

BagMOOV: A novel ensemble for heart disease prediction bootstrap aggregation with multi-objective optimized voting

Saba Bashir · Usman Qamar · Farhan Hassan Khan

Received: 21 August 2014 / Accepted: 24 February 2015
© Australasian College of Physical Scientists and Engineers in Medicine 2015

Abstract Conventional clinical decision support systems are based on individual classifiers or simple combination of these classifiers which tend to show moderate performance. This research paper presents a novel classifier ensemble framework based on enhanced bagging approach with multi-objective weighted voting scheme for prediction and analysis of heart disease. The proposed model overcomes the limitations of conventional performance by utilizing an ensemble of five heterogeneous classifiers: Naïve Bayes, linear regression, quadratic discriminant analysis, instance based learner and support vector machines. Five different datasets are used for experimentation, evaluation and validation. The datasets are obtained from publicly available data repositories. Effectiveness of the proposed ensemble is investigated by comparison of results with several classifiers. Prediction results of the proposed ensemble model are assessed by ten fold cross validation and ANOVA statistics. The experimental evaluation shows that the proposed framework deals with all type of attributes and achieved high diagnosis accuracy of 84.16 %, 93.29 % sensitivity, 96.70 % specificity, and 82.15 % f-measure. The f-ratio higher than f-critical and p value less than 0.05 for 95 % confidence interval indicate that the results are extremely statistically significant for most of the datasets.

Keywords Ensemble classifier · Weighted voting · Heart disease · Multi-objective optimization · Prediction · Data mining

Introduction

Computational intelligence has started playing a vital role in medical diagnosis and intelligent decision making. Medical diagnosis procedures can be categorized using intelligent computational classification tasks. Data mining is a process of analyzing and identifying previously unknown and hidden patterns, relationships and knowledge from large datasets that was not possible with traditional techniques [1]. According to recent research, data mining techniques are extremely helpful in the diagnosis of several diseases such as cancer [2], stroke [3], diabetes [4] and heart disease [5].

Several classification techniques are used for heart disease prediction; such as Naïve Bayes, linear regression, neural networks, support vector machine, and kernel density, which results in different levels of precision, recall and accuracy [5–7]. An ensemble approach shows promising results as compared to a single technique [8]; therefore, researchers have been investigating ensemble based data mining approaches for heart disease prediction showing fruitful results.

Figure 1 presents a generic ensemble framework for prediction and evaluation of a disease. It is composed of training set, test set, model builder, ensemble model, prediction and evaluation. This ensemble framework can also be applied to heart disease data. Heart disease datasets, with known class labels, are partitioned into training and test sets. The training set is used to train the classifiers and fed into a model builder which consists of individual

S. Bashir · U. Qamar · F. H. Khan (✉)
Computer Engineering Department, College of Electrical and
Mechanical Engineering, National University of Sciences and
Technology (NUST), Islamabad, Pakistan
e-mail: farhan.hassan@ceme.nust.edu.pk

S. Bashir
e-mail: saba.bashir@ceme.nust.edu.pk

U. Qamar
e-mail: usmanq@ceme.nust.edu.pk

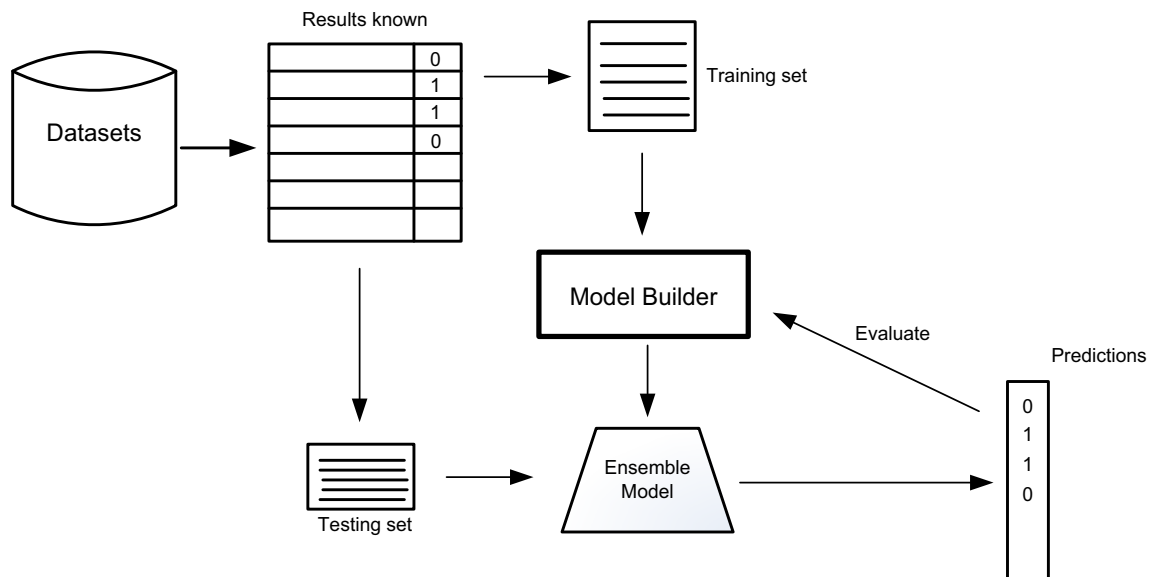


Fig. 1 Generic data mining ensemble framework for prediction and evaluation of disease

classifiers. These trained classifiers are then combined using an **ensemble technique**. This ensemble model is then executed on the test set and prediction is computed. Finally, the performance is evaluated by comparing the predicted results with other classifiers and ensembles.

This research paper presents a framework for intelligent heart disease decision support system (DSS) using a novel ensemble approach based on **heterogeneous machine learning techniques**. The proposed framework results in reliable performance for heart disease prediction as compared to other classifiers. It integrates multiple heterogeneous classifiers using enhanced bagging with multi-objective optimized weighted voting scheme.

Research contributions

Healthcare industry is continuously making an effort to reduce medical errors and provide patient care and safety. Adverse reactions can occur if a disease is not diagnosed accurately. A DSS can assist health professionals for decision making tasks such as diagnosis of heart disease from patient's data. However, in medical applications decision quality is of crucial importance. Therefore, high accuracy is of prime importance in heart disease classification and prediction. In this research we present a complete DSS for accurate diagnosis of heart disease. It can be used by health professionals for diagnosis of heart disease from patient's data. Whilst human decision-making performance can be suboptimal and deteriorate as the complexity of the problem increases, the proposed framework can help healthcare professionals to make correct decisions.

The main contributions of the proposed research are summarized as follows:

- A novel ensemble approach is proposed which uses Bootstrap Aggregation (Bagging) with multi-objective optimized weighted vote for diagnosis of heart disease.
- The proposed framework overcomes the limitations of conventional performance by utilizing an ensemble of five heterogeneous classifiers: Naïve Bayes, linear regression, quadratic discriminant analysis, instance based learner and support vector machines. However, the proposed framework can be implemented using any set of classifiers which may be homogenous, heterogeneous or a combination of both.
- We compare the proposed ensemble approach with existing classifiers and ensembles to prove the superiority of our technique.

The rest of the paper is organized as follows: "**Literature review**" section is related to literature review. The proposed approach is defined in "**The proposed framework**" section and "**Dataset description**" section describes dataset information. "**Results and discussion**" section presents the results and discussion sections. Finally, "**Real-time implementation of the proposed framework**" section summarizes the work which has been done.

Literature review

There are various machine learning techniques that are widely accepted for heart disease analysis and prediction. A clinical decision support framework involves many

machine learning algorithms. Table 1 shows comparison of accuracy, sensitivity and specificity for different techniques used for heart disease diagnosis. It is observed from Table 1 that most of the work is focused on applying a single classification technique for heart disease analysis and prediction. Pattekari and Parveen [9] used Naïve Bayes for heart disease prediction. The proposed technique uses a single classifier and works only on categorical data. The technique can be improved by using other data mining techniques such as time series, clustering and association mining whereas the results can be improved by considering other data types as well. Peter and Somasundaram [10] proposed a cardiovascular disease risk prediction framework by using data mining and pattern recognition techniques. Analysis of results indicates that Naïve Bayes has better accuracy as compared to other techniques. The proposed technique limits the use of only numerical attribute set and input of attribute set is in ASCII file format.

Ghumbre et al. [11] proposed a system that used radial based function network structure and support vector machine for heart disease prediction. The results indicate that the accuracy of support vector machine is as good as radial based function network. The technique is sensitive to data acquisition method used. Chitra and Seenivasagam [12] used a supervised learning algorithm for prediction of heart disease at early stages. The proposed classifier is based on cascaded neural network (CNN) with hidden neurons. High specificity and sensitivity values show that the technique has a high probability of predicting healthy individuals and patients with heart disease.

Chen et al. [13] presented a framework for heart disease analysis and prediction using learning vector quantization (LVQ) algorithm. It uses ROC curve to display results and achieved 80 % accuracy. Enhancements can be made by

using text mining techniques along with data mining. Text mining has the capability to mine unstructured data available in heart disease datasets. Jabbar et al. [14] used association mining and genetic algorithm for heart disease prediction. High values of interestingness measure and accuracy were achieved. The framework used entire attribute set as input which can be further improved by feature reduction, selecting only those features that contribute towards the diagnosis of the disease.

Valente et al. [15] used multivariate linear regression (MLR) to study the relationship between MLR spatial activation patterns and behavioral ratings. Model coefficients are used to perform mapping and sought-after links are provided among different activities. It is concluded from the experimentation that multiple linear regression models are good for target modeling and it deals with high dimensional data. Rizk-Jackson et al. [16] proposed a framework using support vector regression and linear regression techniques in order to generate quantitative measurements of disease progression. Different neuroimaging measures were used to correlate the established measures of disease progression. It is concluded from results that there are different neuroimaging measures that are based on multivariate measurements and disease-state biomarkers can be established successfully. Maroco et al. [17] used different data mining techniques to improve accuracy, sensitivity and specificity of results generated from neuropsychological testing. The proposed technique compares linear discriminant analysis, quadratic discriminant analysis and logistic regression to other seven non parametric classifiers. Five fold cross validation is used to obtain the statistical distribution of results.

As the ensemble approach outperforms individual classifiers, many such approaches have been introduced in

Table 1 Comparison of accuracy, sensitivity and specificity for different machine learning techniques

Author/year/reference	Technique	Specificity (%)	Sensitivity (%)	Accuracy (%)
Chen et al. 2011 [13]	Artificial neural network	70	85	80
Das 2009 [5]	Neural network ensemble	95.91	80.95	89.01
Ghumbre et al. 2011 [11]	Support vector machine	88.50	84.06	85.05
	Radial basis function	82.10	82.40	82.24
Chitra et al. 2013 [12]	Cascaded neural network	87	83	85
	Support vector machine	77.5	85.5	82
Shouman et al. 2011 [38]	Nine voting equal frequency discretization gain ratio decision tree	85.2	77.9	84.1
Tu et al. 2009 [39]	J4.8 decision tree	84.48	72.01	78.9
	Bagging algorithm	86.64	74.93	81.41
Shouman et al. 2013 [40]	Gain ratio decision tree	81.6	75.6	79.1
	Naïve Bayes	80.8	78	83.5
	K nearest neighbor	85.1	76.7	83.2
Shouman et al. 2012 [41]	K mean clustering with Naïve Bayes algorithm	76.59	69.93	78.62

recent decade. Das et al. [5] presented an ensemble based method for diagnosis of heart disease. It combines different neural networks that are trained using same dataset and produce higher generalization. Only one heart disease database is used for the proposed technique and more datasets are required for verification of results. Helmy et al. [18] proposed an ensemble framework based on SVM, ANN and ANFIS to generate high prediction accuracy. Individual classifiers were trained using bagging algorithm and results reveal that heterogeneous ensemble has better results as compared to individual classifiers. It used only two datasets for result verification and does not provide any cross validation technique.

The proposed framework

The proposed framework consists of two main components.

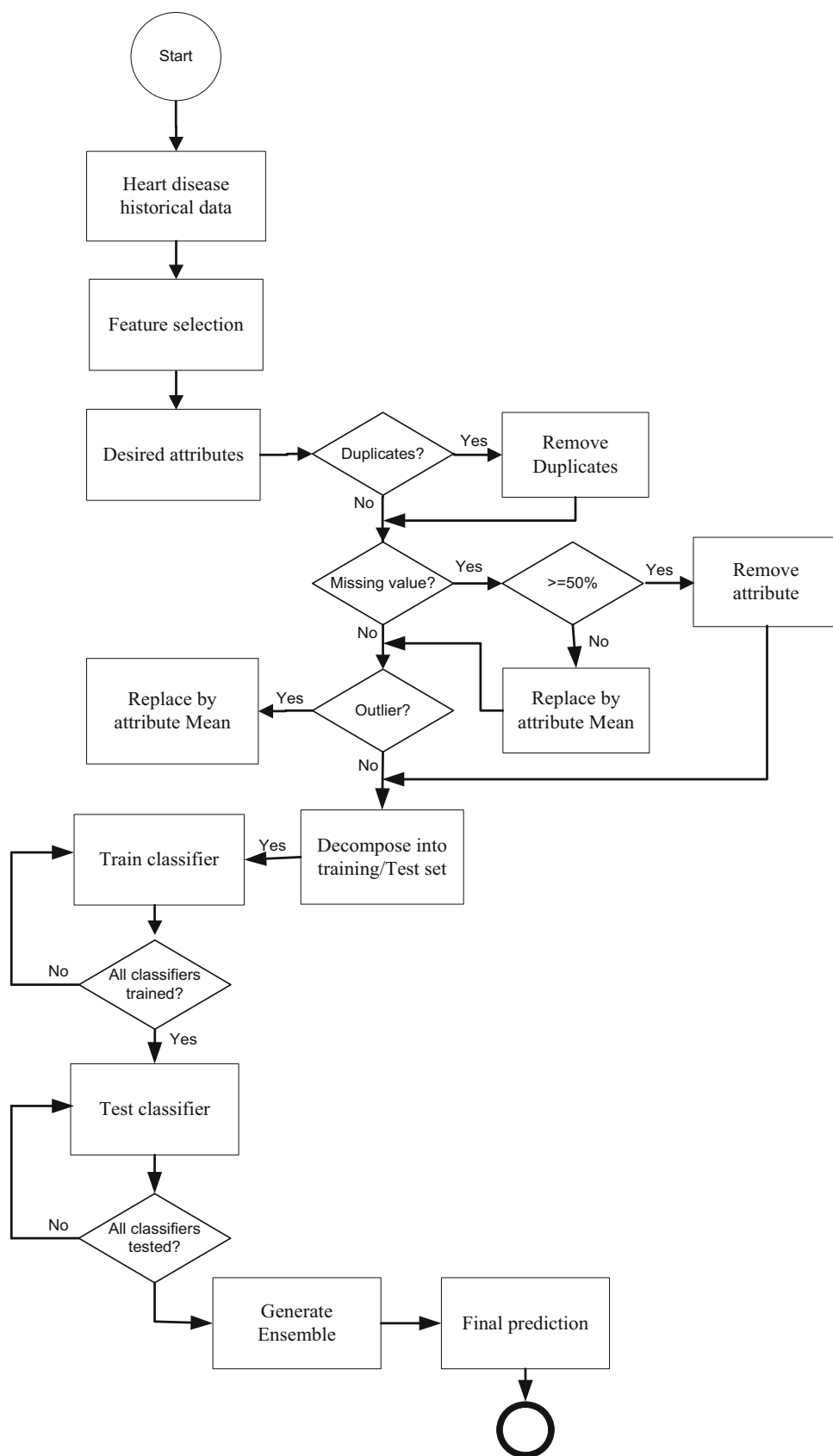
- (a) Data acquisition and pre-processing module.
- (b) BagMOOV, the ensemble model for heart disease prediction.

The data acquisition process obtains data from different repositories, performs data partition and variable selection. Pre-processing steps involve: missing value imputation, outlier detection, feature selection and class label identification. This is then followed by training each classifier using the training set and finally, the proposed ensemble model, BagMOOV combines five different classifiers. For each classifier, the weight is calculated based on F-measure of the training dataset. The final output of the ensemble classifier is the label with highest weighted vote [8, 19]. The flowchart of the proposed approach is given in Fig. 2.

Data acquisition and pre-processing module

The basic purpose of data acquisition and pre-processing module is to obtain data from different heart disease repositories and then refine them into a form that is suitable for subsequent analysis. Each dataset holds the feature space that will ultimately differentiate the data into healthy individuals and sick (heart disease) patients. Each dataset has different set of attributes and data types. The data is then divided into training set and test set by data partition component. Ten fold cross validation is used to partition the dataset into ten mutually exclusive partitions. The biasness is avoided by randomly selecting the samples from each partition. The partitioning results in reduction of computation time for preliminary model runs. The pre-processing phase involves multiple steps that are applied on each dataset sequentially. It includes feature selection, missing value imputation, noise removal and outlier detection.

- **Feature selection** This process involves feature reduction by selecting only those attributes which contribute towards final prediction of disease. The rejected features will not be used for subsequent modules and analysis. There are multiple steps involved in the process of feature selection and identification [20]. The generation and selection procedures are two of the most important steps. The generation procedure involves generation features subset whereas selection procedure will evaluate these features on the basis of different criteria. The generation procedure can result in an empty set, subset based on randomly selected attributes or a set based on all attributes. Forward selection is used in case of empty set which iteratively adds the attributes in feature set whereas backward elimination is used in case of all attributes, which iteratively eliminates the irrelevant attributes from feature set. The relevancy of an attribute is measured by wrapper approaches. The main focus of a wrapper approach is classification accuracy [21]. The estimation accuracy of each feature set is calculated that is a candidate for adding or removing from the dataset. We have used cross validation for accuracy estimation of each feature set of the training set. The feature selection process continues until pre-specified number of features is achieved or some threshold criteria are attained [21]. The proposed ensemble framework performs feature selection for each dataset individually. We have used benchmark heart disease datasets in the research and they do not contain any irrelevant features as the respective publishers have already processed them. Therefore, the entire feature set of each dataset will be used for subsequent analysis. The defined feature selection method will be used for any other dataset that may contain irrelevant attributes.
- **Noise removal** Noise is referred to as random error or variance in a measured attribute. There are multiple techniques for noise removal such as regression analysis, binning and clustering. The proposed pre-processing involves noise removal using binning and the refined data is passed onto the next process. Benchmark heart disease datasets have been used that do not contain any noise because they are already processed by the respective publishers. For other datasets, the noise removal method will be used.
- **Outlier detection** Outlier is a type of noise and they are attribute values that fall above or below a defined range. The outlier detection procedure removes outliers from each attribute. Inter-Quartile Range (IQR) is used to detect outliers and any value not in the range of ± 1.5 IQR will be replaced with attribute mean for continuous attributes and mod for categorical values. No outliers were detected in the datasets used in this research.

Fig. 2 Flow chart of proposed framework

- **Missing value imputation** Missing data in medical datasets must be handled carefully because they have a serious effect on conclusions and interpretation of data. The proposed pre-processing module also involves missing data handling and **missing values are replaced by the mean/mode of each attribute** depending on the data type. **If the missing values for a particular attribute are more than 50 % of all instances, that attribute will be automatically discarded.** Mean substitution is a conservative procedure as the distribution mean as a whole does not change and researchers don't have to guess at missing values. In this research we have used **group mean substitution** instead of simple mean substitution. This is because in medical datasets, we have both male and female patients, as an example the use of menopause are recorded only for women; it is not possible to impute appropriate values for men. Therefore we impute a missing value using the class-conditional mean of the feature (i.e., the mean feature value of all points within the same class as the instance with the missing value). For example if the case with a missing value is a male patient with hypertension, the mean value for male patient with hypertension is calculated and inserted in place of the missing value.

BagMOOV ensemble

The proposed ensemble uses Bootstrap Aggregation (Bagging) with multi-objective optimized weighted vote based technique. The ensemble consists of combination of heterogeneous classifiers which are Naïve Bayes (NB), linear regression (LR), quadratic discriminant analysis (QDA), instance based learner (IBL) and support vector machine (SVM). However, the proposed framework can be implemented using any set of classifiers which may be homogenous, heterogeneous or a combination of both. The description of each individual classifier and the proposed weighted voting ensemble approach are explained in this section.

Base classifiers

A. Naïve Bayes (NB) classifier Naïve Bayes classifier depends on the **hypothesis that presence or absence of a disease is independent of the feature space**. Various supervised learning algorithms can be used to train the probability model [22].

It requires a small training dataset and only the attributes of given class are required instead of entire covariance matrix as they are independent of each other [23]. Following formula is used to classify the given problem:

$$P(C_k|X) = P(C_k) \times \frac{P(X|C_k)}{P(X)} \quad (1)$$

where X is an example that needs to be classified, C_k is a possible class and $P(C_k|X)$ is the probability of vector X belonging to class C_k .

B. Linear regression (LR) Regression is one of the most common techniques used for prediction. It determines the relationship between set of independent variables and a dependent variable in order to perform prediction for dependent variable. The prior relationship identifies the future outcome [24]. A simple regression is where only one variable is used as independent variable whereas multiple regression uses more than one independent variables to predict the value of dependent variable. Regression models are used to determine graphical relationship between variables. The regression model can be defined by following formula [25]:

$$y_i = \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \varepsilon_i = x_i^T \beta + \varepsilon_i \quad (2)$$

where $i = 1, \dots, n$. T presents transpose of x and it is used to calculate the inner product between x_i and β . Combining above n equations and representing in vector form as follows:

$$y = X\beta + \varepsilon \quad (3)$$

where **y represents dependent variable, X represents independent variables, β shows regression coefficients which are p -dimensional parameter vectors and ε_i is error term.**

Multiple attributes given in heart disease datasets are considered as independent variables whereas output class (healthy/sick) is considered as dependent variable.

C. Quadratic discriminant analysis (QDA) It is a machine learning classifier that uses **quadratic surface to separate two or more classes**. It assumes that each class has **normal distribution** and does not require any parameters to tune the algorithm. QDA allows each class to have its own covariance matrix and tends to fit the data best as compared to linear discriminant analysis (LDA). It is inherently multiclass and has closed-form solution [26]. Following formula is used to determine quadratic discriminant function:

$$\delta_k(x) = -\frac{1}{2} \log \left| \sum_k \right| - \frac{1}{2} (x - \mu_k)^T \sum_k^{-1} (x - \mu_k) + \log \pi_k \quad (4)$$

where **μ represents mean of each class, k is the number of classes, \sum_k shows covariance matrix for k class and π_k is the prior probability of class k .** Following classification rule is used to classify given datasets:

$$G(x) = \arg \max_{\delta_k(x)} \quad (5)$$

A class which maximizes the quadratic discriminant function will be considered as output class. The discriminant analysis has reduced error rate and uses multiple dependent variables to determine output class. Moreover, the interpretation between-groups is easier to interpret and calculate.

D. Instance based learner (IBL) Instance based Learner compares each new instance with the instances stored in memory from the training set. It is also termed as lazy learning and works on the principal of **k nearest neighbor (kNN) approach**. Its advantage over other data mining techniques is that the **model automatically adapts to unseen data**. The nearest neighbor is identified by a distance function which is selected depending upon the data type of attributes. Following formula is used to calculate the **distance between two feature vectors** [27]:

$$d(x_i, x_j) = \sum_{q \in Q} (X_{iq} - X_{jq})^2 + \sum_{c \in C} L_c(X_{ic} - X_{jc}) \quad (6)$$

where L_c defines the distance between two categorical attributes in the form of $M \times M$ matrix, Q represents set of quantitative features and C stands for set of categorical features. Let an instance x_i has the k nearest neighbors represented by N_i and the distance is denoted by d then the **majority voting scheme is used to determine the votes $V_i(t)$ having the label t for all neighbors of x_i** . Formally, it can be written as [27]:

$$V_i(t) = \sum_{k \in N_i} I(t, y_k) \quad (7)$$

where I represents the indicator function and $I(t, y_k) = 1$; if $t = y_k$ and $(t, y_k) = 0$; otherwise.

E. Support vector machine (SVM) SVM is supervised learning algorithm that is used for binary classification. A prediction model is constructed for each input test set and produces output in the form of two classes making it non probabilistic binary classifier [11]. SVM finds linear maximum margin hyper-plane. It is defined by a weight vector w and bias b which is hyperplane distance from center. The non-linear separation of dataset is performed by using a kernel function. The following classification rule is used by SVM classifier for solving the given problem:

$$\text{Sgn}(f(x, w, b)) \quad (8)$$

$$f(x, w, b) = \langle w \cdot x \rangle + b \quad (9)$$

where $f(w, b)$ presents maximum margin hyperplane for the complex problem and x denotes the example to be classified. In our study, we use two attributes for each dataset, selected on the basis of highest information gain.

Each individual classifier which is used by the ensemble is trained using the training data in order to make them

useful for heart disease prediction. The feature space and the resultant class labels of each dataset are known to each trained classifier, which then has the capability to predict healthy and sick individuals.

Bootstrap aggregation

Bagging stands for Bootstrap Aggregation, which **combines the results of base classifiers treating each model with equal weight (vote) to generate final prediction**. In order to generate better prediction results, each base classifier is trained using randomly drawn sample sets (bootstrap samples) with replacement from original training set [25]. The proposed ensemble approach is an enhancement of bagging algorithm and it can be divided into two stages. At first stage, original training set for each heart disease dataset is divided into **multiple bootstrap sets with replacement**. In order to create bootstrap samples from training set of size m , **t multinomial trials are performed, where one of the m examples is drawn for each trial. The probability of each example to be drawn in each trial is $1/m$** . The proposed ensemble algorithm chooses a sample r from 1 to m and the r th training example is added to bootstrap training set S . Moreover, it is possible that some

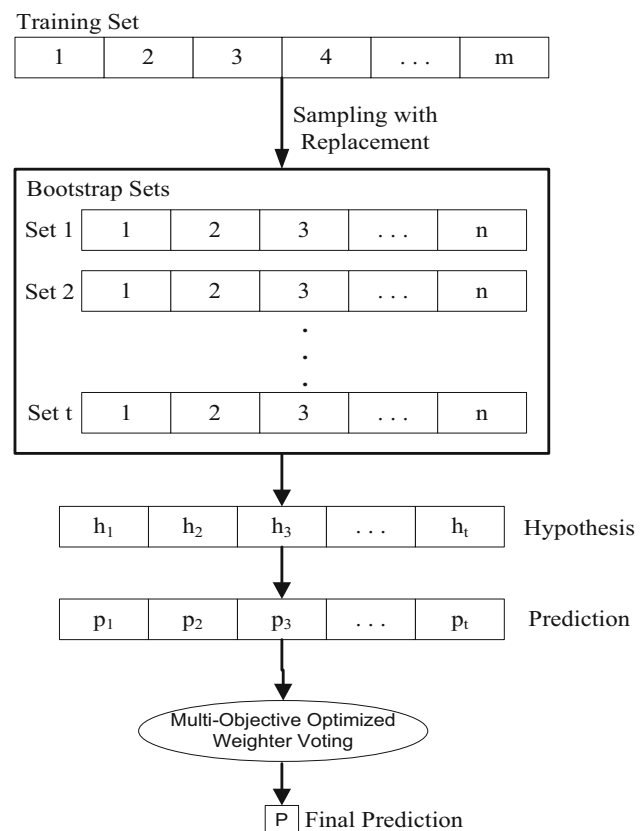


Fig. 3 The flowchart of proposed BagMOOV algorithm

Table 2 Description of five heart disease datasets from UCI repository

No.	Dataset	Attributes	Instances	No. of attributes
1	SPECT	F1...F22, (partial diagnosis 1...22, binary), goal	267	23 (22 binary + 1 binary class)
2	SPECTF	F1R...F22R (count in ROI 1...22 in rest), F1S...F22S (count in ROI 1...22 in stress), goal	267	45 (44 continuous + 1 binary class)
3	Heart disease	Age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal, goal	303	14 (6 real, 1 ordered, 3 binary, 3 nominal, 1 binary class)
4	Statlog	Age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal, goal	270	14 (6 real, 1 ordered, 3 binary, 3 nominal, 1 binary class)
5	Eric	Age, chest_pain, rest_bpress, blood_sugar, rest_electro, max_heart_rate, exercise_angina, goal	209	8 (7 real, 1 binary class)

Table 3 Traditional Confusion Matrix

Predicted Class	Known class	
	A	B
A	True positives	False negatives
B	False positives	True negatives

of the training examples will not be selected in bootstrap training sets whereas others may be chosen one time or more. At second stage, classifiers' training is performed using bootstrap training sets generated during the first stage.

We argue that each classifier in the bagging approach should not have the same weight as each classifier has a different individual performance level. Therefore, we propose to use multi-objective optimized weighting scheme instead of simple voting. The BagMOOV ensemble will return a function $h(d)$ that classifies new samples into class y **having the highest weight from the base models h_1, h_2, \dots, h_t** . t bootstrap training sets are created which have some differences with each other. The ensemble model will perform better than individual classifier if the difference among bootstrap training sets induces a noticeable difference among M individual classifiers, generating reasonably good performance by each of them. [28] proposed that a bagged ensemble approach outperforms each base classifier if the base classifiers are trained using sample sets where differences in sample training sets induce a significant difference in the base classifiers. The flowchart of proposed BagMOOV ensemble algorithm is given in Fig. 3. It shows that training set is divided into multiple datasets using bootstrap aggregation method, then proposed technique is applied on these datasets and final prediction is obtained.

Multi-objective optimization criteria The proposed ensemble classifier is based on the principle of multi-objective optimization where we try to optimize multiple goals simultaneously. Formally, it can be stated as [19]: Find the number of vectors V_k where $V_k =$

$\{v_1, v_2, v_3, \dots, v_n\}$ for each classifier C_k and $C_k = \{C_1, C_2, C_3, \dots, C_n\}$ such that simultaneously optimize the N objective criteria, while satisfying the constraints, if any. The multi-objective optimization focuses on the maximization problem which states that a solution v_i will always dominate a solution v_j if for all $K \in 1, 2, 3, \dots, N$, $f_k(v_i) > f_k(v_j)$ where f_k is an objective function. The maximization problem stands true for each objective function used for the proposed classifier ensemble technique.

Selection of objectives The choice of objective functions should be as much contradictory as possible in order to achieve high performance of weighted voting for ensemble classifier. We have used precision and recall as two objective functions. The recall tries to increase the number of healthy samples while precision tries to increase the number of correct healthy samples as much as possible. The f -measure is then calculated for the training set using the precision and recall for each classifier. F -measure results in a value (weight) that has the highest precision and recall combination. The weights are then normalized by applying min-max normalization using the following formula [29]:

$$V'_i = \frac{V_i - \min_A}{\max_A - \min_A} (\text{new_max}_A - \text{new_min}_A) + \text{new_min}_A \quad (10)$$

where \min_A and \max_A are the min and max values for attribute A in the training set, respectively.

new_min_A is normalized minimum value for attribute A , specified as 0.1 and new_max_A is normalized maximum value for attribute A , specified as 1.0.

Precision presents percentage of healthy samples labeled as healthy by the classifier and it is defined as:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \quad (11)$$

Recall presents the relevant instances that are retrieved by the classifier and is defined as:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \quad (12)$$

The F-measure is the weighted average of recall and precision, represented by:

$$\text{F-Measure} = 2 \times \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \quad (13)$$

Both objective functions (Precision, Recall) achieve two different classification qualities and often there is inverse relationship between them where one is increased at the cost of reducing the other. For example, an information retrieval system such as a search engine increases the recall by retrieving more files whereas the precision is decreased by increasing number of irrelevant files that are retrieved. Thus the motivation of the proposed ensemble classifier is to simultaneously optimize the two objectives.

Working of the proposed ensemble

Working of the proposed ensemble classifier can be easily understood by the following example.

1. Suppose that the classifier training is performed for training data and f-measure is calculated. Naïve Bayes (NB), linear regression (LR), quadratic discriminant analysis (QDA), instance based learner (IBL) and support vector machine (SVM) are used as individual classifiers. Following f-measure results are generated for each classifier: NB = 60 %, LR = 70 %, QDA = 80 %, IBL = 85 %, SVM = 90 %
2. Now, votes are calculated using the formula given in Eq. (10) having new_max = 1.0, new_min = 0.1, max = 100, min = 0. The resultant weights are as follows: NB = 0.6, LR = 0.7, QDA = 0.8, IBL = 0.9, SVM = 0.9
3. Suppose, the classifiers have predicted the following classes for a test instance: NB = Class 0, LR = Class 0, QDA = Class 1, IBL = Class 0, SVM = Class 1
4. The weighted vote based ensemble classifier will generate the following prediction results:
Class 0: NB + LR + IBL $\rightarrow 0.6 + 0.7 + 0.9 \rightarrow 2.4$,
Class 1: QDA + SVM $\rightarrow 0.8 + 0.9 \rightarrow 1.7$
5. Hence, according to weighted vote based ensemble classifier class 0 has higher value as compared to class 1. Therefore, the test instance will be classified as Class 0.

Dataset description

Five different datasets have been used in the proposed research. SPECT, SPECTF, Heart disease and Statlog

datasets are downloaded from UCI machine learning repository¹ and Eric dataset is downloaded from ricco.² Each database contains a feature set and a column indicating the class label. The class label of each dataset is replaced with 0 and 1 in order to maintain consistency. Table 2 shows statistics of five heart disease datasets that are used for experimentation and analysis. It shows dataset name, number of attributes, number of instances and attributes name for each dataset (SPECT, SPECTF, Heart disease, Eric and Statlog).

Results and discussion

The experiments are conducted on five different heart disease datasets having different set of attributes. The classifier evaluation is performed using test sets of each dataset and then results are analysed. **tenfold cross validation is used to alleviate the insufficiency of samples**, dividing each dataset into training set and test set. The learning of each classifier is performed on training set and then they are analysed using the test set. Five classifiers (NB, LR, QDA, IBL and SVM) are first trained using the training sets and then they are tested on a separate unseen test set.

The following evaluation measures are used to assess the performance of the proposed ensemble.

Statistical significance

The performance of the proposed ensemble model as well as of individual classifiers is evaluated by calculating the Confusion matrix, Sensitivity, Specificity, F-Measure and Accuracy of five datasets. The mathematical equations for sensitivity, specificity and f-measure are given in (14), (15) and (16) respectively.

$$\text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \quad (14)$$

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{False Positives} + \text{True Negatives}} \quad (15)$$

$$\text{F-Measure} = \frac{2 \times \text{Sensitivity} \times \text{Specificity}}{\text{Sensitivity} + \text{Specificity}} \quad (16)$$

The accuracy measure presents the proportion of instances that are correctly classified. Mathematically,

¹ <<http://archive.ics.uci.edu/ml/datasets.html>> [last Accessed: Sep 25 2013].

² <<http://archive.ics.uci.edu/ml/datasets.html>> [last Accessed: Sep 25 2013].

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{True Positives} + \text{False Positives} + \text{True Negatives} + \text{False Negatives}} \quad (17)$$

A **confusion matrix** used for the comparison of the actual classification of data set with the number of correct and incorrect predictions made by the model. The number of rows and columns represent the number of classes. It can be used to calculate the accuracy of each classifier [30]. The traditional confusion matrix is defined in Table 3. It shows information about actual and predicted classifications done by classification framework.

The ensemble model is applied on each test set which processes each instance individually. Each instance of test set is classified into healthy or sick class labels. **The ten confusion matrices obtained from each fold of cross validation are then summed up into final confusion matrix.** The average prediction results for all subsets are calculated and then analysed to verify the superiority of proposed ensemble. We have used multiple heart disease datasets in order to show that proposed ensemble model has robust performance when applied on different kind of medical heart disease datasets. Different input features/attributes

are selected from each dataset in order to generate training sets and test sets. Table 4 shows the comparison of accuracy, sensitivity, specificity and F-measure results of the proposed ensemble classifier for all datasets with individual classifier techniques such as Naïve Bayes (NB), Support vector machine (SVM), Linear Regression (LR), Quadratic discriminant analysis (QDA) and k-nearest neighbour (KNN). It is recognizable from Table 4 that, the proposed ensemble model produces significant results when compared with individual classifiers. Highest accuracy level is achieved for all the datasets when compared to other individual classifiers.

Table 5 shows comparison of accuracy, sensitivity, specificity and F-measure for proposed ensemble classifier with other ensemble classifiers such as Bagging [31], AdaBoost [32], Stacking [33] as well as with neural network ensemble [5]. The comparison of results show that proposed ensemble model performed much better than traditional classification techniques. Again, highest

Table 4 Accuracy, Sensitivity, Specificity and F-Measure comparison of Proposed Ensemble vs Individual Classifiers

Classifiers	Acc	Sen	Spec	F-M	Acc	Sen	Spec	F-M
Cleveland dataset (%)					Eric dataset (%)			
NB	77.23	81.71	71.94	76.51	68.90	77.78	57.61	66.19
SVM	80.86	93.90	65.47	77.15	78.47	89.74	64.13	74.81
LR	83.50	88.41	77.70	82.71	77.99	88.89	64.13	74.51
QDA	65.68	68.29	62.59	65.32	46.41	10.26	92.39	18.46
kNN	64.36	68.90	58.99	63.56	65.55	68.38	61.96	65.01
BagMOOV	84.16	93.29	73.38	82.15	80.86	86.32	73.91	79.64
SPECT dataset (%)					SPECTF dataset (%)			
NB	80.52	76.36	81.60	78.90	78.28	23.64	92.45	37.65
SVM	67.04	85.45	62.26	72.04	79.40	0.00	100	0.00
LR	83.15	38.18	94.81	54.44	78.28	9.09	96.23	16.61
QDA	83.52	36.36	95.75	52.71	20.60	100	0.00	0.00
kNN	79.40	7.27	98.11	13.54	71.91	36.36	81.13	50.22
BagMOOV	82.02	27.27	96.2	42.50	78.28	7.27	96.70	13.53
Statlog dataset (%)								
NB	78.52	82.00	74.17	77.89				
SVM	81.85	94.67	65.83	77.66				
LR	82.59	87.33	76.67	81.65				
QDA	68.15	64.00	73.33	68.35				
kNN	65.56	68.67	61.67	64.98				
BagMOOV	84.07	92.00	74.17	82.13				

Acc accuracy, Spec specificity, Sen Sensitivity, F-M F-measure, NB Naïve Bayes, SVM support vector machine, LR linear regression, QDA quadratic discriminant analysis, kNN k nearest neighbor

Table 5 Accuracy, Sensitivity, Specificity and F-Measure comparison of Proposed Ensemble vs Other Ensembles

Ensembles	Acc	Sen	Spec	F-M	Acc	Sen	Spec	F-M
Cleveland dataset (%)					Eric dataset (%)			
Bagging	74.59	81.10	66.91	73.32	73.21	76.92	68.48	72.46
Adaboost	64.36	68.90	58.99	63.56	65.07	68.38	60.87	64.40
Stacking	82.51	88.41	75.54	81.47	79.43	87.18	69.57	77.38
NNE	80.86	82.93	78.42	80.61	77.03	79.49	73.91	76.60
BagMOOV	84.16	93.29	73.38	82.15	80.86	86.32	73.91	79.64
SPECT dataset (%)					SPECTF dataset (%)			
Bagging	79.40	0.00	100.00	0.00	71.16	90.91	66.04	76.50
Adaboost	78.28	7.27	96.70	13.53	71.91	36.36	81.13	50.22
Stacking	79.40	0.00	100.00	0.00	70.41	90.91	65.09	75.87
NNE	79.03	47.27	87.26	61.32	77.53	47.27	85.38	60.85
BagMOOV	82.02	27.27	96.23	42.50	78.28	7.27	96.70	13.53
Statlog dataset (%)								
Bagging	73.33	80.00	65.00	71.72				
Adaboost	65.93	69.33	61.67	65.28				
Stacking	82.59	90.00	73.33	80.82				
NNE	78.15	77.33	79.17	78.24				
BagMOOV	84.07	92.00	74.17	82.13				

Acc accuracy, *Spec* specificity, *Sen* sensitivity, *F-M* F-measure, *NNE* neural network ensemble

Table 6 Accuracy comparison of machine learning techniques for Cleveland heart disease dataset

Reference	Year	Technique	Accuracy (%)
Shouman et al. [40]	2013	Gain ratio decision tree	79.1
		Naïve Bayes	83.5
		K nearest neighbor	83.2
Chaurasia et al. [42]	2013	CART	83.4
		ID3	72.9
		Decision table	82.5
Sunday et al. [43]	2012	WAC	84
		Naïve bayes	78
Soni. et al. [44]	2011	WAC	57.75
		CBA	58.28
		CMAR	53.64
		CPAR	52.32
		Artificial neural network	80
Chen et al. [13]	2011	Artificial neural network	80
Proposed technique		BagMOOV Model	84.16

accuracy level is achieved for all the datasets when compared to ensembles. SVM shows high sensitivity and low accuracy for almost all datasets except SPECT dataset where 0 % sensitivity and 79.4 % accuracy is achieved. It is possible for a classifier to have low accuracy with high sensitivity because accuracy is derived from both sensitivity and specificity. If any of the sensitivity or specificity is high, accuracy can be biased towards highest value; and if both are high then accuracy will be high [34]. Moreover, even both sensitivity and specificity are high, it does not necessarily mean that accuracy will be equally high as

well. Accuracy also depends on the factor that how common the disease is in selected population. A diagnosis for rare conditions in the population of interest may result in high sensitivity and specificity, but low accuracy [35].

The proposed technique shows a consistent accuracy level of around 82 % whereas other classifiers are not stable as observed in the results. The proposed model increased the classification and prediction quality by improving sensitivity and specificity, as a result of which f-measure and accuracy are enhanced to a reasonable margin as compared to conventional models. The proposed

Table 7 ANOVA Statistics for heart disease datasets

Classifiers	Variation	SS	df	MS	F	F crit	P value	SS	df	MS	F	F crit	P value
Cleveland dataset							Eric dataset						
NB	BG	0.7277	1	0.7277	4.6918	3.8569	0.0307	1.4952	1	1.4952	8.0645	3.8639	0.0047
	WG	93.6832	604	0.1551				77.1292	416	0.1854			
SVM	BG	0.5957	1	0.5957	3.8864	3.8569	0.0491	0.0598	1	0.0598	0.3678	3.8639	0.5446
	WG	92.5809	604	0.1533				67.6555	416	0.1626			
LR	BG	0.0066	1	0.0066	0.0485	3.8569	0.8257	0.6914	1	0.6914	3.8973	3.8639	0.0490
	WG	82.1452	604	0.1360				73.7990	416	0.1774			
QDA	BG	5.17492	1	5.1749	28.7549	3.8569	0.0001	12.4019	1	12.4019	61.1820	3.8639	0.0001
	WG	108.6997	604	0.1800				84.3254	416	0.2027			
kNN	BG	5.9406	1	5.9406	32.6487	3.8569	0.0001	2.4498	1	2.4498	12.8123	3.8639	0.0004
	WG	109.9010	604	0.1820				79.5407	416	0.1912			
SPECT dataset							SPECTF dataset						
NB	BG	0.6760	1	0.6760	4.0158	3.8590	0.0456	1.0787	1	1.0787	5.6140	3.8590	0.0182
	WG	89.5581	532	0.1683				102.2172	532	0.1921			
SVM	BG	2.9963	1	2.9963	16.2047	3.8590	0.0001	0.7491	1	0.7491	3.9607	3.8590	0.0471
	WG	98.3670	532	0.1849				100.6142	532	0.1891			
LR	BG	1.0787	1	1.0787	6.2405	3.8590	0.0128	1.5749	1	1.5749	8.0522	3.8590	0.0047
	WG	91.9551	532	0.1729				104.0524	532	0.1956			
QDA	BG	1.5749	1	1.5749	8.8977	3.8590	0.0030	44.4120	1	44.4120	265.2618	3.8590	0.0001
	WG	94.1648	532	0.1770				89.0712	532	0.1674			
kNN	BG	1.7996	1	1.7996	10.0783	3.8590	0.0016	2.1648	1	2.1648	10.8956	3.8590	0.0010
	WG	94.9963	532	0.1786				105.7004	532	0.1987			
Statlog dataset													
NB	BG	0.4167	1	0.4167	2.7440	3.8588	0.0982						
	WG	81.6926	538	0.1519									
SVM	BG	0.7407	1	0.7407	4.7189	3.8588	0.0303						
	WG	84.4519	538	0.1570									
LR	BG	0.0296	1	0.0296	0.2126	3.8588	0.6449						
	WG	74.9704	538	0.1394									
QDA	BG	3.4241	1	3.4241	19.4403	3.8588	0.0001						
	WG	94.7593	538	0.1761									
kNN	BG	4.6296	1	4.6296	25.6464	3.8588	0.0001						
	WG	97.1185	538	0.1805									

NB Naïve Bayes, SVM support vector machine, LR linear regression, QDA quadratic discriminant analysis, kNN k nearest neighbor, BG between groups, WG within group, SS sum of squares, df degree of freedom, MS mean square, F F-statistic, F-crit F-critical

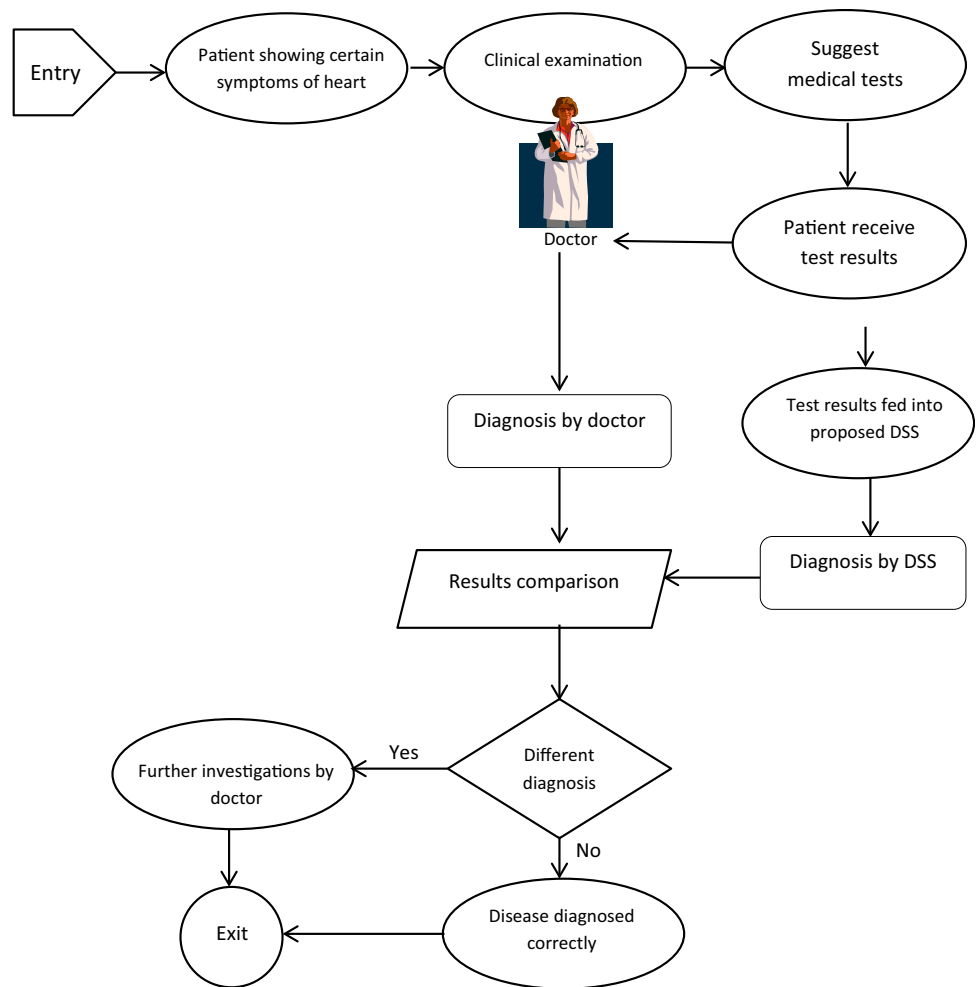
ensemble model achieved best accuracy of 84.16 % with 93.29 % sensitivity, 96.70 % specificity, and 82.15 % f-measure.

Table 6 shows accuracy comparison of proposed Bag-MOOV ensemble technique with state of art techniques. The analysis of table indicates that proposed ensemble model achieved higher accuracy of disease classification for Cleveland heart disease dataset.

Another method used to show the significance of results is ANOVA (Analysis of Variance) statistics. It is a

statistical test used to measure the difference between group means and their variations such as variations among and between groups [36]. The results of ANOVA statistics for each dataset are given in Table 7. The f-ratio and p value obtained from ANOVA statistics indicate that the results are extremely statistically significant for most of the datasets at 95 % confidence interval. It is clear from analysis of ANOVA statistics that proposed framework results are statistically significant when compared with other classifiers.

Fig. 4 Proposed heart disease prediction DSS for real-time clinical practice



Discussion

Heterogeneous classifier ensemble model is used by combining entirely different type of classifiers to achieve a **higher level of diversity**. The diversity parameter can be determined by the extent to which each individual classifier disagrees about the probability distributions for the test datasets. The Naïve Bayes classifier considers each attribute independently without taking into account the relation between them whereas the proposed ensemble model can handle dependency and relation between given attribute set by using Instance based learning algorithm where neighbors determine the class label for unknown instance. The **Linear regression model** determines the statistical relationship between various independent and dependent variables and achieves optimal results. It **limits the prediction of numeric output** where Naïve Bayes Classifier overcomes this limitation. The **quadratic discriminant analysis model** can handle multiple dependent variables to determine output class and reduce error rate. Moreover, it makes very easy to determine between-group differences.

The individual IBL algorithm has some limitations such as it is computationally intensive and requires lot of storage space. The **ensemble model** has resolved the storage problem by selecting only necessary and useful features for heart disease analysis and prediction. The **SVM algorithm** performs the feature selection by using only subset of data chosen based on information gain. Thus, all of the five selected classifiers complement each other very well. In any scenario where one classifier has some limitation, other classifier overcomes it and as a result, higher ensemble performance is achieved. The proposed ensemble shows stable performance which is a significant assessment for accurately determining the heart disease patients. The individual classifiers did not achieve such stability. Moreover, traditional ensemble classifiers did not achieve such performance results when applied on heart disease datasets.

Linear classification is a useful learning tool for prediction and analysis. We have used linear classification models for base classifiers. The **benefit of using linear models** is that training and testing speed is much faster as compared to non-linear models [37] and can be used

Table 8 RIC dataset features for heart disease

Patient_ID	2DEchoResult_Part1	BP-MA_mmHg-uppLim
Age	2DEchoResult_Part2	BP-MA_mmHg-lowLim
Gender	2DEchoResult_Disease1	AffectedArea1
Protocol	2DEchoResult_Disease2	AffectedArea2
BMI	P_Complain1	AffectedArea3
Known_Disease1	P_Complain2	LV_Myocardium
Known_Disease2	P_Complain3	Defect_Size
Known_Disease3	RestingECGResult1	Defected_AreaSize
FstMI_Type	HeartRate-BI_BPM	Defect_Segment
Angiography_Result1	HeartRate-MA_BPM	Via/Non-via
2DEchoResult	BP-BI_mmHg-uppLim	IsDefected
2DEchoResult_Part	BP-BI_mmHg-lowLim	I_LVEF

Table 9 An example of if-then rules generated by proposed ensemble framework

Rules	Diagnosis
If thal < 4.5 and cp < 3.5 and oldpeak < 2.5 and chol < 272 and age < 57	Class = 0
If thal ≥ 3.5 and cp ≥ 3.5 and ca ≥ 0.5 and exang ≥ 0.16	Class = 1
If thal < 45 and cp ≥ 3.5 and oldpeak ≥ 2.5	Class = 1
If cp ≥ 3.5 and ca < 0.5 and bps < 145 and age < 66 and fbs < 0.5	Class = 0
If thal ≥ 4.5 and cp < 3.5 and ca ≥ 0.5 and exang < 0.5	Class = 0

effectively for large scale applications. Moreover, they directly work on data in the original input space. The training of linear classifier is also more efficient [37].

The fundamental advantage of proposed BagMOOV ensemble technique is improved accuracy and efficiency for heart disease classification and prediction as compared to other state of art techniques. It can help healthcare professionals to make valuable decisions related to the diagnosis of heart disease. It can provide several benefits to healthcare resources such as effective management of patient's data, smarter treatment techniques, improved patient care, recognize high-risk patients and health policy planning etc. Moreover, the performance of unstable learning algorithms can be improved using proposed technique. The proposed BagMOOV approach utilizes bagging approach with weighted voting scheme where bagging works as a bias and variance reduction method of base classifiers. The mean square error (MSE) of base classifiers is reduced due to its smoothing effect and it works as a smoothing operation which improves the predictive performance of classification and regression methods. The proposed method is more resilient to noise than boosting and other ensemble techniques and can be trivially parallelizable and more amendable to build the large ensemble. All kinds of variables can be handled by the proposed ensemble model such as interval-scaled, categorical, continuous, real and binary variables. The analysis of ANOVA statistics also proved that proposed model is effective for heart disease diagnosis and can be effectively utilized in real time environment. The combination of multiple techniques using

bootstrap aggregation and weighted voting method makes the model complex and interpretation becomes difficult.

Real-time implementation of the proposed framework

This section discusses how the proposed DSS can be used in real-time environment and with real-life biological test data. The workflow of the DSS is shown in Fig. 4 where the adaption of DSS is quite beneficial for heart disease diagnosis. A patient showing symptoms for heart disease goes to doctor. The doctor performs clinical examination and suggests medical tests related to heart disease such as cholesterol level, blood pressure, heart rate, defect size etc. The patient receives tests results and goes to the doctor again. The doctor diagnoses the diseases based on test results, knowledge and experience. Moreover, in order to attain benefits of proposed DSS, he also enters the data to DSS. The system makes a disease prediction for that particular data. The doctor then compares the prediction made by him and the proposed DSS. If these results are same, this adds weight to the diagnosis performed by the doctor but if they are different, further investigations are performed by doctor.

It should be noted, that an intelligent DSS is not a replacement for doctor or practitioner but it can help them to gather and interpret information and build a foundation for decision-making related to particular disease. There are multiple ways in which the proposed DSS can help both the doctors as well as individual patient. For example:

- *Diagnose by regularly interpreting and monitoring patient data* The proposed DSS can implement rules and patterns for individual patients on the basis of clinical parameters and **warning can be generated in case of rule violations.**
- *Heart disease management using benchmarks and alerts* **A deviation from normal value such as high heart beat reading could result in an intervention before the condition worsens.**

Evaluation of the proposed framework in real-time environment

In order to evaluate the proposed system, **collaboration was sought with Rawalpindi Institute of Cardiology** to apply the proposed framework in real-time environment and on real-life data. Rawalpindi Institute of Cardiology³ is one of the major tertiary cardiac care centers in Pakistan. This 272-bedded hospital provides care for the cardiac patients from over the country. It equipped with Coronary Care units, surgical ITC, Departments of Cardiac Electrophysiology, Echocardiography, Exercise Tolerance Test and Nuclear Cardiology. They kindly agreed to allow the use of proposed DSS for research purposes only under the strict supervision of a team of medical experts and their information technology team.

The first step was to build a knowledge from which the classifiers maybe trained for prediction of heart disease. A team of five based on medical practitioners and doctors helped us to define the medical knowledge in order to classify the healthy and heart disease patients. A patient will come to the doctor to be diagnosed, the medical knowledge was stated in natural language and was written as follows: **A person is having heart disease if a person has high blood pressure (BP) of 180/100 and heart rate of 100 beats per minute (BPM), a strong prior history of cardiovascular ailment, echo results are abnormal and resting electrocardiographic results are positive then there are strong chances of heart disease. On the other hand if a person has normal blood pressure (BP) of 120/80 and heart rate of 60 beats per minute (BPM), echo results are normal and no prior history of cardiovascular disease then there are less chances of heart disease.**

In accordance with this knowledge, the features mentioned with the parameters were put in a dataset along with the diagnosis done by the doctor. All patient names and other identifying tags were anonymized. This process was repeated for **138 patients**. The input dataset attributes for proposed DSS are given in Table 8.

³ http://en.wikipedia.org/wiki/Rawalpindi_Institute_of_Cardiology [Last accessed on 8th December, 2014].

Table 10 Diagnosis comparison of Individual patients

Patient_ID	By doctor	Prediction by BagMOOV
1	1	1
2	1	1
3	1	0
4	1	1
5	1	1
6	0	0
7	1	1
8	0	0
9	1	1
10	1	1

Table 11 Confusion Matrix for BagMOOV

		Classified by doctor	
		Class 0 (healthy)	Class 1 (sick)
Predicted by BagMoov	Class 0 (Healthy)	36	8
	Class 1 (Sick)	13	81

These feature sets and values from the dataset were then used to write the “if–then” rules as well as training the base classifiers to be used by the proposed system. A sample of these rules is given in Table 9. Training of the classifiers is a onetime process.

Now that the classifiers were trained, in the second step DSS was used for predication of heart disease. Just like before, a patient will come to the doctor for diagnosis. The patient data was feed into the DSS and the prediction performed by the DSS was discussed by a panel of doctors in order to verify the accuracy of disease prediction. Moreover at the end of each discussion, the recommendations provided by the proposed DSS were compared with panel’s decisions in order to determine whether the two recommendations matched. The proposed DSS was evaluated on 138 patients.

Analysis of results

The prediction done by the panel of doctors for each patient was matched with prediction done by proposed framework and accuracy was calculated. This process is shown for first ten patients in Table 10. For rest of the patients, the results are shown in Appendix (see Table 14).

The confusion matrix for the 138 patients is shown in Table 11. It shows high level of agreement between the doctor’s diagnosis and proposed ensemble framework. In case of discrepancies, we identified the reasons why proposed DSS provided different decisions. In general the

Table 12 Comparison with Individual Classifiers for RIC Patients

		Class 0	Class 1	Accuracy (%)	Sensitivity (%)	Specificity (%)	F-measure (%)
NB	Class 0	24	14	71.74	48.98	84.27	61.95
	Class 1	25	75				
SVM	Class 0	37	14	81.16	75.51	84.27	79.65
	Class 1	12	75				
Linear Reg	Class 0	34	15	78.26	69.39	83.15	75.65
	Class 1	15	74				
QDA	Class 0	29	10	78.26	59.18	88.76	71.02
	Class 1	20	79				
kNN	Class 0	34	13	79.71	69.39	85.39	76.56
	Class 1	15	76				
BagMOOV	Class 0	36	8	84.78	73.47	91.01	81.30
	Class 1	13	81				

Table 13 Comparison with Ensembles for RIC Patients

		Class 0	Class 1	Accuracy (%)	Sensitivity (%)	Specificity (%)	F-measure (%)
Bagging	Class 0	36	14	80.43	73.47	84.27	78.50
	Class 1	13	75				
Adaboost	Class 0	34	13	79.71	69.39	85.39	76.56
	Class 1	15	76				
Stacking	Class 0	36	14	80.43	73.47	84.27	78.50
	Class 1	13	75				
NNE	Class 0	38	14	81.88	77.55	84.27	80.77
	Class 1	11	75				
BagMOOV	Class 0	36	8	84.78	73.47	91.01	81.30
	Class 1	13	81				

main reason behind was outlier values for some features that results in deviation pattern. For the 138 patients, the proposed DSS produced an accuracy of 84.78 % with 73.47 % sensitivity, 91.01 % specificity and 81.30 % f-measure.

Table 12 shows the accuracy, sensitivity, specificity and f-measure of proposed framework for real-time prediction against other classifiers. Highest accuracy level is achieved by the proposed DSS for all the datasets when compared to other individual classifiers.

Table 13 shows the accuracy, sensitivity, specificity and f-measure of proposed framework for real-time prediction against other ensemble models. Again, highest accuracy level is achieved for all the datasets when compared to ensembles.


These results again reflect the effectiveness of the proposed DSS. For each patient, the proposed ensemble framework can also store the state of care process such as recommendation done by doctors, patients history related to heart disease and diagnosis of disease type.

Conclusions and future work

Ensemble methods were first proposed about 10 years ago in the field of data mining and machine learning. The use of ensemble technique in the field of medical domain plays a vital role for disease prediction and classification. **Heart disease is one of the major causes of death and it should be diagnosed at early stages.** This research paper presents a novel ensemble approach using bagging with multi-objective optimized weighted voting applied on heart disease datasets in order to improve the classification and disease prediction accuracy. It is **based on five heterogeneous classifiers to achieve diversity among individual classifiers with respect to misclassified instances.** The base classifiers used are Naïve Bayes, Linear Regression, Quadratic Discriminant Analysis, Support Vector Machine and Instance based Learning. Data pre-processing is performed before model construction in order to remove anomalies in data.

Five heart disease datasets are obtained from UCI data repository to perform experimentation and results

evaluation. Different parameters are used to show the significance of results such as ANOVA statistics, p value, confusion matrices, accuracy, sensitivity, specificity and f-measure. Each parameter results in high significance of the proposed ensemble approach. The proposed ensemble classifier is compared with single classifiers as well as with ensemble classifiers. Significant results are achieved from the classifiers comparison. Also, the comparison of results exhibit that proposed ensemble approach outperforms other approaches. It is also concluded that proposed bagging method (BagMOOV) increases disease prediction accuracy.

Future research directions include enhancements of individual classifier to be used in a voting ensemble and application of the proposed algorithm on different diseases like diabetes and cancer for classification and prediction. We also plan to apply BagMOOV for multi-disease classification and compare the results with other ensemble techniques such as bagging, boosting, Adaboost, stacking, etc. 

Acknowledgments We are grateful to Rawalpindi Institute of Cardiology for their support in using the proposed DSS for research purposes only under the strict supervision of a team of medical experts and their information technology team.

Appendix

See Table 14.

Table 14 Diagnosis comparison of doctor analysis versus BagMOOV prediction

Patient_ID	Diagnosis by doctor	Prediction by BagMOOV
1	1	1
2	1	1
3	1	0
4	1	1
5	1	1
6	0	0
7	1	1
8	0	0
9	1	1
10	1	1
11	1	1
12	1	1
13	1	1
14	1	1
15	1	1
16	1	0
17	1	1
18	1	1
19	1	0
20	1	1

Table 14 continued

Patient_ID	Diagnosis by doctor	Prediction by BagMOOV
21	0	1
22	0	0
23	0	1
24	0	1
25	1	1
26	1	1
27	0	1
28	1	1
29	1	1
30	1	1
31	1	1
32	1	1
33	1	1
34	1	1
35	0	0
36	1	1
37	0	0
38	1	1
39	0	0
40	1	1
41	1	1
42	0	0
43	1	1
44	0	0
45	0	1
46	1	1
47	1	1
48	1	1
49	1	0
50	1	1
51	1	1
52	1	1
53	1	1
54	1	1
55	0	0
56	0	0
57	1	1
58	1	1
59	0	0
60	1	1
61	1	1
62	0	0
63	0	0
64	0	0
65	0	0
66	0	0
67	0	0
68	0	1

Table 14 continued

Patient_ID	Diagnosis by doctor	Prediction by BagMOOV
69	1	1
70	1	1
71	0	1
72	1	1
73	1	1
74	0	1
75	1	1
76	1	1
77	0	0
78	1	1
79	1	1
80	0	1
81	1	1
82	0	0
83	1	1
84	1	1
85	0	0
86	1	0
87	0	0
88	0	1
89	0	0
90	1	1
91	0	0
92	0	0
93	1	1
94	0	0
95	1	1
96	1	0
97	1	1
98	0	0
99	1	1
100	1	1
101	1	1
102	1	1
103	0	1
104	1	0
105	0	0
106	1	1
107	0	1
108	1	1
109	1	1
110	0	0
111	1	1
112	1	1
113	1	1
114	0	0
115	1	1
116	1	1

Table 14 continued

Patient_ID	Diagnosis by doctor	Prediction by BagMOOV
117	1	1
118	1	1
119	0	1
120	0	0
121	1	1
122	1	0
123	1	1
124	1	1
125	0	0
126	1	1
127	0	0
128	0	0
129	0	0
130	1	1
131	1	1
132	1	1
133	0	0
134	1	1
135	1	1
136	0	0
137	1	1
138	1	1

References

1. Rajkumar A, Reena GS (2010) Diagnosis of heart disease using data mining algorithm. Glob J Comput Sci Technol 10(10):38
2. Porter T, Green B (2009) Identifying diabetic patients: a data mining approach. In: Americas conference on information systems
3. Panzarasa S et al. (2010) Data mining techniques for analyzing stroke care processes. In: Proceedings of the 13th world congress on medical informatics
4. Li L, Tang H, Wu Z, Gong J, Gruidl M, Zou J Tockman M, Clark RA (2004) Data mining techniques for cancer detection using serum proteomic profiling. In: Artificial intelligence in medicine, Elsevier
5. Das R, Turkoglu I, Sengur A (2009) Effective diagnosis of heart disease through neural networks ensembles. In: Expert Systems with Applications, Elsevier, pp. 7675–7680
6. Srinivas K, Rani BK, Govrdhan A (2010) Applications of data mining techniques in healthcare and prediction of heart attacks. Int J Comput Sci Eng (IJCSE) 2:250–255
7. Shouman M, Turner T, Stocker R (2012) Using data mining techniques in heart disease diagnosis and treatment. 978-1-4673-0484-9/12, IEEE
8. Zhang L, Zhou WD (2011) Sparse ensembles using weighted combination methods based on linear programming. Pattern Recognit 44:97–106
9. Pattekari SA, Parveen A (2012) Prediction system for heart disease using Naïve Bayes. Int J Adv Computer Math Sci 3(3):290–294

10. Peter TJ, Somasundaram K (2012) An empirical study on prediction of heart disease using classification data mining techniques. In: IEEE-International conference on advances in engineering, science and management (ICAESM-2012)
11. Ghumbre S, Patil C, Ghatol A (2011) **Heart disease diagnosis using support vector machine**. In: International conference on computer science and information technology (ICCSIT') Pattaya
12. Chitra R, Seenivasagam DV (2013) **Heart disease prediction system using supervised learning classifier**. Int J Softw Eng Soft Comput 3(1):01–07
13. Chen AH, Huang SY, Hong PS, Cheng CH, Lin EJ (2011) HDPS: **heart disease prediction system**. In: Computing in cardiology
14. Jabbar MA, Chandra P, Deekshatulu BL (2012) **Heart disease prediction system using associative classification and genetic algorithm**. In: International conference on emerging trends in electrical, electronics and communication technologies-ICECIT
15. Valente G, Castellanos AL, Vanacor EG, Formisan OE (2014) Multivariate linear regression of high-dimensional fMRI data with multiple target variables. Hum brain mapp 35(2):2163–2177
16. Rizk-Jackson A, Stoffers D, Sheldon S, Kuperman J, Dale A, Goldstein J, Corey-Bloom J, Poldrack RA, Aron AR (2011) Evaluating imaging biomarkers for neurodegeneration in pre-symptomatic Huntington's disease using machine learning techniques. NeuroImage 56(2):788–796
17. Maroco J, Silva D, Rodrigues A, Guerreiro M, Santana I, Mendonça AD (2011) Data mining methods in the prediction of Dementia: **A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests**. BMC Res Notes 4(1):299
18. Helmy T, Rahman SM, Hossain MI, Abdelraheem A (2013) Non-linear heterogeneous ensemble model for permeability prediction of oil reservoirs. Arab J Sci Eng 38:1379–1395
19. Saha S, Ekbal A (2013) **Combining multiple classifiers using vote based classifier ensemble technique** for named entity recognition. Data Knowl Eng 85:15–39
20. Mokaddem S, Atmani B, Mokaddem M (2013) **Supervised feature selection for diagnosis of coronary artery disease based on genetic algorithm**. In: First international conference on computational science and engineering (CSE-2013)
21. Kohavi R, John GH (1997) **Wrappers for feature subset selection**. Artif Intell 97(1):273–324
22. Patil RR (2014) **Heart disease prediction system using Naive Bayes and Jelinek-mercer smoothing**. Int J Adv Res Comput Commun Eng
23. Palaniappan S, Awang R (2008) Intelligent heart disease prediction system using data mining techniques. In: International conference on computer system and applications. AICCSA, pp 108–115
24. Mehra A (2003) Statistical sampling and regression: simple linear regression. PreMBA analytical methods. Columbia Business School and Columbia University
25. Weiss SM, Kulikowski CA (1991) Computer systems that learn: **classification and prediction methods from statistics, neural nets, machine learning, and expert systems**. Morgan Kaufman, San Mateo
26. STAT55-Data mining (2014) The Pennsylvania State University
27. Uguroglu S, Carbonell J, Doyle M, Biederman R (2012) Cost-sensitive risk stratification in the diagnosis of heart disease. In: Proceedings of the twenty-fourth innovative applications of artificial intelligence conference
28. Breiman L (1994) Bagging Predictors, Technical Report 421, Department of Statistics, University of California, Berkeley
29. Jain M, Dua P, Lukiw WJ (2013) Data adaptive rule-based classification system for Alzheimer classification. J Comput Sci Syst Biol 6:291–297
30. Peter TJ, Somasundaram K (2012) An empirical study on prediction of heart disease using classification data mining techniques. In: IEEE-international conference on advances in engineering, science and management
31. Tu MC, Shin D, Shin D (2009) **Effective diagnosis of heart disease through Bagging approach**. In: 2nd international conference on biomedical engineering and informatics
32. Pai P, Li L, Hung W (2014) Using ADABOOST and rough set theory for Debris flow disaster. Water Resour Manag 28(4):1143–1155
33. Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning. Data mining, inference and prediction, 2nd edn. Springer series in statistics
34. BLA (2009) Sensitivity, specificity, accuracy and the relationship between them. Bioinformatics
35. Palaniappan S, Awang R (2008) Intelligent heart disease prediction system using data mining techniques. IJCSNS Int J Comput Sci Netw Secur, 8(8)
36. Gelman A (2008) Variance, analysis of. The new Palgrave dictionary of economics, 2nd edn. Palgrave Macmillan, Basingstoke, Hampshire New York
37. Yuan G, Ho C, Lin C (2012) Recent advances of large-scale linear classification. Proc IEEE 100(9):2584–2603
38. Shouman M, Turner T, Stocker R (2011) Using decision tree for diagnosing heart disease patients. In: Proceedings of the 9th Australasian data mining conference, Ballarat, Australia
39. Tu MC, Shin D et al (2009) **Effective diagnosis of heart disease through bagging approach**. In: 2nd international conference on biomedical engineering and informatics. IEEE, pp 1–4
40. Shouman M, Turner T, Stocker R (2013) Integrating clustering with different data mining techniques in the diagnosis of heart disease. J Comput Sci Eng 20(1)
41. Shouman M, Turner T, Stocker R (2012) **Integrating Naive Bayes and K-means clustering with different initial centroid selection methods in the diagnosis of heart disease patients**. Glob J Comput Sci Technol 125–137
42. Chaurasia V, Pal S (2013) Early prediction of heart diseases using data mining techniques. Caribb J Sci Technol 1:208–217
43. Sunday NA, Latha PP (2013) Performance analysis of classification data mining techniques over heart disease database. Int J Eng Sci Adv Technol 2(3):470–478
44. Soni J, Ansari U, Sharma D (2011) Intelligent and effective heart disease prediction system using weighted associative classifiers. Int J Computer Sci Eng (IJCSE) 3(6):2385–2392