Research Reports in Clinical Cardiology

Dovepress
open access to scientific and medical research

Open Access Full Text Article

METHODOLOGY

# Ensemble approach for developing a smart heart disease prediction system using classification algorithms



This article was published in the following Dove Press journal:
Research Reports in Clinical Cardiology

Mustafa Jan[1]
Akber A Awan[2]
Muhammad S Khalid[1]
Salman Nisar[1]

[1]Department of Industrial and Manufacturing Engineering, Pakistan Navy Engineering College, National University of Sciences and Technology, Karachi, Pakistan; [2]Department of Management and Information System, Pakistan Navy Engineering College, National University of Sciences and Technology, Karachi, Pakistan

**Abstract:** In health care informatics, the predictive modeling solution for cardiovascular risk estimation is extremely challenging. Thus, the attempt to clinically screen the medical databases and predictive modeling through soft computing tools is regarded as a valuable and economical option for medical practitioners. Therefore, the soft computing tools are today's need in health care application, which can perform data analysis and modeling, and they can assist the physician to make right and prompt clinical decisions. Extracting patterns that tie predictor's variables in a health science database is the topic of data mining. Existing data mining techniques are appropriate to model complex, dynamic processes. In this work, we propose Ensemble model approach for combining the predictive ability of a multiple classifiers' model for better prediction accuracy. In this study, ensemble learning combines five classifiers' model approaches, including support vector machine, artificial neural network, Naïve Bayesian, regression analysis, and random forest, to predict and diagnose the recurrence of cardiovascular disease. The cardiovascular data records of Cleveland and Hungarian were extracted from the UCI repository. Experimental results demonstrated that the ensemble model is a superior approach in terms of high predictive accuracy and reliability of diagnostics performance. In addition to this, this study also presents a smart heart disease prediction system as a valuable, economical and prompt predictive option having friendly graphical user interface, which is scalable and expandable.

**Keywords:** ensemble methods, smart heart disease prediction system, data mining model, classification techniques

## Introduction

In the modern world, cardiovascular diseases are declared as the highest-ranking disease in terms of causing million of deaths every year worldwide. Moreover, with the aging of the world's population and the demographic changes projected, it is expected that cardiovascular disease will continue to be a major health care challenge in the years to come. Estimates show that the number of cardiovascular deaths will increase both in developed and developing countries in the years to come.[1] It remains a difficulty, in particular with cardio vascular diseases (CVDs), that despite having treatment options available for disease, the burden remains unacceptably high because of an inability to match patients to treatments that are most appropriate for them individually.[2] In this regard, the soft computational-based tools can aid physicians in prompt and accurate decision-making. Thus, the data analytic application, a current and trending research, can aid physicians and doctors in making better and prompt decision by exploiting mining in data.[2]

Correspondence: Mustafa Jan
Department of Industrial and Manufacturing Engineering, Pakistan Navy Engineering College, National University of Science and Technology, Habib Ibrahim Rehmatullah Road, Karachi, Pakistan
Tel +92 313 260 6121
Email mattari.26@gmail.com

submit your manuscript | www.dovepress.com
Dovepress
http://dx.doi.org/10.2147/RRCC.S172035

The cardiovascular risk predictive modeling has attained prime importance in clinical research and patient care. The prompt, precise and accurate health risk assessment methods, to ensure that health implications are timely assessed, demand a commitment to the use of technological and computational-based health risk assessment models. Present clinical and pharmaceutical environments are data intensive. A large amount of patient data in the form of patient's description, clinical reports, laboratory tests, physician notes and hospital administrative data are being collected routinely.[3,4] With advances in recording and storage technologies, and with the increasing use of soft computational-based health record systems in hospitals, it is now possible to process larger volumes of data for deriving knowledge.

Hence, the pursuit for the fitting approaches to assess cardiovascular data and thus devising diagnostics and prognostics mechanisms to prevent CVD complications in individuals is a continuing quest. Due to one of the above-mentioned reasons, the developing predictive modeling solutions for cardiovascular risk estimation has become extremely challenging in health care informatics. In addition, handling various types of health care problems with desired accuracy and error-free results while dealing very highly nonlinear relationships between predictors and input variables demands soft computational intelligence-based automation in health care informatics. Soft intelligence plays a crucial role in medical diagnosis and intelligent decision making. The use of data mining (DM) model enables computational intelligence in diagnostics processes. DM is the computational intelligence-based process of extracting meaningful data from the set of large amount of data.

DM is a rapidly growing field in a wide range of health science applications. Appropriate DM-based classification techniques and smart heart disease prediction systems can lead toward quality health care in terms of accuracy and low economical health care services. The main motivation behind digitization of health data and utilization of soft computing tools is to lower the cost of health care and reduce the number of preventable errors. The conceptual methodology of soft computational-based DM framework for health informatics is presented in Figure 1.

Among various DM techniques, such as clustering, association rule classification and regression, the classification is one of the most important techniques used for categorization of data patterns. In DM, basically the classification-based machine learning algorithms are used to predict membership function for labeling CVD data instances. Classification is a data analysis technique that extracts labels describing important data classes. The classifier's model is represented as classification rules, decision trees or mathematical formulae, and it is termed as supervised learning. The model is used for classifying future or unknown objects. The classification algorithm predicts disease categorical class (eg, negative and positive) and build classifier model based on the training set. If the accuracy of the model is acceptable, the model can be applied to classify data tuples whose class labels are unknown. The classification comprises two basic steps of learning and classification. In learning, training data are analyzed by classification algorithm and classifier's model is built. In the classification phase, test data are utilized to estimate the accuracy of the classification model.

A healthy number of researchers have been applying various algorithms and techniques such as classification, clustering, regression analysis, artificial neural networks (ANNs), decision trees, genetic algorithm (GA), KNN methods, single DM model and hybrid and ensemble approaches to assist health care professionals with improved accuracy in the diagnosis of heart disease. In this study, the research quest of how the burden

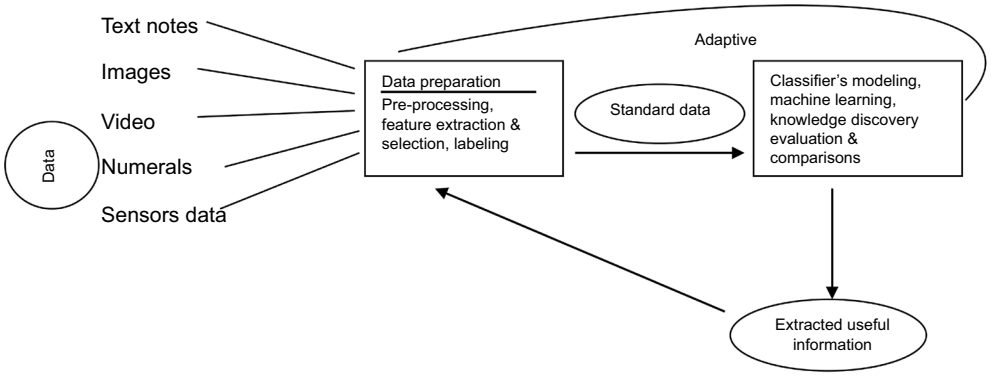


**Figure 1** Conceptual methodology for an adaptive DM model.

of coronary artery disease can be significantly reduced through soft computational methods is explored. The general problem statement of this study is to develop ensemble approach-based classifier's model that can be applied to CVD data sets to improve model prediction's outcomes. We propose a multi-classifier-based predictive model, an ensemble approach for combining the predictive ability of a multiple classifiers' model for better prediction accuracy and reliability. In addition to this, the study presents a prototype intelligent heart disease prediction system based on an ensemble approach while using different classifiers, namely, Naïve Bayesian, neural networks, support vector machine (SVM), random forest (RF) and logistics regression analysis. The proposed prediction system is user interface based, having the ability of scaling and expansion as per user's further requirement. The system has been implemented on the WEKA 3.8 version platform.

## Literature review

Various DM techniques introduced in recent years for prediction of heart diseases have been analyzed and reviewed in this section. The wide range of methodologies and techniques in DM is one of the greatest strengths that can be applied to various health science problems.[5] Researchers have applied various DM techniques such as association rule mining, clustering, classification to improve diagnosis of diseases with good accuracy and low probability of errors. Existing literature indicates that DM via classification plays an effective role in heart disease predictive mode as compared to clustering, association rule and regression. The review also highlights some studies using only single DM technique for diagnosis of heart disease, while many of the other research work have utilized ensemble/hybrid DM approaches in the thirst of better accuracy and reliability of model.

Lee et al proposed a statistical and classification-based approach to develop the multi-parametric linear and nonlinear type relationship for heart rate variability (HRV).[16] The experimental analyses using linear and nonlinear parameters of HRV were performed while applying Naïve Bayesian, association rules, decision tree and SVM classifiers. SVM performed better as compared to other classifiers. Tan and Teoh presented a hybrid approach using classification and GAs based on a wrapper approach. Based on the attribute subset represented by GA, the SVM classified the patterns into desired classes. UCI Machine Learning Repository set of data was utilized and analysis demonstrated the effectiveness of the GA–SVM hybrid approach. The effectiveness of the GA–SVM hybrid model in the multi-class domain was augmented by obtained average accuracy of 84.07%.

A neuro-fuzzy-based inference system for the prediction of heart disease was proposed as a new approach by Parthiban and Subramanian.[6] The neural network adaptive capabilities and the fuzzy logic qualitative approach were integrated with GA to predict heart disease. The performance of the co-active neuro-fuzzy model was evaluated by performance measures, and the results showed reasonable potential in predictive modeling of the heart disease. Anbarasi et al[7] presented enhanced prediction capability of heart disease using feature subset selection through GA. Originally, 13 attributes were involved in predictive analysis; the GA was applied to get optimal subset features. Decision tree, Naïve Bayesian and clustering were utilized for data categorization. WEKA tool was utilized for experimental analysis with 909 data instances. The prediction model was tested by the K-fold cross-validation method, and the prediction capability of the model was enhanced through GA.

Yan and Zheng[8] proposed a real-coded GA-based system for the diagnosis of heart diseases while applying critical clinical features sub-setting. The prediction system was modeled for the diagnosis of five major heart diseases, while utilizing 352 heart disease data instances with their corresponding diagnosis weights for supporting or denying the diagnosis of each heart disease. It provided a reasonably high accuracy for the practical heart disease prediction system. Austin et al developed the alternative classification schemes based on the machine learning literature and DM, which includes bootstrap aggregation bagging, RFs and boosting and SVMs. It was concluded from research that modern DM and ensemble methods offer advantages of high accuracy and efficiency.

Abdullah and Rajalaxmi[9] presented a DM model using RF classifier to improve accuracy and investigation related to coronary heart disease (CHD). Different CHD events including angina and acute myocardial infarction (AMI) and bypass graft surgery were investigated in the study. Experimental analysis showed that an ensemble approach of RF classification algorithm can be very effective in predictive mode for CHD. Lafta et al[10] of University of Southern Queensland developed an intelligent prediction system, based on the innovative time series prediction algorithm. Based on each patient's medical data, the system provided the decision support for medical practitioners. The system has the ability to improve the precision and performance through enhanced ability of the algorithm. Assari et al[11] presented the study for early diagnoses of heart disease by assessing related heart

disease risk factors in individuals After applying different DM techniques on selected data set, the SVM technique was found to be appealing with a highest accuracy of 84.33%. Finally, the model was developed based on extracted rules and main heart disease diagnosis indices. In employed classification techniques, Thal, Ca and Cp were marked as the most significant indices on average. This technique is expected to be applied on any application data set implemented in future.

The literature review unveils and indicates emerging and new areas of DM techniques involved in providing quality health care services at economical cost. It is evident from the abovementioned literature review that the amount of patient data in the form of diagnosis, medications and patient details is effectively dealt by DM techniques. The comparative evaluation of classification techniques in terms of finding the most feasible categorization method for making a cardiovascular prediction model with minimum probability of errors is today's research pursuit. The dimensionality of the heart database is high generally, and so, the selection of significant attributes for better diagnosis of heart disease is also a very challenging task for getting better diagnosis.[12,13] The reliability of the model for the prediction of heart disease with various risk factors is very much concerned. In this context, the ensemble approach and the hybrid DM model have revealed promising results in the diagnosis of heart disease.

Many ensemble techniques have been proposed over the past many years, resulting in new ways to address the issues of reliability and accuracy together. In the abovementioned context, various studies and prediction systems/methodologies have been introduced till now. So, in perspective of future and present works, a framework with improved mining of CVD data can be a better way in early prediction. Based on all these, the objective of the research is to exploit the efficacy of the hybridized model using an ensemble approach in heart disease prediction. The aim of this study is to develop a smart heart disease prediction system, which is applicable in terms of accuracy, reliability and practical utility.

## Materials and methodology

Methodology provides a framework for undertaking the proposed DM modeling. The methodology is a system comprising steps that transform raw data into recognized data patterns to extract knowledge for users. The DM methodology framework breaks down the mining process of cardiovascular data into phases as shown in Figure 2. It shows an iterative DM process for implementing machine learning

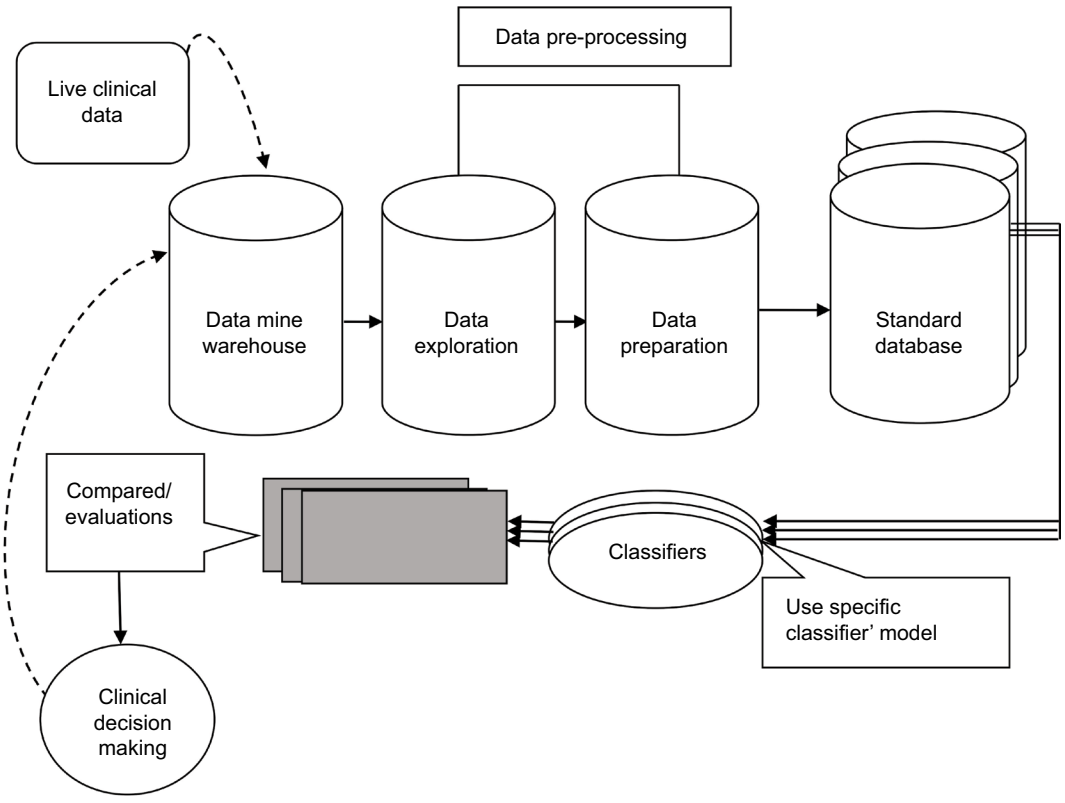


**Figure 2** Methodology for mining heart disease data.

methods on the cardiovascular data set taken for application. The proposed methodology includes steps, referred to as the preprocessing stage where the exhaustive exploration of the data is carried out. It will account for dealing with missing values, balancing data and normalizing attributes depending on algorithms used. After pre-processing of data is performed, predictive modeling of the data is carried out using classification models and ensemble approach. Finally, prescriptive modeling is undertaken, where the predictive model is evaluated in terms of performance and accuracy using various performance metrics. Accordingly, the modifications are either made to the data preparation stage or the predictive modeling strategy is amended.

## Data set for experiments

The data set for this research was taken from UCI data repository.[14] Data accessed from the UCI Machine Learning Repository is freely available. From the abovementioned DM warehouse, the Cleveland and Hungarian data were selected. This database contains 76 attributes, and after neglecting redundant and irrelevant attributes, 14 attributes were selected. Table 1 represents the list of 14 attributes and their brief description. In particular, the Cleveland and Hungarian databases have been used by many researchers and found to be suitable for developing a mining model, because of lesser missing values and outliers. The data were cleaned and preprocessed before they were submitted to the proposed algorithm for training and testing. Hence, 590 are the valid instances for supervised machine-learning model building. Table 1 shows the selected important risk factors from databases and their corresponding values.

Attribute selection or feature sub-setting technique was applied for further reduction of data to make patterns easier and understandable, but found negligible effects on performance measures of the model engaged in this study. In view of the above, all the 13 attributes were taken into the consideration for developing a classifier's model and obtaining CVD predictive outcome. The classification techniques engaged in this study are Naïve Bayesian, neural network, SVM, decision tree-based RF algorithm and regression analysis. The ensemble DM approach was utilized for evaluating the classification algorithms engaged. The WEKA DM tool was used to build the model. In these experiments, 10-fold cross-validations were employed to partition the data set into training and test sets; this fulfills the requirement of model training and testing purpose. As a result, the accuracy rate obtained from this experiment was above 93% for all classifier's models applied in the study.

## Classifiers used for experiments

### Naïve Bayesian

It is a probabilistic classifier based on Baye's theorem specified by the prior probabilities of its root nodes. The Bayes theorem is given in Equation 1 and normalization constant is given in Equation 2. It proves to be an optimal algorithm in terms of minimization of generalized error. It can handle statistical-based machine learning for feature vectors $\underline{y} = [y_1, y_2, \ldots, y_l]^T$ and assign the label for feature vector based on maximal probable among available classes $\{XX_1, X_2, \ldots, X_M\}$. It means that feature "y" belongs to $X_i$class, when posterior probability $P(X_i|\underline{y})$ is maximum ie $\underline{y} \rightarrow X_i : P(X_i|\underline{y})$ Max. The Bayesian classification problem may be formulated by a-posterior probabilities that assign the class label $\omega_i$ to sample X such that $P(X_i|\underline{y})$ is maximal. The Bayesian classification problem may be formulated by a-posterior probabilities that assign the class label $\omega_i$ to sample $X$ such that $P(X_i|\underline{y})$ is maximal.

Likelihood Prior

$$P(X_i|\underline{y}) = \frac{p(\underline{y}|X_i)P(X_i)}{p(\underline{y})} \quad (1)$$

Normalization Constant

$$p(\underline{y}) = \sum_{i=1}^{2} p(\underline{y}|X_i)P(X_i) \quad (2)$$

Application of Bayes' rule with the mutual exclusivity in diseases and the conditional independence in findings is known as the Naïve Bayesian Approach. It is a probabilistic classifier based on Bayes' theorem with strong independence assumptions between the features. Naïve Bayesian classifier despite its simplicity, it surprisingly performs well and often outperforms in complex classification. Simple Naïve Bayesian can be implemented by plugging in the following main Bayes' formula:

$$P(X_1, X_2, \cdots, X_n \mid Y) = P(X_1 \mid Y)\, P(X_2 \mid Y) \cdots P(X_n \mid Y) \quad (3)$$

The abovementioned Naïve Bayesian network produces a mathematical model, which is used for modeling the complicated relations of random variables of disease attributes and decision outcome. The algorithm uses the formula to calculate conditional probability with respect to disease condition attributes value and decision attribute value. Based on prior knowledge, the algorithm classifies the decision attribute into labels assigned, and hence the conditional support is computed for each variable attribute.[15]

**Table 1** Factors related to heart disease patient with its type and description

| S no | Input variables | Description | Options |
|---|---|---|---|
| 1 | Age | Age in years | Continuous value |
| 2 | Sex | 1 = male, 0= female | Male, female |
| 3 | Cp | Chest pain type | Chest pain type. Values from 1 to 4.<br>1: typical angina.<br>2: atypical angina.<br>3: non-anginal pain.<br>4: asymptomatic. |
| 4 | Trestbps (blood pressure) | Resting blood pressure in mmHg | Continuous value in mmHg |
| 5 | Chol (cholesterol) | Serum cholesterol in mm/dL | Continuous value in mm/dL |
| 6 | Fbs (fasting blood sugar ) | Fasting blood sugar in mg/dL | Fasting blood sugar attributes value "1" for greater than 120 mg/dL, else the attribute value is 0 (false).<br>Value 1 = true.<br>Value 0 = false. |
| 7 | Restecg (ECG) | Electrocardiographic results (ECG result) | Resting electrocardiographic results value ranging from 0 to 2.<br>0: normal.<br>1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of >0.05 mV).<br>2: showing probable or definite left ventricular hypertrophy |
| 8 | Thalach (heart rate) | Maximum heart rate | Continuous value categorized into normal and abnormal |
| 9 | Exang | Exercise Induced angina | Exercise induced angina. Values from 0 to 1.<br>Value 1 = yes.<br>Value 0 = no. |
| 10 | Old peak | ST depression induced by exercise relative to rest | Continuous values |
| 11 | Slope | Slope of the peak exercise ST segment | Measure of slope for peak exercise. Values can be 1, 2, or 3.<br>Value 1: up sloping.<br>Value 2: flat.<br>Value 3: down sloping. |
| 12 | Ca | Number of major vessels colored by fluoroscopy | Number of major vessels from 0 to 3 |
| 13 | Thal | Heart rate of patient | Represents heart rate of the patient. It can take values 3, 6, or 7.<br>Value 3 = normal.<br>Value 6 = fixed defect.<br>Value 7 = reversible defect. |
| 14 | Class | Class labels (predicted outcome) | Contains a numeric value between 0 and 1. Each value represents heart disease or absence of disease.<br>Value 0: absence of heart disease.<br>Value 1: presence of heart disease. |

## Neural networks

The network comprises a large number of extremely interconnected processing elements called neurons. For health science application, the ANN is configured for data classification. In practical applications, neural network generates highly accurate results and has the ability to derive meaning from complicated data and detect trends that are too complex to be noticed by other computer techniques. It has adaptive learning feature with real-time operation. For adaptive feature, it has to be configured and trained by feeding learning patterns with adjustments of weights according to learning rule. The multilayer feed-forward neural network can be trained by a

back propagation algorithm. In the algorithm, each node's weight is increased or decreased slightly as the error varies. The neural network simple architecture is shown in Figure 3. The weights and the transfer function (activation function) drive the behavior of an ANN. The sigmoid transfer function has significant resemblance to real neurons as compared to linear units, but it must be considered as rough approximations of real.

The feed forward neural network can be trained for heart diseases using a back propagation learning algorithm with momentum. The clinical data are made input to start off the ANN model with algorithm running inside as a core. After

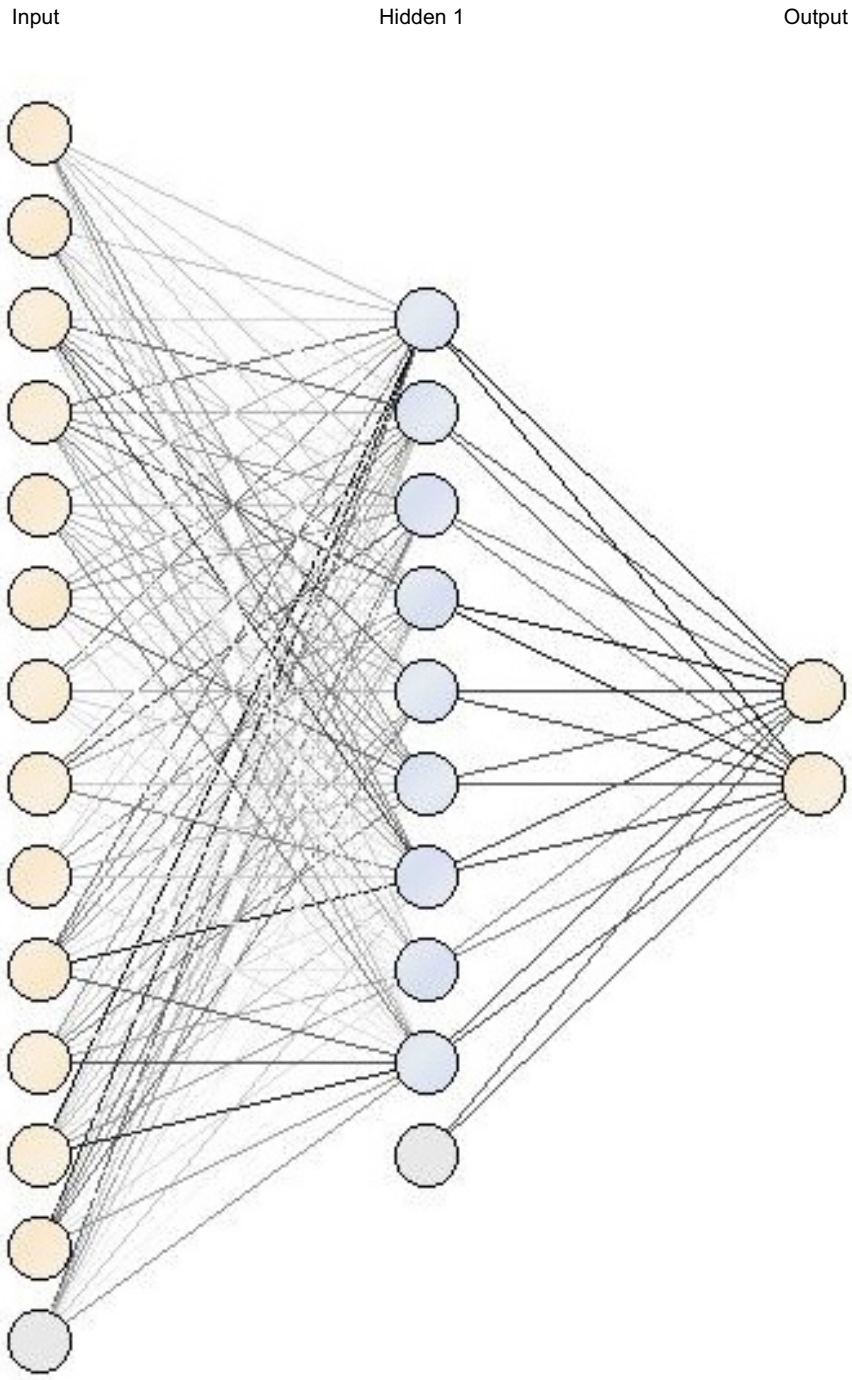


**Figure 3** Neural network model architecture.

the training model, the computational steps of a neural network algorithm start the classification of clinical data into two equal parts randomly. One is used for testing, and the other is used for training.

### SVMs

SVM is a rough realization of structure-based minimization of risk and a categorization of linear/non-linear data. SVMs maximize the margin around the separating hyperplane. Each subset of training samples (support vectors) specifies decision boundary functions. SVM comprises three steps, the support vector creation, formation of maximal distance between points found and perpendicular decision boundary. This is the basic type of SVM called linear SVM. The linear SVM hyperplane's concept is shown in Figure 4.

The maximal margin linear classifier is an inapplicable approach for many practical problems. For practical application, where non-linear separable data set is to be separated by hyperplane, the non-linear data are to be mapped to another feature space through Kernel functions. In original input space, the data points are separated by hyperplane, and Kernel functions map the non-linear training samples to high dimensional space. Then, an algorithm search for the best hyperplane to separate the transformed data into two different classes is carried out. The margin of the hyperplane is maximized for classification purpose while minimizing the classification errors. The algorithm predicts the risk of heart disease in multi-dimensional hyperplane and categorizes the data into different labels optimally by creating the margin between data clusters.

### RF

RF, one of the most accurate machine learning algorithms, is a decision tree-based ensemble classifier approach which contains flowchart like tree structure. RF is a combination of tree-structured classifiers {h(x, n)}, where for "x" data input and "n" are distributed random trees for the classification of data. Each one of the decision tree in random tree-structured forest, cast a vote that indicates the decision about class of data. RF uses the Gini-index for determining the final class in each tree. This algorithm chooses optimal attributes from "M" total number of input attributes at random for each tree. With this selected attribute, the best possible split is created using the Gini index to develop a decision tree model. This is an iterative process for each of the branches until the terminating nodes are too small to split further. For data set X having "n" classes, Gini-index, Gini(X) can be defined by:

$$Gini(X) = \sum_{i=1}^{n} (Rj)^2 \qquad (4)$$

where "Rj" is the relative frequency of class j in data set "X". In RF, the split at which the Gini index is lowest is chosen at the split value.

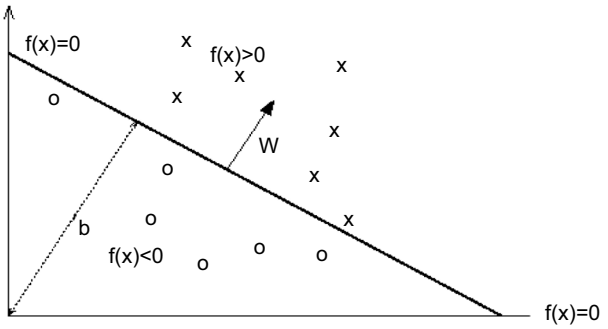


**Figure 4** SVM hyper plane concept.
**Abbreviation:** SVM, support vector machine.

### Classification via regression analysis

There are various different types of regressions in statistics, but the basic idea is that a model is created that maps values from predictors in such a way that the lowest error occurs in making a prediction. The relationship in the form of single equation between dependent variables and other independent variables is evaluated or attempted by regression analysis as statistical modeling. The term regression is related to analyses of independent variable and dependent variable to develop a model having relationship between predictors and outcome variables. Regression can be generalized by linear regression and mainly used for estimating multi-class-dependent variables. The linear relationship between predictor and outcome takes the form of an equation for a line that best indicates a series of data. In terms of health care application using regression, it is used to fit a curve to medical data in which the dependent variable (disease prediction) is binary or discrete.

## Ensemble DM approach

In order to have more reliable and accurate prediction results, ensemble method is a well-proven approach practiced in research for attaining highly accurate classification of data by hybridizing different classifiers. The improved prediction performance is a well-known in-built feature of ensemble methodology. This study proposes a weighted vote-based classifier ensemble technique, overcoming the limitations

of conventional DM techniques by employing the ensemble of five heterogeneous classifiers: Naive Bayesian, neural network, RF, SVMs, and classification via regression analysis. We have used two benchmark heart disease data sets taken from UCI repository namely Cleveland and Hungarian.

Basic idea in the ensemble method is to train a set of classifiers and to test each classifier for measuring performance. Given a data set $D=\{x_1,x_2,\ldots,x_n\}$ and their corresponding labels $L=\{l_1,l_2,\ldots,l_n\}$, an ensemble approach computation is as follows. A selected set of classifiers $\{f_1,f_2,\ldots,f_k\}$, each of which maps data to a class label: $f_j(x)=$ li. A combination of classifiers f(x) through an ensemble approach is adopted which minimizes generalization error: in terms of minimized probability of error occurrence $f(x)= w_1f_1(x)+ w_2f_2(x)+\ldots+ w_kf_k(x)$.

A classifier model Mi is trained by data set Di, and testing is performed for data samples X. The baseline accuracy and prediction probability for positivity of disease occurrence are set as required for each classifier's vote to be valid for reliable labeled classes. With respect to baseline accuracy and prediction probability taken for each classifier, vote will be given to classifier having prediction probability and overall classifier's accuracy above the baseline, which is set as 93% in this study. In this study, each classifier Mi returns its class "X" (true positive result [TPR]) prediction probability, the each classifier "M" vote counted as "1", if having more than 93% TPR and overall classifier's accuracy. Consequently, the assigned class X by the most votes of classifiers is taken as reliable and valid results for positivity of heart disease (TPR). For perfect analysis of heart disease, the output of each algorithm with accuracy for TPR more than 93% is taken as one vote. The vote of each classifier's model is added and compared with a threshold value of 4 votes at least for reliable TPR. If the addition of that classifier's vote makes figure greater than or equal to 4.0, then the TPR is validated, and treatment is recommended accordingly. The Ensemble model augmented by classifiers' votes is shown in Figure 5.

## Experimental analysis and results

In this study, the WEKA DM tool was chosen for experimental analysis of CVD data sets and performance evaluation of each classifier applied for categorization of data sets. The Cleveland and Hungarian data set used for experimental analysis is available in UCI Machine Learning Repository. The performance measures of model have been presented at model comparison's table shown in Table 2. The WEKA mining experiments were done while using all 14 attributes for calculating performance measures of each classifier's model. The visualization of CVD patient attributes is given by the WEKA visualization tool as shown in Figure 6. In Figure 6, we can see the complete CVD patient data and can just visualize a particular attribute vs number of instances. The two colors indicate the instance's class (negative/positive CVD prediction). The *x*-axis shows the number of instances, and the *y*-axis shows the value for particular attribute.

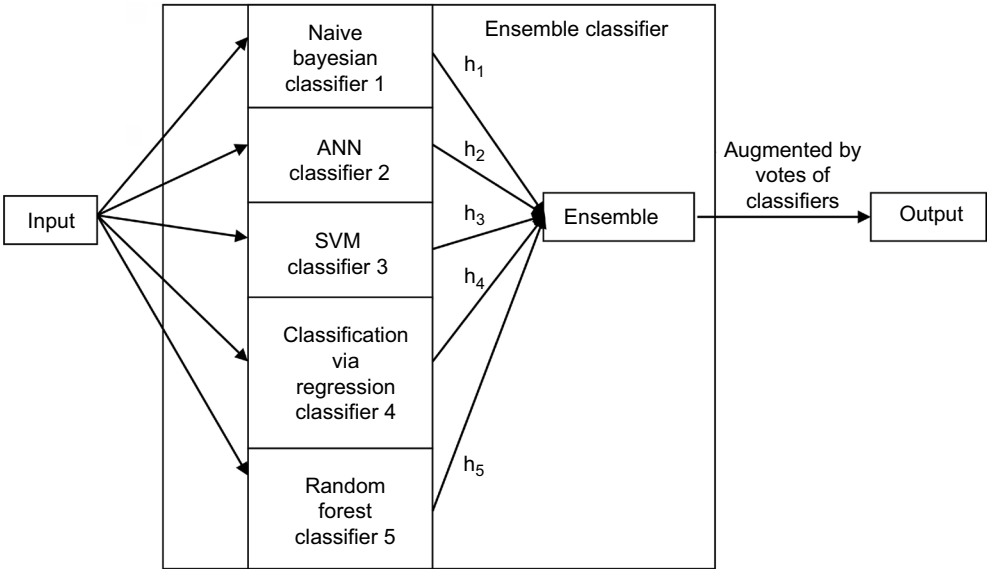


**Figure 5** Ensemble vote model augmented by votes of classifiers.
**Abbreviations:** ANN, artificial neural network; SVM, support vector machine.

**Table 2** Summary of implemented algorithms' performance

| Algorithms' performance measures | Naïve Bayesian | ANN | SVM | Random forest | Logistic regression |
|---|---|---|---|---|---|
| Accuracy (%) | 93.223 | 94.9153 | 98.135 | 98.136 | 93.22 |
| TP rate | 0.932 | 0.949 | 0.981 | 0.981 | 0.932 |
| FP rate | 0.071 | 0.037 | 0.023 | 0.023 | 0.071 |
| Precision | 0.932 | 0.955 | 0.981 | 0.981 | 0.932 |
| F-measure | 0.932 | 0.949 | 0.983 | 0.981 | 0.932 |
| ROC curve | 0.982 | 0.984 | 0.903 | 0.984 | 0.932 |
| Kappa statistics | 0.8612 | 0.8975 | 0.9613 | 0.9613 | 0.8612 |
| RMS error | 0.2242 | 0.2226 | 0.1373 | 0.1373 | 0.24 |

**Abbreviations:** ANN, Artificial Neural Network; FP, false positive; ROC, receiver operating characteristic; RMS, root mean square; SVM, support vector machine; TP, true positive.

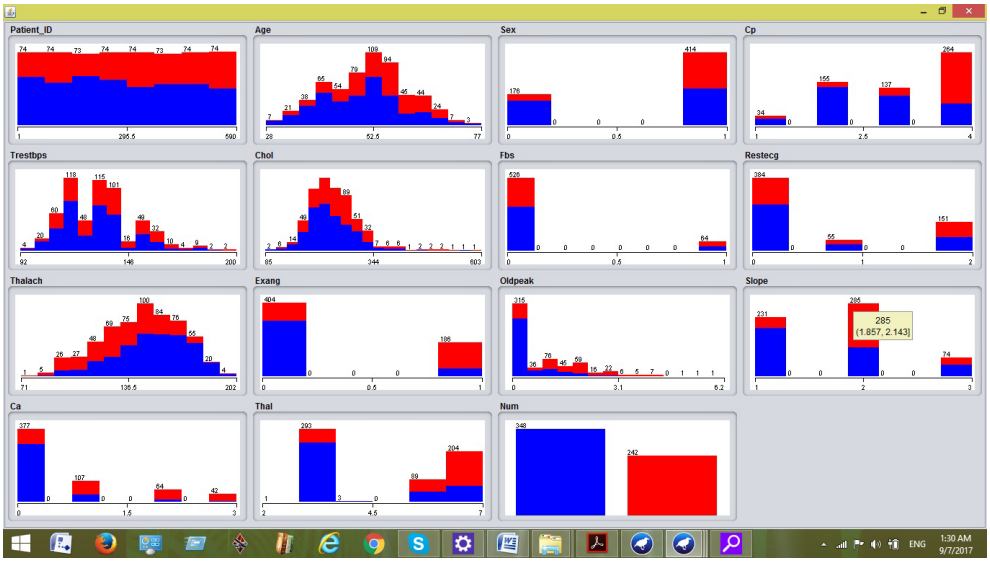


**Figure 6** Visualization of the heart disease patient's attributes vs number of data instances.
**Notes:** Ca, Number of major vessels colored by fluoroscopy; Chol, serum cholesterol in mm/dL; Cp, chest pain type; Exang, exercise-induced angina; Fbs, fasting blood sugar in mg/dL; Oldpeak, ST depression induced by exercise relative to rest; Restecg, Electrocardiographic results (ECG result); Thal, heart rate of patient; Thalach, maximum heart rate; Trestbps, resting blood pressure in mmHg.

Mainly five classification algorithms were used for developing a DM model. For sampling the training and testing data set, 10-fold cross-validation was applied. The performance and the accuracy of each experiment are evaluated through performance measures such as true positive rate, precision, F-measure, receiver operating characteristic (ROC) area, Kappa statistics and root mean square (RMS) error. The same measures have been used for comparative analysis of implemented algorithms. After the experiments, the next step is to compare algorithms used in these experiments for highlighting the best one in terms of disease prediction probability and classifier's accuracy. Having a look at the results, it becomes apparent that the goal to produce an ensemble classifier for early diagnostic screening with required level of accuracy is successful. A correlation between accuracy and the amount of attributes used in the creation of the classifier was found. In general, more attributes give greater accuracy as visualized by results. With respect to ROC area as performance measure, an optimal/perfect classifier will score 1 on this test, so this does make our results looking more promising with results more than 0.9 ROC for all classifiers. The comparative performance summary of implemented algorithms is given in Table 2.

In general, the results of all the implemented algorithms are much better by all algorithms with specially the RF ensemble algorithm leading in accuracy and prediction probability. The accuracy of implemented algorithms on the given heart disease data set is presented in Table 2, and the

lowest accuracy is 93.22% for regression analysis and the highest accuracy is 98.17% for the RF algorithm. Based on the abovementioned results and comparisons with respect to the selected performance measures, the RF ensemble algorithm and SVM performed well and further hybridization through voting of each algorithm with more than 93% prediction probability has enhanced reliability of the prediction system. More emphasis is given to select the algorithms having high true positive rate, as being the core measure for early diagnosis of heart disease.

## Working of the smart heart disease prediction system

The working of the proposed model as a smart heart disease prediction system is shown in Figure 7, and its graphical user interface (GUI) is shown in Figure 8. It consists of training/testing data set and user input as the application data set. WEKA DM tool's environment has been used to implement the heart disease prediction system. The components used in the smart prediction system are CVD data instances from Cleveland and Hungarian sources, five different classification algorithm and performance measures' methods for evaluation. These CVD data consist of the two class labels, ie, true positive/true negative for heart disease prediction and its corresponding value.

The training data set in the Attribute-Relation File Format (ARFF) comprising 14 attributes including class label has been utilized as input to WEKA. The heart disease prediction system accepts input from the user through a GUI, and its database will be maintained by the SQL server R2 tool at the back end. The user input obtained from GUI as excel is converted into ARFF through the WEKA tool. When a new data instance related to CVD patients comes in the input, the prediction system based on ensemble classifiers' model classifies the new instance. This is the real-time testing of the proposed model, where we can check the accuracy and applicability of the prediction system. After the prediction, we will get the class labels for new instance (CVD patient data). As a generalized prediction system, we can use this model for predictive analysis of different data sets having slight modifications in the predictive model.

## Conclusion and future work

An ensemble approach-based prototype smart heart disease prediction model has been proposed as a system while utilizing RF trees, SVM Naïve Bayesian, neural networks

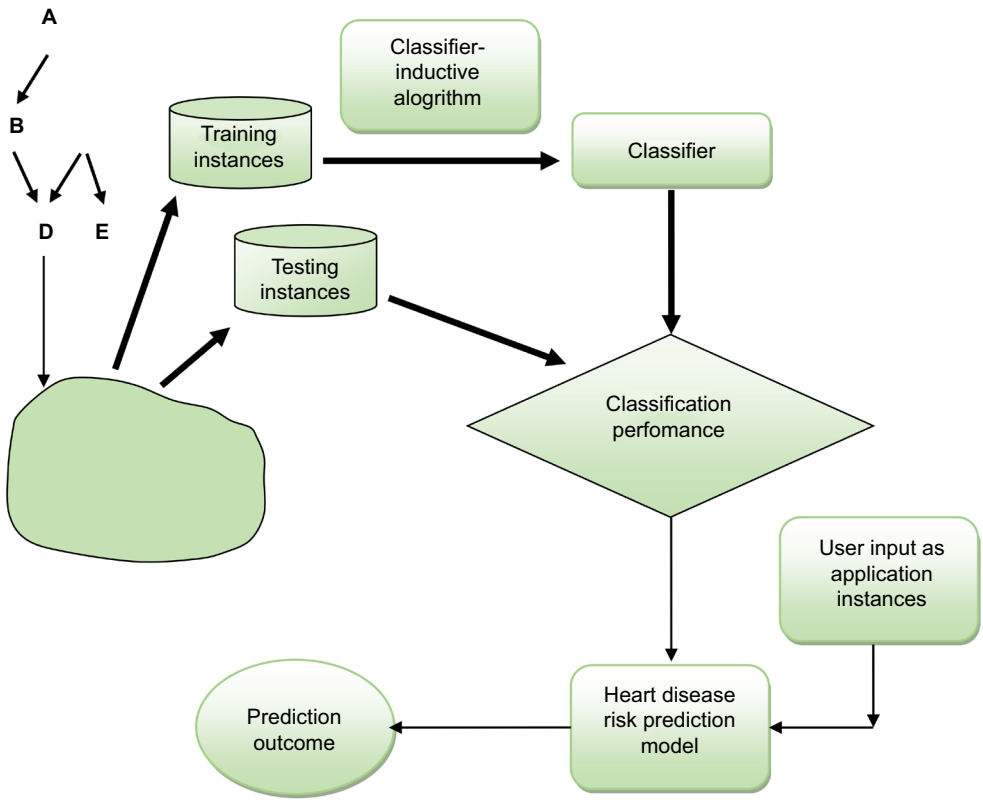


**Figure 7** Working model-proposed smart heart disease prediction system.

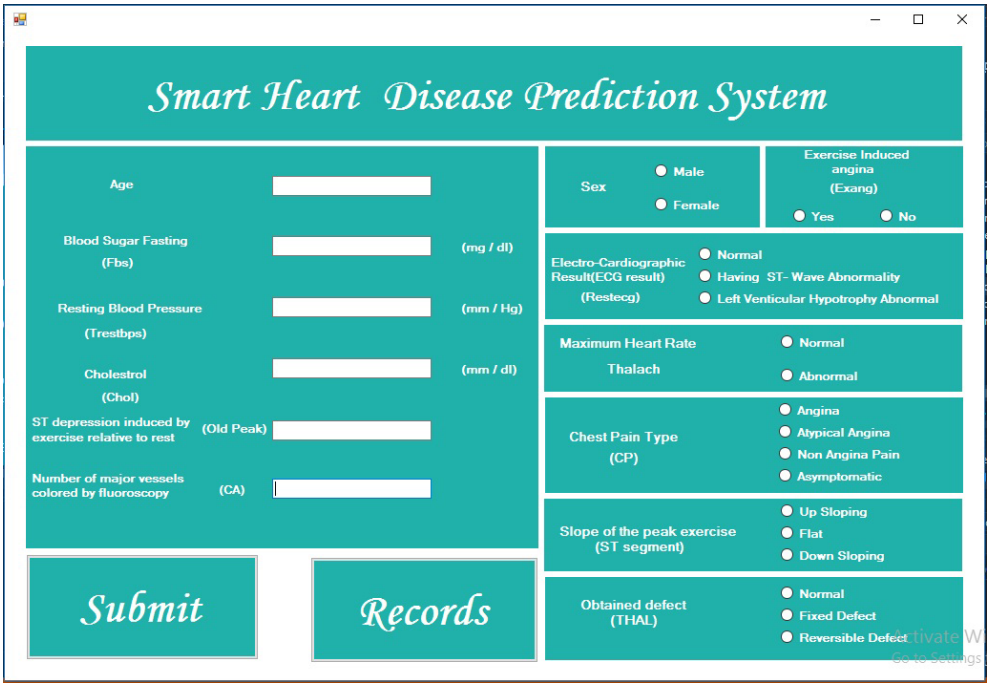


**Figure 8** GUI for smart heart disease prediction system's input.
**Abbreviation:** GUI, graphical user interface.

and logistic regression analysis-based classifiers. The proposed system is GUI-based, user-friendly, scalable, reliable and expandable system, which has been implemented on the WEKA platform. The proposed working model can also help in reducing treatment costs by providing Initial diagnostics in time. The model can also serve the purpose of training tool for medical students and will be a soft diagnostic tool available for physician and cardiologist. General physicians can utilize this tool for initial diagnosis of cardiopatients. There are many possible improvements that could be explored to improve the scalability and accuracy of this prediction system. As we have developed a generalized system, in future we can use this system for the analysis of different data sets. The performance of the health's diagnosis can be improved significantly by handling numerous class labels in the prediction process, and it can be another positive direction of research. In DM warehouse, generally, the dimensionality of the heart database is high, so identification and selection of significant attributes for better diagnosis of heart disease are very challenging tasks for future research.

## Acknowledgments

The authors thank Allah the most beneficent and most merciful.

## Disclosure

The authors report no conflicts of interest in this work.

## References

1. Reddy KS. Cardiovascular disease in non-Western countries. *N Engl J Med*. 2004;350(24):2438–2440.
2. Myerburg RJ. Sudden cardiac death: exploring the limits of our knowledge. *J Cardiovasc Electrophysiol*. 2001;12(3):369–381.
3. Ordonez C, Omiecinski E. *Mining Constrained Association Rules to Predict Heart Disease, IEEE*. San Jose, CA, USA: Published in International Conference on Data Mining (ICDM); 2001:433–440.
4. Sidney C, Smith MD. Current and future directions of cardiovascular risk prediction,. *Am J Cardiol*. 2006;97(2A):28A–32A.
5. Labib NM. Data Mining for Cancer Management in Egypt Case Study: Childhood Acute Lymphoblastic Leukemia, *Int J Med Health Pharm Biomed Eng*. 2007;1(8).
6. Parthiban L, Subramanian R. Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm. *International Journal of Biological and Medical Sciences*. 2007;1(5):278–281. Available from: https://waset.org/publications/14149/intelligent-heart-disease-prediction-system-using-canfis-and-genetic-algorithm. Accessed November 01, 2018.
7. Anbarasi M, Anupriya E, Iyenga NCHSN. Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm. *Int J Eng Sci Technol*. 2010;2(10):5370–5376.
8. Yan H, Zheng J. Selecting critical clinical features for heart diseases diagnosis with a real-coded genetic algorithm. *J Appl Soft Comput*. 2008;8:1105–1111.
9. Abdullah AS, Rajalaxmi RR. 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) Proceedings published in International Journal of Computer Applications® (IJCA)*; 2012.

10. Lafta R, YanLi, Tseng VS. An Intelligent Recommender System based on Short Term Risk Prediction for Heart Disease patients. *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*. Singapore: IEEE; 2015.
11. Assari R. Parham Azimi, Mohammad RT. Heart Disease Diagnosis Using Data Mining Techniques. *Int J Econ Manag Sci*. 2017;6:3.
12. Zriqat Isra'a Ahmed, Altamimi AM. Mohammad Azzeh, A Comparative Study for Predicting Heart Diseases Using Data Mining Classification Methods. *Int J Comput Sci Inform Secu.* 2016;14(12).
13. Shrivastava A, Tomar SS. A hybrid framework for heart disease prediction: review and analysis. *Int J Adv Technol Eng Explor*. 2016;3(15):21–27.
14. UCI Machine Learning Repository. Available from: https://archive.ics.uci.edu/ml/index.php. Accessed November 01, 2018.
15. Dangare CS, Apte SS. Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques. *Int J Comput Appl*. 2012;47(10):44–48.
16. Lee HG, Noh KY, Ryu KH. Mining Biosignal Data: Coronary Artery Disease Diagnosis using Linear and Nonlinear Features of HRV. In: Washio T, Zhou ZH, Huang JZ, et al, editors. *Emerging Technologies in Knowledge Discovery and Data Mining. PAKDD 2007, Lecture Notes in Computer Science*. Vol 4819. Germany: Springer, Berlin, Heidelberg; 2007:218–228.