Asian Journal of Research in Computer Science

12(2): 57-68, 2021; Article no.AJRCOS.61760
ISSN: 2581-8260

Efficient Data Mining Techniques for Heart Disease Prediction and Comparative Analysis of Classification Algorithms

Md. Ashikur Rahman Khan[1*], Masudur Rahman[1], Jayed Us Salehin[1], Md. Saiful Islam[1] and Md. Fazle Rabbi[1]

[1]Noakhali Science and Technology University, Noakhali, Bangladesh.

Authors' contributions

This work was carried out in collaboration among authors. Author MR designed the study, managed the data collection and literature searches, performed the prediction and comparative analysis and wrote the first draft of the manuscript. Author MARK supervised the work. Authors JUS and MSI co-supervised the research. Others are helped for data collection, designing the method, writing research, managed and analysis of the study. All authors read and approved the final manuscript.

Article Information

DOI: 10.9734/AJRCOS/2021/v12i230281
Editor(s):
(1) Dr. Manish Mahajan, CGC College of Engineering, India.
Reviewers:
(1) G. Suseendran, Vels Institute of Science Technology & Advanced Studies (VISTAS), India.
(2) M. R. Ebenezar Jebarani, Sathyabama Institute of Science and Technology, India.
Complete Peer review History: https://www.sdiarticle4.com/review-history/61760

Original Research Article

Received 23 July 2020
Accepted 30 September 2020
Published 16 October 2021

ABSTRACT

Data mining techniques are used to extract interesting patterns and discover meaningful knowledge from huge amount of data. There has been increasing in usage of data mining techniques on medical data for determining useful trends and patterns that are used in analysis and decision making. About eighty percent of human deaths occurred in low and middle-income countries due to heart diseases. The healthcare industry generates large amount of heart disease data which are not organized. These data make the prediction process more complicated and voluminous. Data mining provides the techniques for fast and accurate transformation of data into useful information for heart diseases prediction. The main objectives of this research is to predict heart diseases more accurately using Naïve Bayes, J48 Decision Tree, Neural Network, Random Forest classification algorithms and compare the performance of classifiers. The research uses raw dataset for performance analysis and the analysis is based on Weka Tool. This research also shows best technique from them which is Random Forest on the basis of accuracy and execution time.

**Corresponding author: E-mail: ashik.nstu@yahoo.com;*

Keywords: KDD; WEKA; MLP; CSV; ARFF.

1. INTRODUCTION

The number of diseases in the world increasing day by day as a result the amount of medical data in the health sector also increasing and it becomes one of the challenges in medical science. World health organization has estimated 17.5 million people died from cardio vascular diseases in 2012, representing 31 percent of all global deaths. Out of these, an estimated 7.4 million were due to coronary heart disease and 6.7 million were due to stroke. According to the WHO country profile for 2018, cardiovascular disease alone kills 2.56 lakh people in Bangladesh accounting for 30 per cent of deaths caused by Non-Communicable Diseases. However, 90% of those deaths were estimated to be preventable if patients have correctly been diagnosed early and they improved their habits such as: healthy eating, exercise, and alike [1].

In the diagnosis of heart disease large number of work is carried out, researchers have been investigating the use of data mining techniques to help professionals. Data mining has recently become one of the most advanced and encouraging fields for the extraction and manipulation of data to produce useful information. It is an assortment of algorithmic techniques to extract instructive patterns from raw data. Data mining uses a series of pattern recognition technologies and statistical and mathematical techniques to discover the possible rules or relationships that govern the data in the databases. The Data mining tasks can be classified into two categories Predictive and Descriptive. A Predictive model makes a prediction about values of data using known results found from different data and its goal is to identify strong links between variables of a data table. Predictive model data mining tasks involves the classification, prediction, time series analysis and regression.

Nowadays, many hospitals keep their present data in electronic form through some hospital database management system. These systems generate large volume of data on daily basis. This data may be in form of free text, structured as in databases or in form of images [2]. This data may be used to extract meaning information which may be used for decision making. Data mining which is one of the KDD (Knowledge discovery in database) concentrates on finding meaningful patterns from large datasets. These patterns can be further analyzed and the result can be used for further valuable decision making and analysis. It has helped to confirm the best prediction technique in terms of its accuracy and error rate on the specific dataset. Knowledge Discovery (KD) uses data mining generally consists of some phases: Data Selection, Data preprocessing, Data transformation, Data mining, Pattern evaluation, Knowledge presentation. The following Fig. 1 provides the graphical description about KDD working process.

Researchers have been applying different data mining techniques to help health care professionals with improved accuracy in the diagnosis of heart disease.

1.1 Problem Statements

Today, Healthcare organizations produces huge amounts of multifarious data about hospitals, resources, disease diagnosis, electronic patient records, etc. This increase in data volume automatically requires the data to be retrieved when needed. The large amount of data is critical to be processed and analyzed for knowledge extraction that empowers support for understanding the prevailing circumstances in healthcare industry. With the use of data mining techniques is possible to extract the knowledge and determine interesting and useful patterns. The knowledge gained in this way can be used in the proper in order to improve work efficiency and enhance the quality of decision making.

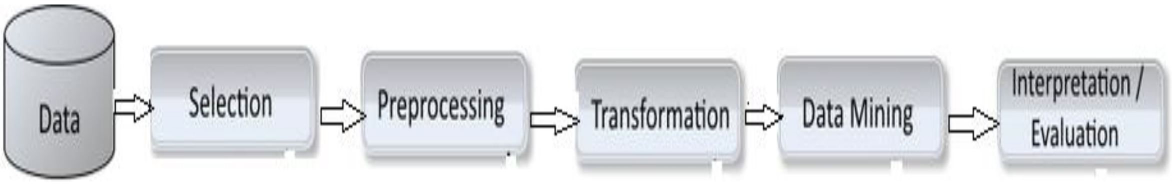

Fig. 1. Knowledge Discovery (KDD) Diagram

A major problem in health science or bioinformatics exploration is in managing the correct diagnosis of certain important information. In this respect, it is critical that the healthcare industry look into how data can be better captured, stored, prepared and mined. Possible directions include the standardization of clinical vocabulary and the sharing of data across organizations to enhance the benefits of healthcare data mining applications. Detection of the disease in the early stage become difficult for large dataset. The accuracy of heart diseases prediction become lower with higher runtime.

Most of the algorithms (like ID3) require that the target attributes have only discrete values because decision trees use the divide and conquer method. If there are more complex interactions among attributes exist then performance of decision trees is low. Their performance is better only when there exist a few highly relevant attributes. On the other hand training or learning process in large dataset is very slow and computationally very expensive. Data sharing is another major problem neither patients nor healthcare organizations are interested in sharing of their private data.

The researchers all over the world are working on this issue. They are trying to develop better algorithm comparatively lower execution time and better accuracy for easily detect this disease than the traditional techniques. We use different classification methods including Decision tree, Bayesian algorithms (Naive Bayes), Neural Network (Multilayer Perceptron) and others classification methods to predict heart diseases easily with lower execution time and higher accuracy. Hence it will provide revolutionary changes in medical sector.

1.2 Objectives

This research paper works with classifier algorithms and provides better result that will directly help to the physicians for accurate diagnosis. The main purposes of this research are as follows:

1. Increasing the accuracy of heart disease prediction and develop a model for heart disease prediction.
2. Filtering raw data to remove noise and redundant data.
3. Reducing the runtime and make comparisons among classifiers based on accuracy and building time.
4. Comparative analysis the result of classifier algorithms.
5. To find out the best classifier algorithm for heart disease prediction from the selected classifiers
6. To extract the predictive class.

2. RELATED WORKS

A lots of research has been done on Heart disease prediction using datamining. Researcher always trying to improve the overall performance of early heart disease detection using data mining techniques. In this context, researchers have been applied different data mining techniques such as: Association Rules Technique, Clustering, and Classification Algorithms to extract significant patterns for prediction of heart diseases with good accuracy.

In 2010, M. Anbarasi et al. Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm was proposed [3]. The Genetic Algorithm is an optimization technique inspired by natural selection and natural genetics. A Heart disease system was developed to predict accurately the presence of heart disease with reduced number of attributes. There are three classifiers Decision tree, Classification with clustering and Naive Bayes were used for diagnosis of patients with heart disease and the Weka data mining tool was used for experiments. These experiments show that the Decision tree has highest accuracy as well as construction time as compare to others Naïve Bayes and Classification via clustering.

In 2011, AH Chen et al. introduced HDPS: Heart Disease Prediction System in which only one data mining algorithm is used that is artificial neural network (ANN) that is used to classify the heart disease based on 13 different attributes [4]. The data set is used in this system is taken from UCI machine learning repository having 303 instances.

In 2012, Chaitrali S. Dangare et al. used data mining neural network approach for prediction about heart disease [5]. This shows about 100% accuracy. Multilayer Perceptron Neural Network (MLPNN) with Back propagation algorithm (BP) was used in the system. WEKA data mining tool is used for the experiments and the data set for this contains 573 records which are divided into two parts training and testing. Total 15 attributes were used in this to increase the accuracy of the prediction.

In 2011, Jyoti Soni et al. introduced an intelligent and effective heart disease prediction system using Weighted Associative Classifiers (WAC), in WAC different weights are assigned to the attributes according to their capability about predicting [6]. The system is implemented by using JAVA platform and benchmark data from online available UCI repository. The data set contain 303 records and 14 different attributes. The results of experiments show about 80% accuracy of WAC.

In 2013, Vikas Chaurasia et al. used different data mining approaches to predict heart diseases. Different experiments were conducted using Weka tool. The results of these experiments were compared with each other using 10 cross validations with and without bagging. Bagging means Bootstrap aggregation which is used to increase the accuracy of the classification. Three techniques were used and compare in this research i.e. Naïve Bayes, J48 Decision Tree and Bagging. The data set was taken from Hungarian Institute of Cardiology, that have 76 raw attributes but only 11 attributes were selected for experiments. The experiments also show that the bagging has the highest accuracy i.e. 85.03% and J48 decision tree and naïve Bayes have 84.35% and 82.31% accuracy respectively [7].

In 2012, Shadab et al. used Naive Bayes technique using 15 attributes in the dataset for the heart diagnosis in heart prediction system [8].

In 2013, Rashedur et al. used Neural network technique using Weka data mining tool and achieved 79.19% and to compare various classification techniques, he used another technique fuzzy logic using TANGRA data mining tool and achieved 83.85% accuracy [9].

In 2010, Bagging algorithms used in many researches works to improve model stability and accuracy of medical data set. My Chau Tu's used bagging algorithm to identify the heart disease [10].

In 2012, Aqueel Ahmed et al. show the classification techniques in data mining and show the performance of classification among them. In this classification accuracy among these data mining has discussed. In this decision tree and SVM perform classification more accurately than the other methods and was able to achieve 91% accuracy [11].

In Moreno et al. applied an association algorithms successfully for prediction in health domain, they discovered huge number of rules (some of them are not interested), in addition generally the performance of association algorithms is low [12].

In 2015, Chaitrali S. et al. showed that Artificial Neural Network outperforms other data mining techniques such as Decision Tree and Naïve Bayes. In this research work, Heart disease prediction system was developed using 15 attributes [13]. The research work included two extra attributes obesity and smoking for efficient diagnosis of heart disease in developing effective heart disease prediction system.

In 2016, Kumaravel et al. have proposed automatic diagnosis system for heart diseases using neural network. In this system ECG data of the patients is used to extract features and 38 input parameters are used to classify 5 major types of heart diseases with accuracy of 63.6 - 82.9% [14].

In 2016, B. Venkatalakshmi et al. performed an analysis on heart disease diagnosis using data mining techniques Naïve Bayes and Decision Tree techniques. Different sessions of experiments were conducted with the same datasets in WEKA 3.6.0 tool. Data set of 294 records with 13 attributes was used and the results revealed that the Naïve Bayes outperformed the Decision tree techniques [15].

In 2016, Aditya Methaila et.al. In their research work focused on using different algorithms and combinations of several target attributes for effective heart attack prediction using data mining. Decision Tree has outperformed with 99.62% accuracy by using 15 attributes. Also, the accuracy of the Decision Tree and Bayesian Classification further improves after applying genetic algorithm to reduce the actual data size to get the optimal subset of attribute sufficient for heart disease prediction [16].

In 2018. Benjamin Fredrick David and S. Antony Belcy in their research work, experimental setup has been made for the evaluation of the performance of algorithms with the help of heart disease benchmark dataset retrieved from UCI machine learning repository. It is found that Random Forest algorithm performs best with 81% precision when compared to other algorithms for heart disease prediction [17].

3. METHODOLOGY

3.1 Introduction

In this chapter, the heart disease dataset that had been studied was discussed in detail. We elaborated briefly how the current data had been collected and then moved on to how the datasets were built and formatted to train the classifiers. The current research intends to improve the diagnosing of heart diseases by examining the patient's symptoms using data mining classification methods. To achieve this goal, a literature review was carried out to review the data mining works related to diagnose heart diseases. For this purpose, we need follow the following process carefully.

3.1.1 Attributes selection and description

There are many attributes used for heart disease diagnosis. However, most of the published experiments commonly refer to 13 of them. This research used different types of input attributes for the heart disease are age, sex, chest pain type, blood pressure, cholesterol, blood sugar ECG result, old peak, slop etc. Here is given 13 attributes where some of attributes are nominal and others attributes are numeric and one attribute is class variable. The attributes of the dataset and their description is given in Table 1. The disease is the target class of the dataset denoting the heart disease presence with a yes or a no. Similarly, all the attributes and their values are represented in Table 1.

3.1.2 Data collection

Data are collected from different hospitals of heart patients for the research purpose. Raw data is collected from heart disease patients for making heart disease predictions. These datasets are integrated into a single dataset. It includes 13 attributes and one class value. The heart disease dataset included in this research work consists of total 305 instances.

3.1.3 Data pre-processing

In real world, data are generally incomplete, noisy and inconsistent. For predicting data mining, data in raw form are not best for analysis. Missing values are extremely common in medical records. However, these values must be treated before being used as they may lead to failure classification or incorrect disease prediction [18]. Knowledge Discovery (KDD) is a process that allows automatic scanning of high-volume data in order to find useful patterns that can be considered knowledge about the data Data preprocessing is a first step of the Knowledge discovery in databases (KDD) process. This paper used WEKA tool for that purpose. Weka tool support dataset in ARFF (Attribute-Relation File Format) file format. To do so, we created ARFF file which could be edited by Notepad++. The converted dataset format in ARFF file.

3.1.4 Data analysis and visualization

This step will represent “PRESENT” and “ABSENT” of heart disease among the instances. This will show that among all instances, some instances are classified as present of heart disease and others instances are classified as absent of heart disease. Here, this paper visualized all attributes using WEKA classifier software This figure shows the distribution of the attribute values with respect to the class attribute. In this picture blue color represents “PRESENT” and red color represents “ABSENT” of heart disease among the instances. Here, we can see that among all 305 instances, 174 instances are classified as heart patient and others 131 instances are classified as absent of heart disease.

3.2 Classification Algorithms

Different machine learning classification algorithms and their implementations needed for predicting heart disease. Classification is the process of building a model of classes from a set of records that contain class labels. In this paper we use the following data mining techniques.

3.2.1 Naive bayes

Naive Bayes Algorithm is used to generate mining models. The algorithm calculates the probability of every state of each input column given predictable columns possible states. This algorithm is based on statistics. It is used for estimating the probability of a class value during classification and prediction. The common principle of Naive Bayes algorithm is all Naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature. As the classifier returns probabilities, it is very easy to apply these results to a large number of tasks than if an arbitrary scale was used [19]. Naïve Bayes theorem provides a way of calculating posterior probability P(c|x) from Class prior probability P(c), predictor prior probability P(x) and Like hood P(x|c). The equation is in below:

$$P(c|x)=P(x|c)P(c)/P(x)$$

$$P(c|X)=P(x_1|c)\times P(x_2|c)\times\ldots\times P(x_n|c)\times P(c)$$

3.2.2 Decision tree

Decision Tree techniques has shown useful accuracy in the diagnosis of heart disease. In medical field decision trees determine the sequence of attributes. First, it produces a set of solved cases. Then the whole set is divided into training set and testing set. Where a training set is used for the induction of a decision tree. While the testing set is used to find the accuracy of an obtained solution [20]. Decision Tree has become popular in knowledge discovery because the construction of decision tree classifier does not require any domain knowledge. Successful decision tree model depends upon the data but in general it has good accuracy. The sample decision tree is shown in Fig. 2.

3.2.2.1 J48

J48 algorithm generates the rules for the prediction of the target variable. This paper used J-48 decision tree algorithm. This method removes the least reliable branches. It uses the concept of information entropy. To make the decision the attribute with highest information gain is used and information gain is basically the difference in entropy.

Table 1. Attributes and their descriptions

Attribute Name	Description
Age	patient's age in years
Gender	sex of patient
Chest pain	chest pain
Blood_ Pressure	resting blood pressure (in mm Hg on admission to the hospital)
Chol	serum cholesterol in mg/dl
f_ Blood_ sugar	fasting blood sugar > 120 mg/dl
r_ ECG_ results	resting electrocardiographic results
maxi_ heart_ rate	maximum heart rate achieved
Exercise	Exercise include angina
Oldpeak	ST depression induced by exercise relative to rest
Slope	the slope of the peak exercise ST segment
Number of major vessels	number of major vessels colored by fluoroscopy
Defect type	type of defect
Disease	diagnosis of heart disease

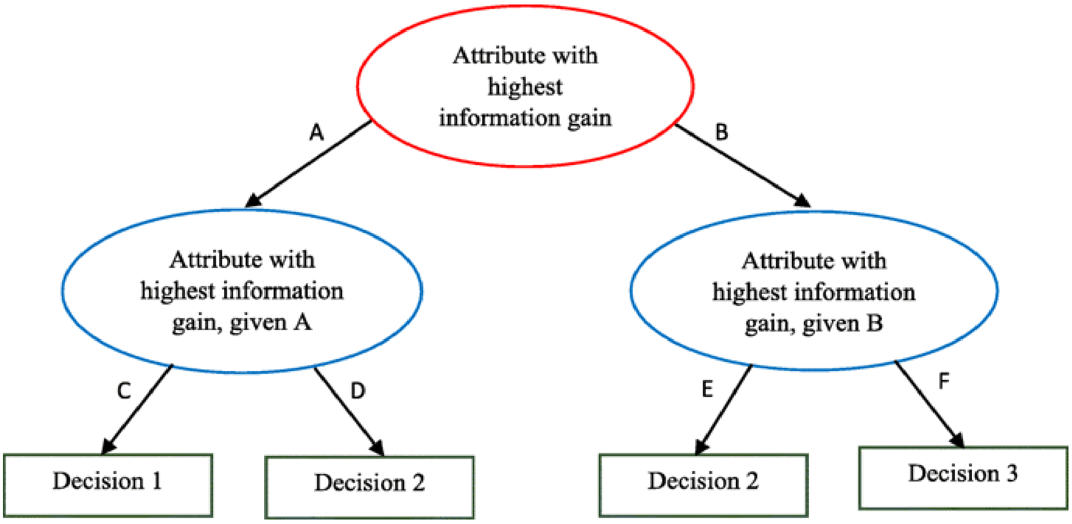
Attribute with highest information gain
A
B
Attribute with highest information gain, given A
Attribute with highest information gain, given B
C
D
E
F
Decision 1
Decision 2
Decision 2
Decision 3

Fig. 2. A simple decision tree

Information gain is used in C4.5 to collect the information in data sets. And this information is used to take the decisions. It is the mathematical tool that algorithm J48 has used to decide, in each tree node, which variable fits better in terms of target variable prediction. This algorithm it generates the rules from which particular identity of that data is generated. The objective is progressively generalization of a decision tree until it gains equilibrium of flexibility and accuracy.

Basic Steps in the Algorithm: [20]

(i) In case the instances belong to the same class the tree represents a leaf so the leaf is returned by labeling with the same class.
(ii) The potential information is calculated for every attribute, given by a test on the attribute. Then the gain in information is calculated that would result from a test on the attribute.
(iii) Then the best attribute is found on the basis of the present selection criterion and that attribute selected for branching.

The objective is to maximize the Gain, dividing by overall entropy due to split argument by value. Because of the outliers this is a significant step to the result.

3.2.2.2 Random forest

Random forests (RF) are combination of tree predictors using decision tree such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest [21]. The generalization error of a forest of tree classifiers depends on the strength of the individual trees in the forest and the correlation between them. They are more robust with respect to noise. It is a supervised classification algorithm used for the prediction and it is considered as the superior due to its large number of trees in the forest giving improved accuracy than decision trees. Typically, the trees are trained independently and the predictions of the trees are combined through averaging. Random forest algorithm can use both for classification and the regression based on the problem domain. To perform the prediction using the trained random forest algorithm, we need to pass the test features through the rules of each randomly created trees.

3.2.3 Neural network (multilayer perceptron)

Artificial Neural Networks are the human neurons type network structure which consists of number of nodes that are connected through directional links where each node represents a processing unit and the links between them specify the casual relation between them. This classification technique is becoming powerful tool in data mining and may be used for different purposes in descriptive and predictive data mining. Multilayer Perceptron classifier is based on back propagation algorithm to classify instances of data. It is a feed forward neural network with one input and output layer with several possible hidden layers that are totally interconnected. It should be learned by a set of weights for predicting the class label of tuples. This neural network consists of three layers namely input layer, one or more hidden layers, and an output layer. The sample artificial neural network is shown in Fig. 3.

To make input layer, the inputs are fed simultaneously into the units. These inputs are passed through the input layer and are then weighted and fed simultaneously to a second layer of neuron like units, which is known as a hidden layer. The outputs of the hidden layer units can be input to another hidden layer, and so on. At the core, back propagation is simply an efficient and exact method for calculating all the derivatives of a single target quantity (such as pattern classification error) with respect to a large set of input.

3.3 Data Mining Tool WEKA

Waikato Environment for Knowledge Analysis (Weka) is an open source Java based platform containing various machine learning algorithms for data mining tasks. For experimentations and implementations Weka is used as the data mining tool. It is a collection of tools for Regression, Clustering, Association, Data pre-processing, visualization. WEKA is a very good data mining tool for the users to classify the accuracy on the basis of datasets by applying different algorithmic approaches and compared in the field of bioinformatics. In WEKA dataset can be preprocessed then fed dataset into learning schema, and analyze the resulting classifier and its performance. It supports CSV (Comma Separated Value) and ARFF (Attribute-Relation file format) value.

3.4 Proposed Model

The heart disease prediction can be performed by following the procedure which is similar to Fig. 4 which specifies the research methodology for building a classification model required for the prediction of the heart diseases in patients. The model forms a fundamental procedure for carrying out the heart disease prediction using any machine learning techniques. First, heart diseases patient records are collected that preprocess and extract some features for analyzing further manipulation of it. Then, Four different classification algorithms used to measure the performance of heart disease. In figure we represent several steps how to implement our proposed model. Those steps are described briefly as follows:

Heart patient records are collected from different hospitals. From there, we have collected 305 heart patient records for analysis. The selected data was checked for noise, inconsistency and missing values. Noises and inconsistencies identified in the dataset were corrected manually, while missing values were replaced with the most similar value. Weka tool used for filtering purpose. This research performed 10-fold cross validation technique on four different classifier algorithms to evaluate predictive models by partitioning the original sample into a training set to train the model, and test set to evaluate it. This paper applied 10-fold cross validation with Decision tree (J48, Random Forest), Bayesian algorithms (Naive Bayes), Neural Network (Multilayer Perceptron) classifiers.

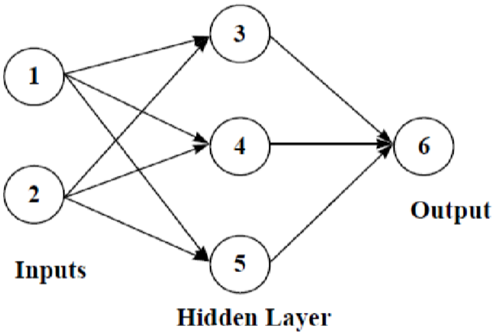

Fig. 3. Sample artificial neural network

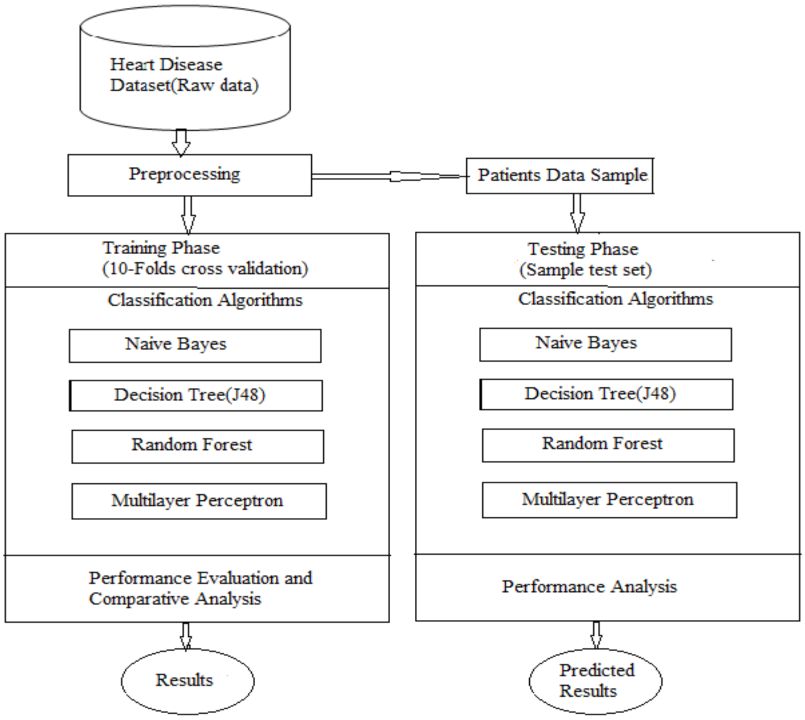

Fig. 4. Block diagram of proposed system

4. RESULTS AND DISCUSSION

4.1 Performance Evaluation of Algorithms

Five common performance measures have been used to evaluate the accuracy of classification algorithms. These measures were selected because they are widely used to assess performance of classification models. For finding these measurements we need to calculate a confusion matrix. The Confusion matrix is one of the most intuitive and easiest metrics used for finding the correctness and accuracy of the model. It is used for Classification problem where the output can be of two or more types of classes (Present and Absent). The confusion matrix, is a table with two dimensions ("Actual" and "Predicted"), and sets of "classes" in both dimensions. Actual classifications are columns and Predicted ones are Rows. A confusion matrix we can create as follows.

The performance evaluation for each classifier models depend on the following terms:- Recall, Error rate (ERR), Accuracy, . Precision, Specificity and F-measures.

4.2 Evaluation and Comparisons of Classification Algorithms

In order to evaluate the performance and usefulness of four different classification algorithms for predicting Heart disease WEKA tool is used. After performing different classification algorithm in WEKA, value of different term for each classification algorithm are listed in a table to measure and investigate the performance on the classification methods. Here Table 3 shows different performance parameters like TP rate, FP rate, and Precision, Recall, F-measure, accuracy, build time and ROC area are for different classifiers.

The Table 3 presents the great accuracy of classifier algorithms on the given heart disease dataset, the lowest accuracy provided by Naïve Bayes is 86.22% and the highest accuracy provided by Random forest is 97.70%. Fig. 5. also shown the graphical presentation of accuracy of all algorithms.

On the other hand, if we look at the building time taken by all classifiers then Neural Network takes 0.45 seconds, Random Forest takes 0.17 seconds, Decision Tree(J48) takes 0.02 seconds and Naïve Bayes takes 0 seconds. Based on the building time, Naïve Bayes takes lowest times and Multilayer Perceptron takes highest times.

Other performances measures like TP rate and FP rate that are also used to compare the results also achieve remarkable performance and are shown in the above table. The TP rate and the FP rate were, (0.862, 0.151) for Naïve Bayes, (0.951, 0.050) for J-48, (0.9777, 0.025) for Random forest, (0.954, 0.052) for Multilayer perceptron respectively. This shows that the Random forest provides the highest TP rate i.e. 0.9777 while Naïve Bayes provides the lowest TP rate. We compare the entire TP rate and FP rate scored by all the algorithms. we found that all of these algorithms were better in predicting positive cases as TP rate in them is always greater than FP rate. Precision, Recall and F-measure of all the algorithms were comparative closer, the highest precision; Recall and F-measures scored are provided by Random forest. If we look at ROC curves, we found that the Multilayer Perceptron and Random forest are relatively close compared to other classifiers' curve value for Random forest classifier found 0.999 which is nearest to the Perfect Classification Point 1.

Based on above results and comparisons we found that the Random forest performs the highest Accuracy, TP rate, Precision, F-measure and ROC curve value. Naïve Bayes also score the fastest execution time as compare to other classifiers. Multilayer Perceptron takes the longest time to build the model.

Table 2. Confusion Matrix for performance analysis

	Actual		
Predicted		Present	Absent
	Present	True Positives (TP)	False Positives (FP)
	Absent	False Negatives (FN)	True Negatives (TN)

Table 3. Performance measurement of all classifiers

Algorithms	TP rate	FP rate	Precision	Recall	F-measure	ROC area	Accuracy (%)	Time(s)
Naïve Bayes	0.862	0.151	0.862	0.862	0.862	0.920	86.22	0.00
J48	0.951	0.050	0.951	0.951	0.951	0.975	95.08	0.02
Random Forest	0.977	0.025	0.977	0.977	0.977	0.999	97.70	0.17
Multilayer Perceptron	0.954	0.052	0.954	0.954	0.954	0.949	95.40	0.45

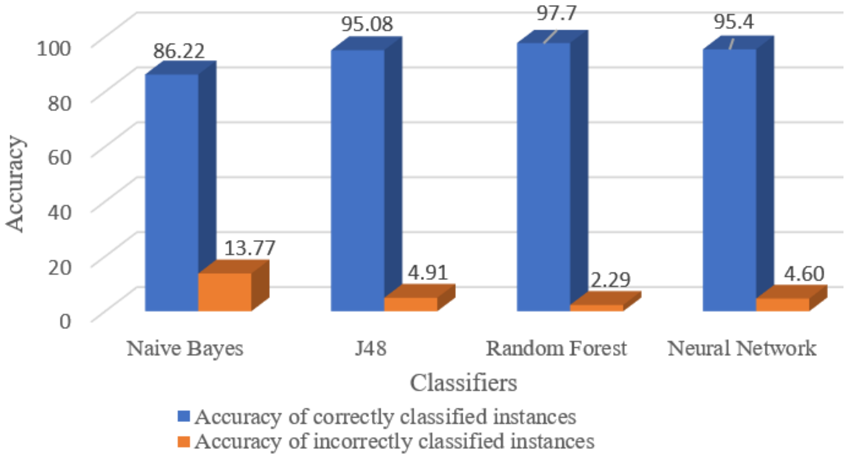

Fig. 5. Graphical representation of classifiers with their Accuracy

4.3 Prediction with Test Data Set

After performing several classification algorithm and ensemble method it has seen that Random Forest shows better result than other classifier. Then created a model with this ensemble method to make prediction on test data set. This research used this model to get the predicted class values of test dataset. After loading the model, we reevaluated the model with test data. This testing dataset contains 10 heart disease patients' instances. Test set was run by WEKA software. Then we apply different classifier algorithms for testing purpose. As a result, different algorithm shows different predictive result. After performing the four classifiers on test dataset, Random forest has shown the better result. Random forest provides 90%, Naïve Bayes provides 70%, Decision tree(j48) provides 80% and Neural network (Multilayer perceptron) provides 80% accuracy. Fig. 6. represents the prediction accuracy of different classifier algorithms with testing data set.

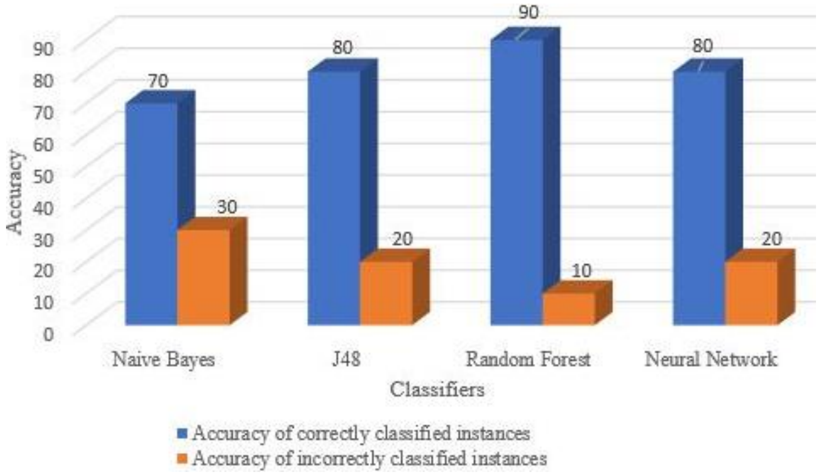

Fig. 6. Graphical representation of prediction accuracy using test data set.

5. CONCLUSION AND FUTURE WORK

5.1 Conclusion

Now a days, in medical diagnosis various data mining techniques are used for organizing numerous data. In this research, different data mining techniques were applied to improve the early detection of heart diseases. This research uses real dataset for training and testing purpose. Generally, four classifiers were utilized using WEKA machine learning tool to predict better accurate results and performance analysis. Comparisons of these algorithms are based on the performance factors of classification accuracy and execution time. Random Forest classifiers algorithm highest accuracy among all that is 97.7049%. is considered as the best classifier algorithm based on performance analysis in the diagnosis of heart disease patients. The experimental results show that we can produce short but accurate prediction list for the heart patients by applying the predictive models to the records of incoming new patients.

5.2 Future Work

The outcomes of this thesis may be used as assistant tool to help in making more consistent diagnosis of heart diseases. This study will also work to identify those patients which needed special attention. As a future work, we will use the research described here as a foundation for the development of effective prediction system to enhance medical care and reduce costs. We will significantly extend the functionality of the current research. It will also possible to plug in the prediction system into other systems such as access control security system to support the security fundamentals of healthcare systems and suggest useful treatment according to disease level.

CONSENT AND ETHICAL APPROVAL

As per international standard or university standard guideline participant consent and ethical approval has been collected and preserved by the authors.

ACKNOWLEDGEMENTS

I would like to thank all of my teachers and special thanks to my supervisor Prof. Dr. Md. Ashikur Rahman Khan and my co-supervisors and others for the continuous support. I would also like to express my gratitude to the Department of Information and Communication Engineering, Noakhali Science and Technology University.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Parvez Ahmad, Saqib Qamar, Syed Qasim Afser Rizvi. Techniques of data mining in healthcare. International Journal of Computer Applications. 2015;120(15):120-127.
2. Sujata Joshi, Mydhili K. Nair. Prediction of heart disease using classification based data mining techniques. Springer. 2015; 2(2):206-210.
3. Anbarasi. Enhanced prediction of heart disease with feature subset selection. International Journal of Engineering Science and Technology. 2010;2(2): 106=110.
4. AH Chen. HDPS: Heart Disease Prediction System. IEEE. 2011;3(2):6-10.
5. Chaitrali, Apte.. A data mining approach for prediction of heart disease using neural networks. International Journal of Computer Engineering & Technology. 2012;3(3)15-17.
6. Jyoti Soni. Intelligent and effective heart disease prediction system using weighted associative classifiers. International Journal on Computer Science and Engineering. 2011;3(6).
7. S. Pal. Data mining approach to detect heart dieses. International Journal of Advanced Computer Science and Information Technology. 2013;2(4).
8. A. Parveen. Prediction system for heart disease using naïve bayes. International Journal of Advanced Computer and Mathematical Sciences. 2012;2(3).
9. Rashedur F, Rahman M. Comparison of various classification techniques using different data mining tools for diabetes diagnosis. Journal of Software Engineering and Applications.
10. My Chau Tu. A comparative study of medical data classification methods based on decision tree and bagging algorithms. Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing. 2009;2(3):12-14.

11. Aqueel Ahmed, S. A. Hossain. Data mining techniques to find out heart diseases. International Journal of Innovative Technology and Exploring Engineering (IJITEE). 2012;1(1): 20-25.
ISSN: 2278-3075.
12. L. Moreno.,V. Fateh. Association Rules: Problems, solutions and new applications. Actas del III Taller Nacional de Minería de Datos y Aprendizaje. 2015;2(1):313-315.
13. A. Chaitrali S. Dangare. Improved study of heart disease prediction system using data mining classification techniques. International Journal of Computer Applications. 2014;47(10):0975 – 888.
14. N. Kumaravel, K. Sabullah., N. Nair. Automatic diagnoses of heart diseases using neural network. In Proceedings of the Fifteenth Biomedical Engineering Conference. 2016;3(2):319-322.
15. M. S. B. Venkatalakshmi. Heart disease diagnosis using predictive data mining. International Journal of Innovative Research in Science. Engineering and Technology. 2016;3(3).
16. D. Methaila. Early heart disease prediction using data mining techniques. CCSEIT, DMDB, ICBB, MoWiN, AIAP. 2016;3(2):53-59.
17. Benjamin Fredrick David H, Antony Belcy S. Heart disease prediction using data mining techniques. International Journal of Innovative Research in Science. Engineering and Technology; 2018.
18. UshaRani. A novel approach for imputation of missing attribute values for efficient mining of medical datasets-class based cluster approach. arXiv preprint arXiv. 2016;312-315.
19. Korting TS. C4. 5 algorithm and multivariate decision trees. Image Processing Division, National Institute for Space Research—INPE; 2018.
20. Vaghela C, Bhatt N, Mistry D. A survey on various classification techniques for clinical decision support system. International Journal of Computer Applications. 2015; 116(23).
21. Liaw A, Matthew Wiener. Classification and regression by random forest. R News. 2002;2(3):18-22.

Peer-review history:
The peer review history for this paper can be accessed here:
https://www.sdiarticle4.com/review-history/61760