A decision support system for heart disease prediction based upon machine learning

Article *in* Journal of Reliable Intelligent Environments · September 2021

DOI: 10.1007/s40860-021-00133-6

4 authors, including:

Pooja Rani
Maharishi Markandeshwar University, Mullana
10 PUBLICATIONS **56** CITATIONS
SEE PROFILE

Rajneesh Kumar Gujral
Maharishi Markandeshwar Deemed University, Mullana
92 PUBLICATIONS **498** CITATIONS
SEE PROFILE

Nada Sid Ahmed
University of Hail
7 PUBLICATIONS **28** CITATIONS
SEE PROFILE

Journal of Reliable Intelligent Environments
https://doi.org/10.1007/s40860-021-00133-6

ORIGINAL ARTICLE

A decision support system for heart disease prediction based upon machine learning

Pooja Rani[1] · Rajneesh Kumar[2] · Nada M. O. Sid Ahmed[3] · Anurag Jain[4]

Received: 5 November 2020 / Accepted: 2 January 2021

Abstract
Detection of heart disease through early-stage symptoms is a great challenge in the current world scenario. If not diagnosed timely then this may become the cause of death. In developing countries where heart specialist doctors are not available in remote, semi-urban, and rural areas; an accurate decision support system can play a vital role in early-stage detection of heart disease. In this paper, the authors have proposed a hybrid decision support system that can assist in the early detection of heart disease based on the clinical parameters of the patient. Authors have used multivariate imputation by chained equations algorithm to handle the missing values. A hybridized feature selection algorithm combining the Genetic Algorithm (GA) and recursive feature elimination has been used for the selection of suitable features from the available dataset. Further for pre-processing of data, SMOTE (Synthetic Minority Oversampling Technique) and standard scalar methods have been used. In the last step of the development of the proposed hybrid system, authors have used support vector machine, naive bayes, logistic regression, random forest, and adaboost classifiers. It has been found that the system has given the most accurate results with random forest classifier. The proposed hybrid system was tested in the simulation environment developed using Python. It was tested on the Cleveland heart disease dataset available at UCI (University of California, Irvine) machine learning repository. It has achieved an accuracy of 86.6%, which is superior to some of the existing heart disease prediction systems found in the literature.

Keywords Decision support system · Clinical data · Heart disease · Machine learning

1 Introduction

Heart is the most important part of the human body which is responsible for pumping oxygen-rich blood to other body parts through a network of arteries and veins. Any type of disorder that affects our heart is heart disease [1]. According to the world health organization report published in 2019, around 17 million people die every year worldwide due to heart disease [2]. There are various types of heart diseases such as coronary artery disease, congenital heart disease, arrhythmia, etc. The patient suffering from heart disease has various symptoms such as chest pain, dizzy sensations, and deep sweating. Smoking, high blood pressure, diabetes, obesity, etc. are the main reasons behind heart disease [3]. Invasive methods of diagnosing the disease are expensive and painful. Therefore, there is a need for a technique that can diagnose heart disease in a non-invasive manner at less cost.

Machine learning techniques can be used to design a decision support system to detect heart disease through clinical data easily and cost-effective manner. This kind of

✉ Rajneesh Kumar
drrajneeshgujral@mmumullana.org

Pooja Rani
pooja.rani@mmumullana.org

Nada M. O. Sid Ahmed
n.sidahmed@uoh.edu.sa

Anurag Jain
anurag.jain@ddn.upes.ac.in

1 MMEC (Research Scholar), MMICT&BM (A.P.), Maharishi Markandeshwar (Deemed To Be University), Mullana, Ambala, Haryana 133207, India

2 Department of Computer Engineering, MMEC, Maharishi Markandeshwar (Deemed To Be University), Mullana, Ambala, Haryana 133207, India

3 College of Computer Science and Engineering, University of Ha'il, Ha'il, Kingdom of Saudi Arabia

4 Virtualization Department, School of Computer Science, University of Petroleum and Energy Studies, Dehradun, India

Published online: 25 January 2021

decision-making system can assist doctors to diagnose the disease at an early stage [4]. We can convert this type of decision support system into a medical chatbot system which can be utilized by the patient himself for early detection of heart disease. Chatbot system will take some clinical information as input and advise patients accordingly based on input. However, these medical chatbot systems must be regulated by regulations to ensure safety and reliability. Therefore it is required to carry out risk management activities to minimize safety risks. ISO 14971 reference standard is adopted for risk management [5, 6]. These medical chatbot systems help patients self-monitor their health status [7].

Clinical decision support systems assist physicians in the diagnosis of the disease and providing appropriate treatment. They should be evaluated to ensure that these systems perform their intended work safely and efficiently [8].

Mobile Health technologies can also be used to embed decision support systems with mobile devices. Mobile Health technologies collect real-time data from patients and provide more efficient health services. It helps improve patient monitoring without visiting the health center [9, 10].

Better monitoring of heart patients can help in reducing the mortality rate. Generally, people approach a heart specialist at a very late stage [11]. The major motivation behind proposing this decision support system is to develop a system, which can be used by any literate person in the absence of a doctor to diagnose heart disease at an early stage. Even this kind of system may assist the doctor while taking the decision. The beauty of this system is that it will depend entirely on clinical data that does not require a heart specialist doctor. A variety of heart disease decision support systems have been found in the literature with varying degrees of accuracy. But most of the researchers have not considered the issue of missing values and feature selection approach collectively. In the proposed hybrid decision support system for heart disease prediction, authors have not only handled the issue of missing value but also handled the feature selection issue efficiently. The proposed hybrid system has tested in the simulation environment developed using Python.

The remaining sections of this paper are organized as follows: Sect. 2 contains a review of the various hybrid heart disease decision support system proposed by different researchers. Section 3 includes a detailed description of the proposed hybrid system. Section 4 includes results and discussions. Conclusion and future scope are written in Sect. 5.

2 Literature review

Different researchers had proposed different decision support systems to predict heart disease making use of various machine learning algorithms. Different decision support systems proposed by different researchers in predicting heart disease are discussed in this section.

Bashir et al. [12] proposed a system for predicting heart disease using an ensemble mechanism making use of five classifiers decision tree induction using information gain, naive bayes, memory-based learner, support vector machine, and decision tree induction using Gini Index. Feature selection was not performed because datasets used by authors contain only relevant attributes. Data preprocessing was performed to remove missing values and outliers.

Olaniyi and Oyedotun [13] developed a heart disease diagnosis system with support vector machine and multilayer perceptron neural network algorithms.

Verma et al. [14] proposed a system for predicting heart disease using a hybrid approach making use of four classifiers multi-layer perceptron (MLP), Fuzzy unordered rule induction algorithm (FURIA), Multinomial logistic regression model (MLR), C4.5 (decision tree algorithm). Feature selection was performed using correlation-based feature subset selection (CFS) combined with Particle Swarm Optimization (PSO). CFS method selects attributes based upon the correlation between attributes. The number of features was reduced from 25 to 5 using feature selection. After performing feature selection with CFS and PSO, incorrectly assigned data points were removed from the data using the *K*-means clustering algorithm. Verma and Srivastava [15] proposed a system for diagnosing coronary artery disease using a neural network model.

Miranda et al. [16] developed a decision support system to detect the risk of cardiovascular disease. Firstly predictor attributes were selected, and then data cleaning was performed to deal with incomplete and inaccurate parts of the data. Data generalization was performed by the conversion of numerical data into categorical data. The classification was done using the naive bayes Classifier.

Wiharto et al. [17] performed the classification of heart disease using the C4.5 algorithm on the data available on UCI repository. The system predicted the output as healthy or level of sickness (level 1, level 2, level 3 and level 4). Dimensionality reduction was done by selecting the relevant features using Information Gain (IG) method. An information gain measure how effective is an attribute to classify data. After oversampling and feature selection, data was used to train the system using the C4.5 algorithm. Using SMOTE and feature selection increased the performance of C4.5.

Jabbar et al. [18] in their proposed system used the random forest to perform heart disease prediction. The chi-square method was used for feature selection to select relevant attributes. The approach proposed by the authors provided better accuracy than the decision tree.

Kim and Kang [19] used a neural network to develop a system to diagnose heart disease. Sensitivity analysis

of features was used to detect the features that were more important for prediction. Features with higher sensitivity were more important features. After the selection of relevant features, correlated features were found by analyzing the change in sensitivity of features with a change in the value of one feature. If the change in the value of one feature changes the sensitivity of another feature more than the average change in sensitivity of all features, it means two features are correlated.

Arabasadi et al. [20] developed a system for predicting heart disease using a neural network optimized by genetic algorithm. It increased accuracy over the classic neural network. Liu et al. [21] developed a system for predicting heart disease using the C4.5 algorithm. Boosting was applied to increase the performance of the system. A hybrid system for diagnosing coronary artery disease was proposed using C4.5, multilayer perceptron, and naive bayes classifier [22].

David and Belcy [23] used random forest, decision tree, and naive bayes classifiers to predict heart disease. Random Forest provided more accurate predictions than decision tree and naive bayes. Haq et al. [24] combined different feature selection algorithms with different classifiers. Data preprocessing was performed using the removal of missing values and using standard and min–max scalar. For selecting important features three algorithms of feature selection were used. The minimal redundancy maximal relevance feature selection algorithm identifies relevant features and removes the duplicate features. Relief feature selection Algorithm select features based upon the weights assigned to features. Least absolute shrinkage and selection operator select features by updating coefficients and removing features whose coefficients become zero. The prediction was done using six machine learning algorithms logistic regression, support vector machine, naive bayes, artificial neural network, decision tree, and *K*-nearest neighbor.

Malav and Kadam [25] performed prediction of heart disease using ANN (artificial neural network) and *K*-means. *K*-means performed the clustering and provided the input to ANN. Poornima and Gladis [26] developed a heart disease prediction system using a hybrid approach. Initially, missing values were removed during the preprocessing of data. The dimensionality of data was reduced using OLPP (orthogonal local preserving projection). The classification was performed using a neural network. The network was trained using LM (Levenberg–Marquardt) and GSO (group search optimization) for setting the weights.

Khourdifi and Bahaj [27] developed a system for heart disease prediction using support vector machine, *K*-nearest neighbor, multilayer perception, random forest, and naive bayes classifiers optimized by ant colony optimization and particle swarm optimization. K-nearest neighbor and random forest provided the highest accuracy.

Ali et al. [28] had selected relevant features using the chi-square statistical method. Selected features were applied to a deep neural network to perform classification by training the network. Network configuration was optimized by the use of an exhaustive grid search method.

Mohan et al. [29] developed a heart disease prediction system using random forest and linear methods. Ali et al. [30] developed a system for the prediction of heart failure using two models of SVM (support vector machine). One model was used for selecting features and another model was used for prediction.70% of data was used for training and 30% of data was used for testing. The model for selecting features was L1 regularized linear SVM. The prediction model was L2 regularized RBF (radial basis function) kernel SVM. The optimization of the two models was done by optimizing the hyperparameters of SVM.

Verma and Mathur [31] developed a system to predict heart disease using deep learning. Relevant features were selected using correlation and cuckoo search algorithms.

Mienye et al. [32] developed an artificial neural network model for heart disease diagnosis. Performance of ANN was optimized using Sparse autoencoder.

Latha and Jeeva [2] developed a model to diagnose heart disease by combining results of Naive Bayes, multilayer perceptron, random forest and Bayes network using majority voting. Terrada et al. [33] developed a system to diagnose heart disease using ANN, AdaBoost, and Decision Tree.

Paragliola and Coronato [34] proposed a model to identify the risk of cardiac events for patients suffering from hypertension. Authors developed a hybrid model using long short-term memory network and convolutional neural network and provided ECG signals as input. The system used time-series signals for early prediction of an increase in hypertension.

Tama et al. [35] developed an ensemble model for heart disease diagnosis. Random forest, gradient boosting, and extreme gradient boosting classifiers were used in constructing the ensemble model.

3 Materials and methods

3.1 Methodology

In this research, the authors have proposed a hybrid decision support system for the prediction of heart disease. This hybrid system consists of three stages: data collection, data pre-processing, and model construction. In the pre-processing stage, missing values are imputed, feature selection is done, feature scaling is performed and class balancing is done. Missing values were imputed using MICE (multivariate imputation by chained equations) algorithm. After that Feature selection is performed using a hybrid GA and RFE

approach. Coefficient of all features is brought to the same value using standard scalar ensuring that each feature has the mean 0 and standard deviation 1. In the dataset, 164 instances are belonging to class 0, and 139 instances belonging to class 1. Class balancing is performed using SMOTE.

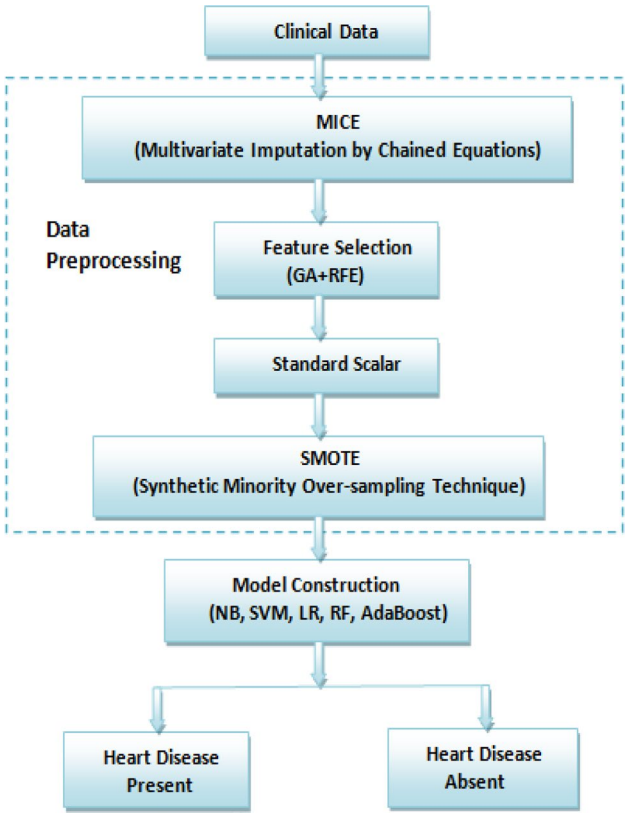

Fig. 1 Proposed hybrid heart disease prediction system

It creates synthetic samples of minor classes resulting in an equal number of samples of both classes [36]. Classification is performed on selected features using NB, SVM, LR, RF, and AdaBoost classifier. Finally, the classifier predicts that a person is having heart disease or not. The methodology of proposed hybrid system for heart disease prediction is shown in Fig. 1.

3.2 Dataset

Cleveland Heart disease dataset obtained from UCI (University of California, Irvine) repository was used for performing the experiments. This dataset is having 14 features out of which eight are categorical features and six are numeric features. Features of the dataset and description of the features are shown in Table 1. Data of patients having age from 29 to 77 are collected in this dataset. Chest pain is a symptom of heart disease. There are four types of chest pain: typical angina, atypical angina, non-angina pain and asymptotic. Feature RBP has a value of resting blood pressure of the patient. SCHOL indicates the cholesterol level of the patient. Level of fasting blood sugar is indicated in FABS. If sugar is above 120 mg/dl then 1 is stored in this feature, otherwise 0 is stored. RECR has electrocardiographic results and the maximum heart rate of the patient is stored in MHR. EIGA has a value of 1 if a person suffers from exercise-induced angina; otherwise 0.ST depression induced by exercise is stored in STD which has possible values of upsloping, downsloping and flat indicated by 0, 1 and 2.SPE is slope of peak exercise. NMVCF contains information about how many major vessels are colored by fluoroscopy. TARG attribute indicates whether a person is suffering from heart disease or not. In this feature, there are five possible values

Table 1 Features of Cleveland heart disease dataset

Feature name	Feature code	Description
Age	AG	Age between 29 and 77
Sex	SX	Male: 1, female: 0
Type of chest pain	CP	Typical angina: 1, atypical angina: 2 non-angina pain: 3, asymptomatic: 4
Resting blood pressure	RBP	Between 94 mm Hg and 200 mm Hg
Serum cholesterol	SCHOL	Between 126 mg/dl and 564 mg/dl
Fasting blood sugar	FABS	FBSR > 120 mg/dl (true:1, false: 0)
Resting electrocardiographic results	RECR	Normal: 0, ST-T wave abnormality: 1, Hypertrophy: 2)
Maximum heart rate achieved	HR	Between 71 and 202
Exercise-induced angina	EIAG	Yes: 1, No: 0
ST depression induced by exercise relative to rest	STD	Up sloping: 1, Flat: 2, downsloping: 3
The slope of the peak exercise ST segment	SPE	Between 0 and 6.2
Number of major vessels (0–3) colored by fluoroscopy	NMVCF	Between 0 and 3
Thallium	THALM	Normal: 3, fixed defect: 6, reversible defect: 7
Target	TARG	Heart disease present: 1, heart disease absent: 0

0 for the absence of heart disease and 1 to 4 for different levels of the disease. Levels 1–4 are merged to indicate the presence of disease. Dataset has 6 instances having missing values. There are four missing values in NMVCF feature and two missing values in THALM feature [37].

3.3 Multivariate imputation by chained equations algorithm (MICE)

Cleveland dataset has six missing values. These Missing values were imputed using MICE algorithm. This algorithm performs imputation multiple times. It assumes that data is missing randomly. In this method, a regression model is used to predict the value of the missing attribute from the remaining attributes of the dataset [38]. The steps of this algorithm are shown in Fig. 2.

3.4 Feature selection

A hybrid approach combining GA and RFE was used for feature selection. Eight features SX, CP, RECR, EIAG, STD, SPE, NMVCF and THALM were selected using this approach.

A genetic algorithm (GA) is based upon the idea behind natural selection. It generates multiple solutions in a single generation. Each solution is known as a chromosome. The set of solutions in a single generation is known as the population. This algorithm iterates over multiple generations to generate a better solution. At each step of the algorithm, genetic operators are applied to chromosomes from the previous generation to create the next generation. Selection, crossover, and mutation are different types of genetic operators used. Selection operator selects the best individuals from each generation. A fitness function evaluates each individual to determine how fit it is compared to other individuals in the population. Chromosomes selected using a selection operator is placed in the mating pool and participate in the production of the next generation. Crossover operator combines individuals placed in mating pools to create better individuals for the next generation. There are different types of crossover operators such as single point, two-point, and multipoint crossover. Individuals in the next generation will be similar to the previous generation if diversity is not brought into the population. This diversity is introduced using mutation operator which makes random changes in the individuals [20]. The steps of the genetic algorithm are shown in Fig. 3.

In performed experiments, authors have used the number of generations as a stopping criterion. Total 18 generations have been used. 80 chromosomes were selected in each generation. 40 best chromosomes were selected in each generation and passed to the next generation. 40 chromosomes were selected randomly. A crossover rate of 5 and a mutation rate of 0.05 was used. Negated Mean square error of the chromosome was used as an objective function to calculate the fitness of the chromosomes. Recursive Feature Elimination (RFE) algorithm recursively removes irrelevant features. A classifier is trained on the training data and the

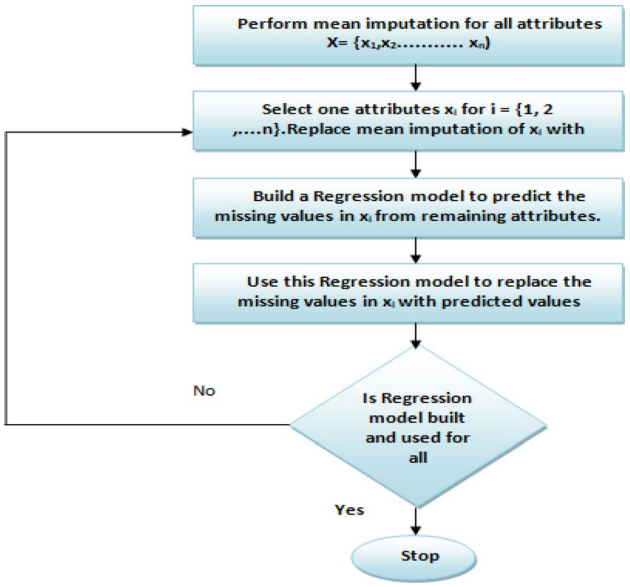

Fig. 2 Multiple imputation chained equations algorithm

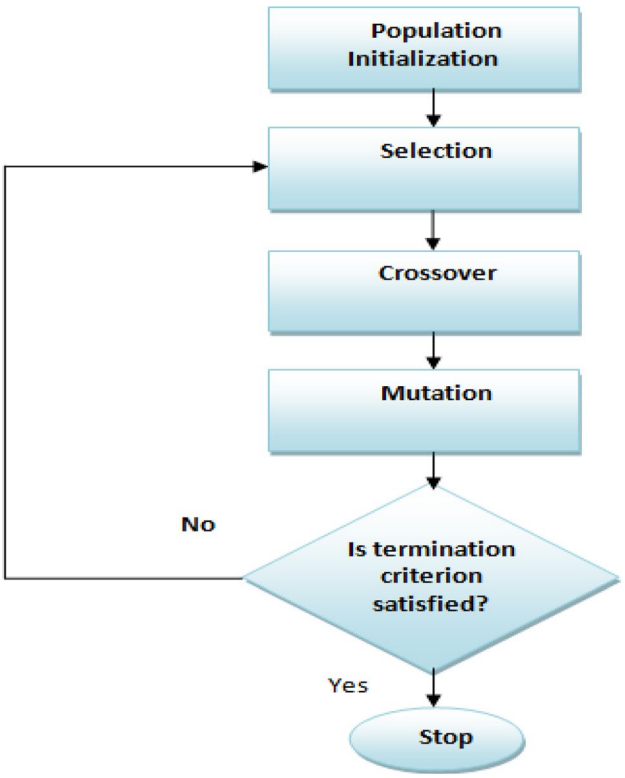

Fig. 3 Genetic algorithm

importance of features is calculated. At each step of the algorithm weakest features are removed and the model is again trained with the remaining subset of features. These steps are iteratively performed until the desired number of features is achieved. The number of features to be retained is passed as a parameter to the algorithm [39]. The working of RFE algorithm is shown in Fig. 4.

3.5 Classification algorithms

Classification algorithms used for performing prediction are discussed in this section.

3.5.1 Naive Bayes (NB)

It is a type of probabilistic classifier that uses bayes theorem. It can perform classification effectively in various types of problems such as categorization of documents, spam filtering, and disease diagnosis. The underlying assumption behind this algorithm is the independence between features that participates in the prediction process. It calculates the posterior probability of the class using the following equation:

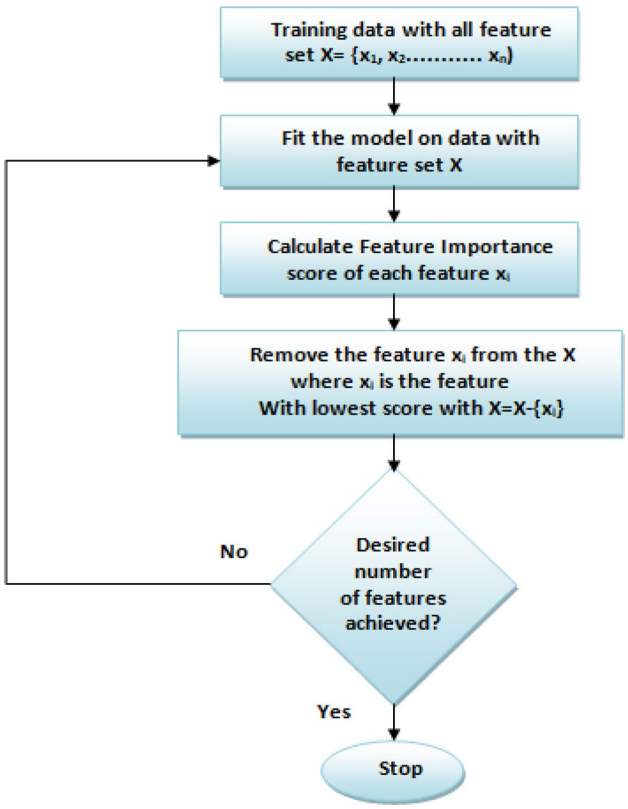

Fig. 4 Recursive Feature Elimination algorithm

$$P(c|I) = P(I|c)P(c)/P(I).$$

Posterior probability denotes the probability of occurring class c with input I. In the above equation, $P(c)$ is prior class probability and $P(I)$ is prior feature probability. $P(I|c)$ is the likelihood which indicates the probability of occurring feature I with class c. This algorithm can also be used for multi-classification problems [40].

3.5.2 Support vector machine (SVM)

It creates a hyperplane to perform classification so that all samples belonging to one class will lie on one side of hyperplane and samples belonging to another class will lie on another side. It optimizes the hyperplane to ensure the maximum distance between the two classes. Support vectors are those data points of classes that are nearest to hyperplane [41].

Hyperplane can be created as given in the following equation:

$$H_0 : w^T x + b = 0.$$

Two more hyperplanes H_1 and H_2 are created in parallel to the constructed hyperplane as given in the following equations:

$$H_1 : \quad w^T x + b = -1,$$

$$H_2 : \quad w^T x + b = 1.$$

Hyperplane should satisfy the constraints given by following equations for each input vector I_j:

$$wI_j + b \geq +1 \quad \text{for } I_j \text{ having class 1,}$$

and,

$$wI_j + b \geq -1 \quad \text{for } I_j \text{ having class 0.}$$

Hyperplane separating two classes in SVM is shown in Fig. 5.

3.5.3 Logistic regression (LR)

This algorithm can be used for binary classification problems to predict the value of a variable Y which can have two possible values 0 or 1. It can also be used for multi-classification problems when Y has more than two possible values. Logistic regression equation given below calculates the probability by which input X should be classified as class 1:

$$P(X) = \frac{\exp\left(\beta_0 + \beta_1 X\right)}{1 + \exp\left(\beta_0 + \beta_1 X\right)}.$$

Here β_0 is bias and β_1 is the weight that is multiplied by input X [41].

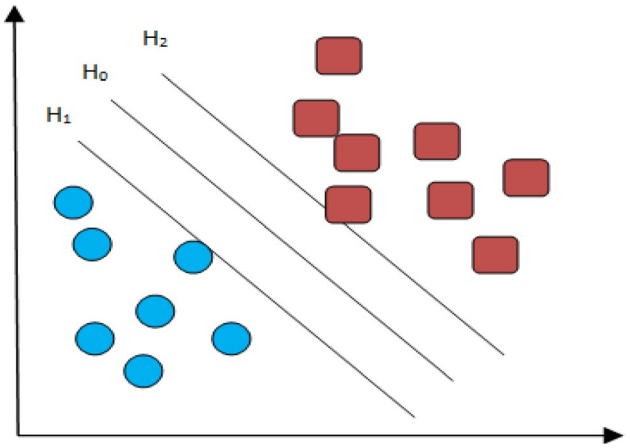

Fig. 5 Hyperplane separating two classes in SVM

3.5.4 Random forest

Random Forest uses the concept of bagging to combine several decision trees to increase prediction capability. In bagging, individual learners are trained independently. In it, multiple samples of data are generated randomly from the original dataset with replacement, and each decision tree is trained on different samples of data. Features are also selected randomly during tree construction. Prediction generated by multiple trees is combined using a majority vote [18]. The working of Random forest is shown in Fig. 6.

Random forest can be tuned for increased accuracy by optimizing parameters such as the number of estimators, minimum size of node and number of features used to split node, etc. In this research, authors had done hyperparameter tuning of Random Forest using RandomizedSearchCV() method.

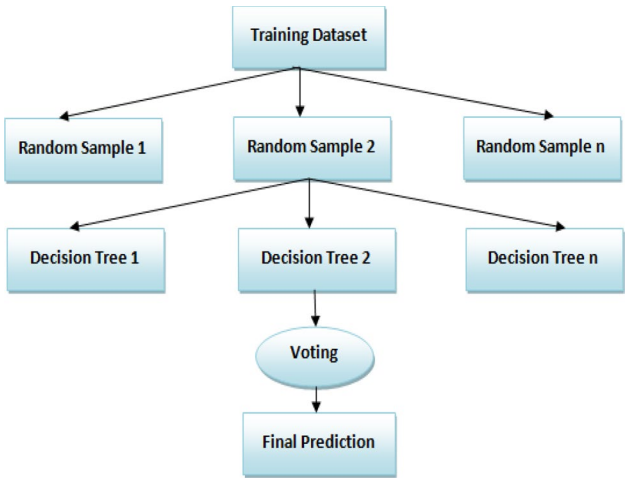

Fig. 6 Random Forest algorithm

3.5.5 Adaboost

Adaboost is known as adaptive boosting algorithm. It uses the concept of boosting which is an ensemble technique used to increase the performance of weak learners. This algorithm firstly trains the classifier on the original dataset. Then multiple copies of the classifier are trained and each copy tries to correct the error occurring from the previous copy.

Each copy of the classifier is trained on a different subset of data. Multiple subsets of dataset are created by assigning weights to data items. An incorrectly classified instance has a higher chance of selecting for the next subset because it is assigned a higher weight. In this way, multiple models are trained one after another sequentially. After that, these weak classifiers are combined using a cost function to produce a strong classifier. Classifiers with higher accuracy are given more weightage in the final prediction. Weak classifier to which boosting is to be applied can be passed as a parameter to Adaboost algorithm. The default classifier used for boosting in Adaboost is decision tree [2]. The working of AdaBoost algorithm is shown in Fig. 7.

3.6 Evaluation parameters

The performance of classifiers was evaluated on the scale of accuracy, sensitivity, specificity, precision, and recall [32, 41]. If a person suffering from the disease is predicted to be a heart disease patient by the system then it is a true positive, otherwise, it is a false negative. Similarly, if a healthy person is predicted to be disease-free then it is a true negative, otherwise it is a false positive.

Accuracy It is a performance parameter that measures the ability of the system to make correct predictions,

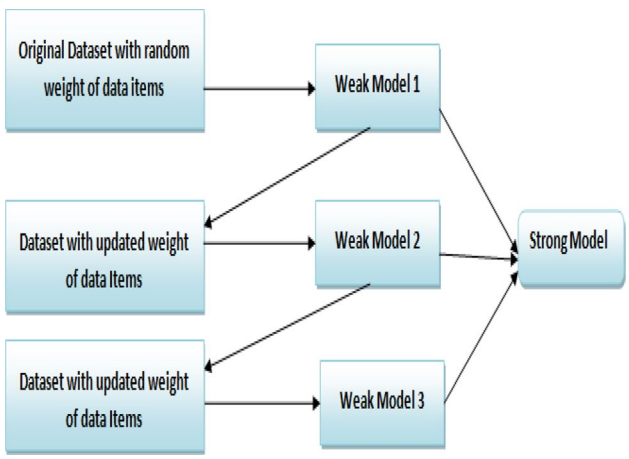

Fig. 7 Adaboost algorithm

$$\text{Accuracy} = \left(\frac{\text{Correct predictions}}{\text{Total predictions}}\right) \times 100.$$

Sensitivity It is a performance parameter that measures the ability of the system to make correct positive predictions,

$$\text{Sensitivity} = \left(\frac{\text{True positives}}{\text{True positives} + \text{false negatives}}\right) \times 100.$$

Specificity It is a performance parameter that measures the ability of the system to make correct negative predictions,

$$\text{Specificity} = \left(\frac{\text{True negatives}}{\text{True negatives} + \text{false positives}}\right) \times 100.$$

Precision Precision measures the capability of a system to produce only relevant results,

$$\text{Precision} = \left(\frac{\text{True positives}}{\text{True positives} + \text{false positives}}\right) \times 100.$$

F-Measure F-Measure combines results of precision and sensitivity using harmonic mean,

$$\text{F - Measure} = 2 \times \frac{\text{Sensitivity} \times \text{precision}}{\text{Sensitivity} + \text{precision}}.$$

4 Results and discussions

Several classification algorithms including Naive Bayes, support vector machine, logistic regression, Random Forest and Adaboost were used to diagnose heart disease in patients. Cleveland dataset from UCI was used to perform experiments. Heart disease was diagnosed using several medical parameters available in the dataset. These parameters were used to perform classification with class 1 indicating that the person has a disease and class 0 indicating that person is disease-free. Dataset was having missing values in 6 instances. These values were imputed using MICE algorithm. Application of this algorithm resulted in a complete dataset with no instance having missing value. System performance was measured on the scale of accuracy, sensitivity, specificity, precision, and F-measure. Results were validated using K-fold cross-validation method with $K = 10$. In this method partitioning of the dataset is done into k groups. Performance of the model is evaluated using $K - 1$ groups for the training of model and one group for testing of model. These steps of evaluating the model are repeated K times each time taking different training and testing groups.

Table 2 Performance of classifiers on full feature set

Classifier	Accuracy	Sensitivity	Specificity	Precision	F-Measure
NB	84.79	80.57	88.41	85.49	82.96
SVM	79.50	74.82	83.53	79.38	77.03
LR	83.80	79.13	87.80	84.61	81.78
RF	83.83	77.69	89.02	85.71	81.50
Adaboost	82.12	79.85	84.14	81.02	80.43

Table 3 Performance improvement using scaling

Classifier	Accuracy	Sensitivity	Specificity	Precision	F-Measure
NB	84.79	80.57	88.41	85.49	82.96
SVM	84.79	79.13	89.63	86.61	82.70
LR	82.50	78.41	85.97	82.57	80.44
RF	83.83	77.69	89.02	85.71	81.50
Adaboost	82.12	79.85	84.14	81.02	80.43

4.1 Performance of classifiers with all features

Firstly the experiments were performed on all features of the dataset without applying any kind of pre-processing or feature selection. The performance of classifiers on the full feature set is shown in Table 2. NB classifier provided the highest performance on the full feature set, whereas SVM provided the lowest performance.

4.2 Performance improvement using scaling

The performances of classifiers were again analyzed after applying the pre-processing technique of scaling. Standard scalar technique was applied to the input dataset. Scaling resulted in a change in the performance of classifiers. Performance of some classifiers increased, whereas the performance of some classifiers decreased. NB, RF, and Adaboost resulted in no change in performance.

Accuracy of SVM increased by 6.66%, whereas the accuracy of LR declined by 1.55%. It resulted in a 5.76% increase in sensitivity, 7.30% increase in specificity, 9.10% increase in precision and 7.36% increase in F-Measure. Results indicate that scaling has a very positive impact on SVM whereas it does not have a positive impact on the performance of other classifiers. The impact of scaling on the performance of classifiers is shown in Table 3.

4.3 Performance improvement using class balancing

After applying scaling, performance of classifiers were further improved by balancing the classes. SMOTE algorithm was applied for class balancing.

Table 4 Performance improvement using class balancing

Classifier	Accuracy	Sensitivity	Specificity	Precision	F-measure
NB	85.07	82.92	87.19	86.62	84.7
SVM	84.16	81.09	87.19	86.36	83.64
LR	83.24	80.48	85.97	85.16	82.75
RF	83.85	81.70	85.97	85.35	83.48
Adaboost	82.34	79.26	85.36	84.41	81.76

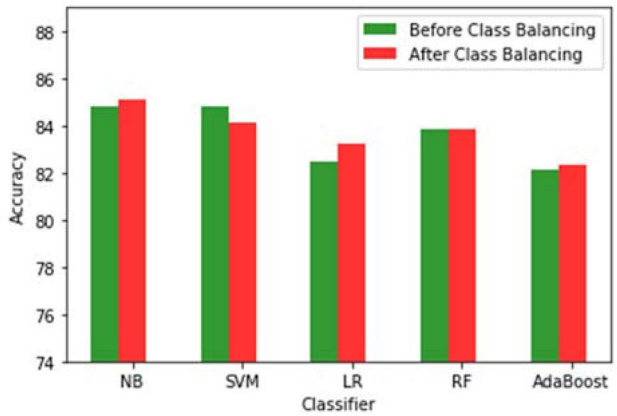

Fig. 8 Increase in accuracy of classifiers using class balancing

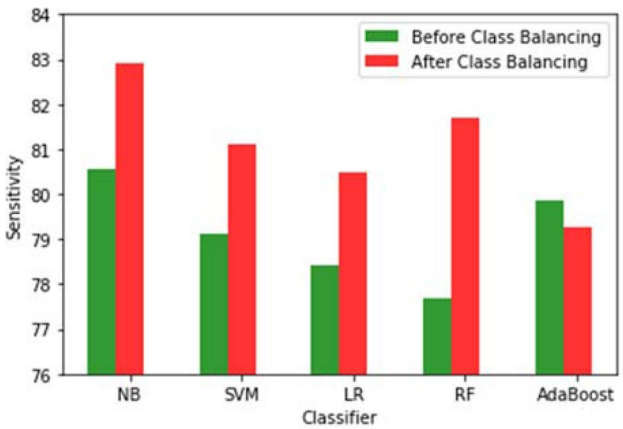

Fig. 9 Increase in sensitivity of classifiers using class balancing

It resulted in 164 instances of both class 0 and class 1. Impact of class balancing on the performance of classifiers is shown in Table 4. An increase in accuracy, sensitivity, specificity, precision, and F-Measure of classifiers with class balancing is shown in Figs. 8, 9, 10, 11 and 12 respectively.

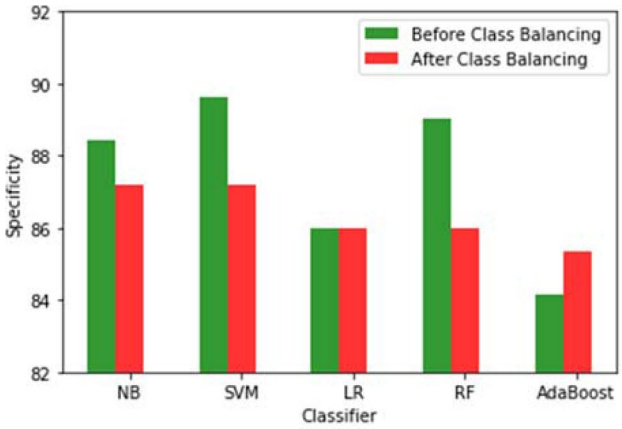

Fig. 10 Increase in specificity of classifiers using class balancing

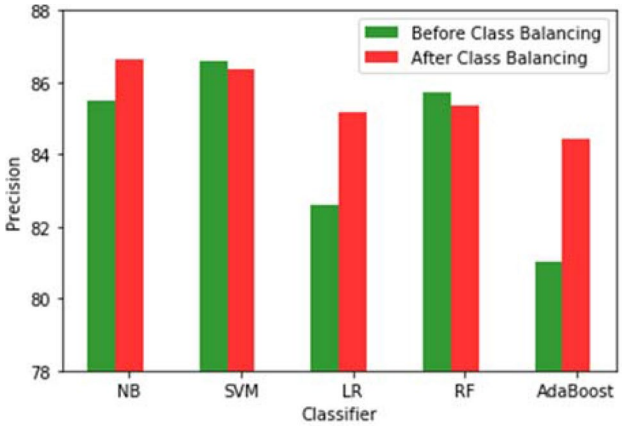

Fig. 11 Increase in precision of classifiers using class balancing

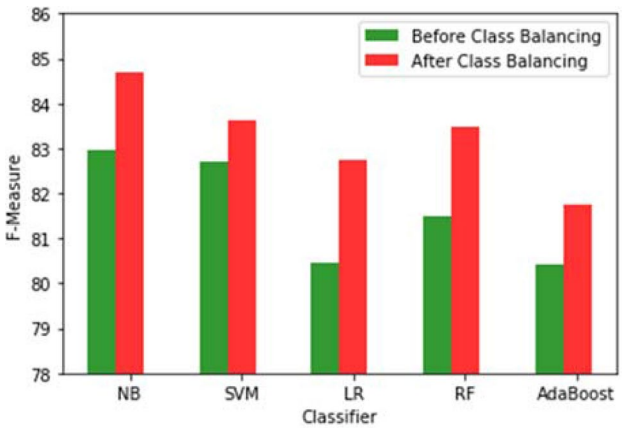

Fig. 12 Increase in F-measure of classifiers using class balancing

Table 5 Performance improvement using feature selection

Classifier	Accuracy	Sensitivity	Specificity	Precision	F-Measure
NB	83.55	82.31	84.75	84.37	83.33
SVM	84.46	81.09	87.80	86.92	83.91
LR	85.07	82.92	87.19	86.62	84.73
RF	86.60	84.14	89.02	88.46	86.25
Adaboost	86.59	83.53	89.63	88.96	86.16

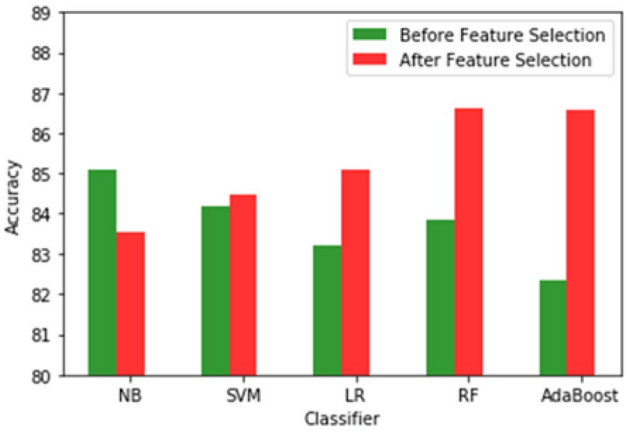

Fig. 13 Increase in accuracy of classifiers using feature selection

4.4 Performance improvement using feature selection

Accuracy of classifiers was further improved using feature selection. Hybrid GARFE algorithm was applied for feature selection. This hybrid mechanism resulted in the selection of eight features out of thirteen features. As shown in Table 5, Feature selection had improved the performance of all classifiers except NB. SVM accuracy increased by only 0.33%. LR accuracy increased by 2.19%. RF accuracy increased by 3.27%. Maximum increase of 5.16% was observed in the accuracy of Adaboost. Best results were achieved with Adaboost and RF. Sensitivity of Adaboost increased by 5.23%, specificity increased by 5%, precision increased by 4% and F-Measure increased by 5.38%. Sensitivity of RF increased by 2.98%, specificity increased by 3.54%, precision increased by 3.64% and F-Measure increased by 3.31%. An improvement in accuracy, sensitivity, specificity, precision, and F-Measure of classifiers with feature selection is shown in Figs. 13,14,15,16 and 17, respectively.

Results indicate that Random Forest provided the highest performance in combination with MICE, GARFE, Scaling and SMOTE. Dimensionality reduction using Feature selection helped in improving the performance of RF to a large extent.

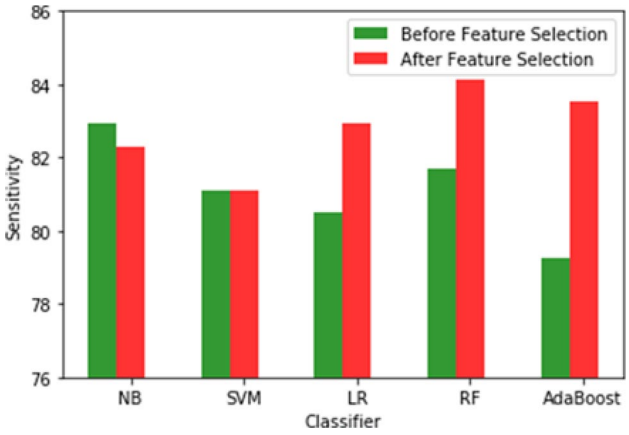

Fig. 14 Increase in sensitivity of classifiers using feature selection

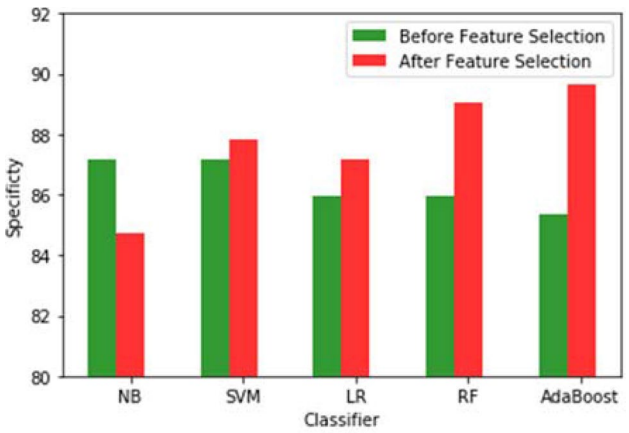

Fig. 15 Increase in the specificity of classifiers using feature selection

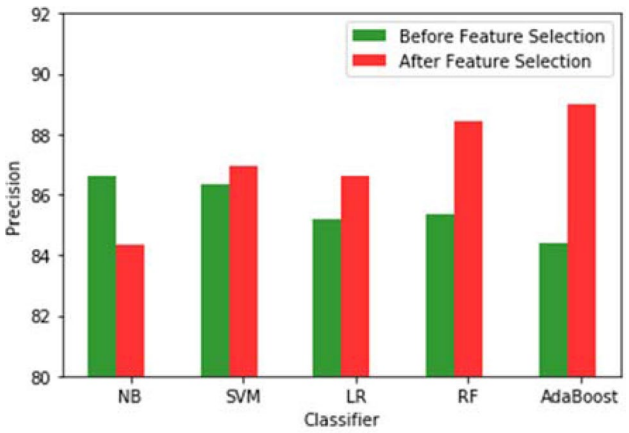

Fig. 16 Increase in the precision of classifiers using feature selection

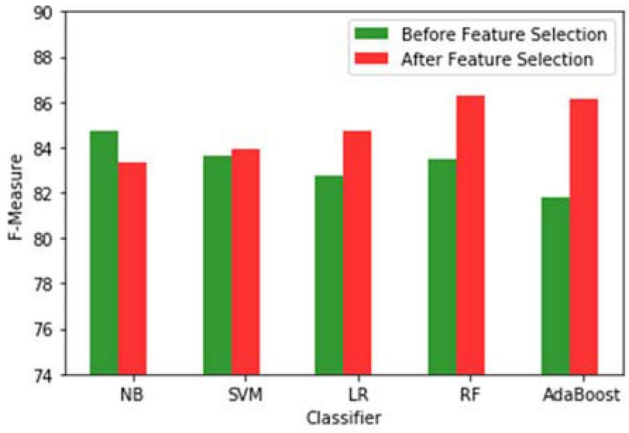

Fig. 17 Increase in F-measure of classifiers using feature selection

A comparison of the proposed system with existing systems is shown in Table 6. The improvement in the accuracy of the proposed system compared to the existing systems is shown in Fig. 18.

5 Conclusion and future scope

The major cause of loss of life in heart disease is a delay in its detection. To minimize this, in this research work, the authors have proposed a hybrid heart disease decision support system. The main contribution of this research is to propose an optimized decision support system for diagnosing heart disease with better accuracy as compared to existing systems. Either it is the stage of missing value, feature selection, or classifier selection; the authors have identified the best algorithms through simulation and used them while proposing the hybrid decision support system. To test and compare the proposed system, the authors have used the cleveland dataset in the simulated environment developed using python. Optimization of the system has been done using RandomizedSearchCV() method. It has shown better performance relative to other hybrid decision support systems found in the literature. Random Forest had given the best accuracy of 86.60%.

The proposed system is not a replacement for a doctor, it can be used inremote and rural areas where heart specialist doctor or other modern medical facilities are not available. Moreover, it can also assist the doctor in taking quick decisions. The proposed system has certain limitations also. It can only diagnose if a person is suffering from heart disease or not. The severity of heart disease can not be diagnosed with this system.

In the future authors have planned to experiment with more methods of feature selection such as Ant Colony optimization, particle swarm optimization to further improve the system performance. Authors have also planned to develop a system to diagnose heart disease using deep learning methods. Furthermore, the authors have also plan to extend the proposed methodology to diagnose other chronic diseases such as chronic kidney disease, diabetes, and cancer. Authors have also planned to deploy the system under the supervision of a doctor to test the performance of the system through real data in real time. In addition, the proposed system can be extended using IoT devices for the collection of clinical parameters in real-time. Clinical data such as ecg, pulse rate, oxygen level, and body temperature can be collected using AD8232, MAX30100, and AMG8833 IR sensors in real-time.

Table 6 Comparison of Proposed hybrid heart disease prediction system with existing systems

Study	Year	Dataset used	Handling of missing values	Feature selection method	Classifiers	Accuracy
Jabbar et al	2016	Cleveland	Rows having missing values were deleted	Chi-square method	Random forest classifier	83.60
Verma and Mathur	2019	Cleveland	Rows having missing values were delete	A hybrid correlation and cuckoo search method	Multilayer perceptron	85.48
Latha and Jeeva	2019	Cleveland	Rows having missing values were deleted	Features were selected by creating different feature subsets randomly	Results of Naive Bayes, Bayes network, multilayer perceptron and random forest classifiers were combined using voting mechanism	85.48
Tama et al	2020	Cleveland	Rows having missing values were deleted	Particle swarm optimization algorithm	Two tier ensemble method using random forest, gradient boosting and extreme gradient boosting classifiers	85.71
Proposed system		Cleveland	Missing values imputed using MICE algorithm	A hybrid method combining GA and RFE algorithms	Random forest classifier	86.60

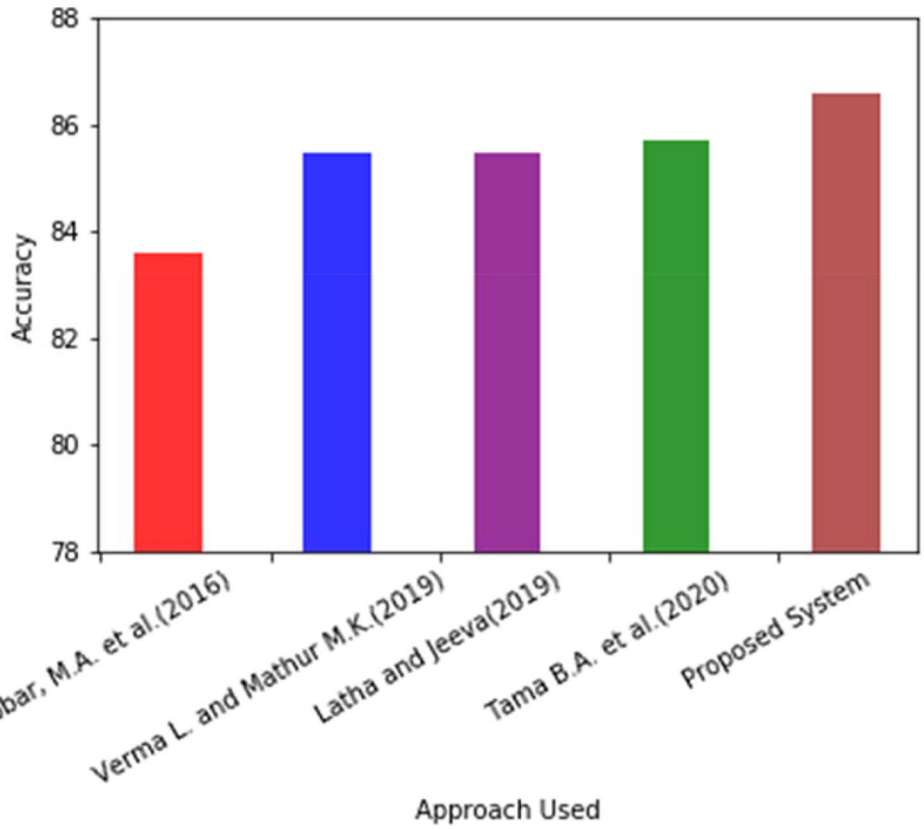

Fig. 18 Improvement in accuracy of the proposed system

Funding Not applicable.

Availability of data and material Not applicable.

Compliance with ethical standards

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Code availability Not applicable.

References

1. Kumar R, Rani P (2020) Comparative analysis of decision support system for heart disease. Adv Math Sci J 9(6):3349–3356. https://doi.org/10.37418/amsj.9.6.15
2. Latha CBC, Jeeva SC (2019) Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. Inf Med Unlocked 16:100203. https://doi.org/10.1016/j.imu.2019.100203
3. Subhadra K, Vikas B (2019) Neural network based intelligent system for predicting heart disease. Int J Innov Technol Explor Eng 8(5):484–487
4. Jain A, Tiwari S, Sapra V (2019) Two-phase heart disease diagnosis system using deep learning. Int J Control Autom 12(5):558–573. http://sersc.org/journals/index.php/IJCA/article/view/2690
5. Coronato A (2018) Engineering high quality medical software: regulations, standards, methodologies and tools for certification. Inst Eng Technol. https://doi.org/10.1049/PBHE012E
6. Coronato A, Cuzzocrea A (2020) An innovative risk assessment methodology for medical information systems. IEEE Trans Knowl Data Eng 13(9):1–14. https://doi.org/10.1109/TKDE.2020.3023553
7. Sartori F, Melen R, Lombardi M, Maggiotto D (2019) Virtual round table knights for the treatment of chronic diseases. J Reliab Intell Environ 5(3):131–143. https://doi.org/10.1007/s40860-019-00089-8
8. Cicotti G (2017) An evidence-based risk-oriented V-model methodology to develop ambient intelligent medical software. J Reliab Intell Environ 3(1):41–53. https://doi.org/10.1007/s40860-017-0039-9
9. Vithanwattana N, Mapp G, George C (2017) Developing a comprehensive information security framework for mHealth: a detailed analysis. J Reliab Intell Environ 3(1):21–39. https://doi.org/10.1007/s40860-017-0038-x
10. Jusob FR, George C, Mapp G (2017) Exploring the need for a suitable privacy framework for mHealth when managing chronic diseases. J Reliab Intell Environ 3(4):243–256. https://doi.org/10.1007/s40860-017-0049-7
11. Ponikowski P, Anker SD, AlHabib KF, Cowie MR, Force TL, Hu S, Jaarsma T, Krum H, Rastogi V, Rohde LE, Samal UC, Shimokawa H, Siswanto BB, Sliwa K, Filippatos G (2014) Heart failure: preventing disease and death worldwide. ESC Heart Fail 1(1):4–25. https://doi.org/10.1002/ehf2.12005

12. Bashir S, Qamar U, Khan FH, Javed MY (2014) MV5: a clinical decision support framework for heart disease prediction using majority vote based classifier ensemble. Arab J Sci Eng 39(11):7771–7783. https://doi.org/10.1007/s13369-014-1315-0
13. Olaniyi EO, Oyedotun OK, Adnan K (2015) Heart diseases diagnosis using neural networks arbitration. Int J Intell Syst Appl 7(12):75–82. https://doi.org/10.5815/ijisa.2015.12.08
14. Verma L, Srivastava S, Negi PC (2016) A hybrid data mining model to predict coronary artery disease cases using non-invasive clinical data. J Med Syst 40(7):1–7. https://doi.org/10.1007/s10916-016-0536-z
15. Verma L, Srivastava S (2016) A data mining model for coronary artery disease detection using noninvasive clinical parameters. Indian J Sci Technol 9(48):1–6. https://doi.org/10.17485/ijst/2016/v9i48/105707
16. Miranda E, Irwansyah E, Amelga AY, Maribondang MM, Salim M (2016) Detection of cardiovascular disease risk's level for adults using naive Bayes classifier. Healthc Inf Res 22(3):196–205. https://doi.org/10.4258/hir.2016.22.3.196
17. Wiharto W, Kusnanto H, Herianto H (2016) Interpretation of clinical data based on C4.5 algorithm for the diagnosis of coronary heart disease. Healthc Inf Res 22(3):186–195. https://doi.org/10.4258/hir.2016.22.3.186
18. Jabbar MA, Deekshatulu BL, Chandra P (2016) Prediction of heart disease using random forest and feature subset selection. In: Innovations in bio-inspired computing and applications. Springer, Cham, pp 187–196. https://doi.org/10.1007/978-3-319-28031-8_16
19. Kim JK, Kang S (2017) Neural network-based coronary heart disease risk prediction using feature correlation analysis. J Healthc Eng 2017:1–13. https://doi.org/10.1155/2017/2780501
20. Arabasadi Z, Alizadehsani R, Roshanzamir M, Moosaei H, Yarifard AA (2017) Computer aided decision making for heart disease detection using hybrid neural network-Genetic algorithm. Comput Methods Programs Biomed 141:19–26. https://doi.org/10.1016/j.cmpb.2017.01.004
21. Liu X, Wang X, Su Q, Zhang M, Zhu Y, Wang Q, Wang Q (2017) A hybrid classification system for heart disease diagnosis based on the RFRS method. Comput Math Methods Med 2017:1–11. https://doi.org/10.1155/2017/8272091
22. Verma L, Srivastava S, Negi PC (2018) An intelligent noninvasive model for coronary artery disease detection. Complex Intell Syst 4(1):11–18. https://doi.org/10.1007/s40747-017-0048-6
23. David H, Belcy SA (2018) Heart disease prediction using data mining techniques. ICTACT J Soft Comput 9(1):1824–1830. https://doi.org/10.1109/I2C2.2017.8321771
24. Haq AU, Li JP, Memon MH, Nazir S, Sun R (2018) A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms. Mobile Inf Syst 2018:1–21. https://doi.org/10.1155/2018/3860146
25. Malav A, Kadam K (2018) A hybrid approach for heart disease prediction using artificial neural network and K-means. Int J Pure Appl Math 118(8):103–110
26. Poornima V, Gladis D (2018) A novel approach for diagnosing heart disease with hybrid classifier. Biomed Res 29:2274–2280. https://doi.org/10.4066/biomedicalresearch.38-18-434
27. Khourdifi Y, Bahaj M (2019) Heart disease prediction and classification using machine learning algorithms optimized by Particle Swarm Optimization and Ant Colony Optimization. Int J Intell Eng Syst 12(1):242–252. https://doi.org/10.22266/ijies2019.0228.24
28. Ali L, Rahman A, Khan A, Zhou M, Javeed A, Khan JA (2019) An automated diagnostic system for heart disease prediction based on Chi square statistical model and optimally configured deep neural network. IEEE Access 7:34938–34945. https://doi.org/10.1109/ACCESS.2019.2904800
29. Mohan S, Thirumalai C, Srivastava G (2019) Effective heart disease prediction using hybrid machine learning techniques. IEEE Access 7:81542–81554. https://doi.org/10.1109/ACCESS.2019.2923707
30. Ali L, Niamat A, Khan JA, Golilarz NA, Xingzhong X, Noor A, Nour R, Bukhari SAC (2019) An optimized stacked support vector machines based expert system for the effective prediction of heart failure. IEEE Access 7:54007–54014. https://doi.org/10.1109/ACCESS.2019.2909969
31. Verma L, Mathur MK (2020) Deep learning based model for decision support with case based reasoning. Int J Innov Technol Explor Eng 8(6C):149–153
32. Mienye ID, Sun Y, Wang Z (2020) Improved sparse autoencoder based artificial neural network approach for prediction of heart disease. Inf Med Unlocked 18:1–5. https://doi.org/10.1016/j.imu.2020.100307
33. Terrada O, Hamida S, Cherradi B, Raihani A, Bouattane O (2020) Supervised machine learning based medical diagnosis support system for prediction of patients with heart disease. Adv Sci Technol Eng Syst J 5(5):269–277. https://doi.org/10.25046/aj050533
34. Paragliola G, Coronato A (2020) An hybrid ECG-based deep network for the early identification of high-risk to major cardiovascular events for hypertension patients. J Biomed Inform. https://doi.org/10.1016/j.jbi.2020.103648
35. Tama BA, Im S, Lee S (2020) Improving an intelligent detection system for coronary heart disease using a two-tier classifier ensemble. Biomed Res Int 2020:1–10. https://doi.org/10.1155/2020/9816142
36. Saez JA, Krawczyk B, Woźniak M (2016) Analyzing the oversampling of different classes and types of examples in multi-class imbalanced datasets. Pattern Recogn 57:164–178. https://doi.org/10.1016/j.patcog.2016.03.012
37. Detrano R (1989) Long Beach and Cleveland Clinic Foundation. VA Medical Center. http://archive.ics.uci.edu/ml/datasets/Heart+Disease. Accessed 10 Sept 2020
38. Azur MJ, Stuart EA, Frangakis C, Leaf PJ (2011) Multiple imputation by chained equations: what is it and how does it work. Int J Methods Psychiatr Res 20(1):40–49
39. Priscila SS, Hemalatha M (2017) Improving the performance of entropy ensembles of neural networks (EENNS) on classification of heart disease prediction. Int J Pure Appl Math 117(7):371–386
40. Dulhare UN (2018) Prediction system for heart disease using Naive Bayes and particle swarm optimization. Biomed Res 29(12):2646–2649. https://doi.org/10.4066/biomedicalresearch.29-18-620
41. Rani P, Kumar R, Jain A, Lamba R (2020) Taxonomy of machine learning algorithms and its applications. J Comput Theror Nanosci 17(6):2509–2514. https://doi.org/10.1166/jctn.2020.8922