



Office of the Controller General of Patents, Designs & Trade Marks  
Department of Industrial Policy & Promotion,  
Ministry of Commerce & Industry,  
Government of India

(<http://ipindia.nic.in/index.htm>)



(<http://ipindia.nic.in/index.htm>)

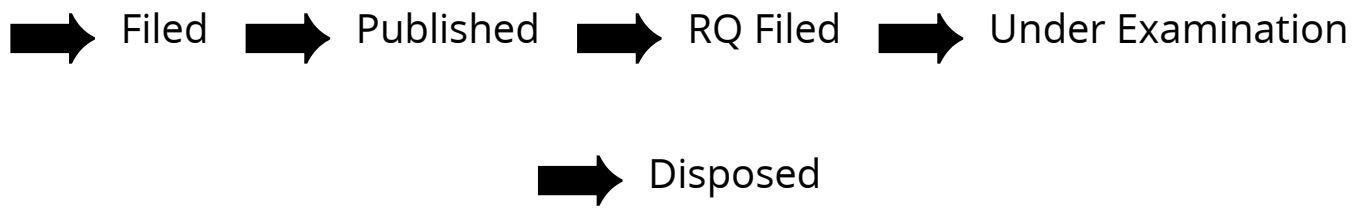
#### Application Details

APPLICATION NUMBER	202011025642
APPLICATION TYPE	ORDINARY APPLICATION
DATE OF FILING	18/06/2020
APPLICANT NAME	1 . Pooja Rani 2 . Dr. Rajneesh Kumar 3 . Dr. Anurag Jain 4 . Tarun Gulati 5 . Rohit Lamba 6 . Dr.Renu Sharma
TITLE OF INVENTION	A HYBRID DECISION SUPPORT SYSTEM FOR HEART DISEASE PREDICTION
FIELD OF INVENTION	COMPUTER SCIENCE
E-MAIL (As Per Record)	ashish.iprindia@hotmail.com
ADDITIONAL-EMAIL (As Per Record)	ipnation@out.com
E-MAIL (UPDATED Online)	
PRIORITY DATE	
REQUEST FOR EXAMINATION DATE	--
PUBLICATION DATE (U/S 11A)	07/08/2020

#### Application Status

APPLICATION STATUS	<b>Awaiting Request for Examination</b>
--------------------	---

[View Documents](#)



In case of any discrepancy in status, kindly contact [ipo-helpdesk@nic.in](mailto:ipo-helpdesk@nic.in)

FORM 2  
THE PATENTS ACT 1970  
(39 of 1970)  
&  
THE PATENT RULES, 2003  
COMPLETE SPECIFICATION  
(See section 10 and rule 13)

**1. TITLE OF THE INVENTION: -**

**A HYBRID DECISION SUPPORT SYSTEM FOR HEART DISEASE PREDICTION**

**2. Applicant(s):-**

Name	Nationality	Address
Pooja Rani	INDIA	Maharishi Markandeshwar Institute of Computer Technology & Business Management, Maharishi Markandeshwar (Deemed to be University), Mullana, Ambala, Haryana, 133207, India
Dr. Rajneesh Kumar	INDIA	Department of Computer Science and Engineering, MMEC, Maharishi Markandeshwar (Deemed to be University), Mullana, Ambala, Haryana, 133207, India
Dr. Anurag Jain	INDIA	Department of Virtualization, School of Computer Science, University of Petroleum and Energy Studies, Dehradun, Uttarakhand, 248007, India
Tarun Gulati	INDIA	Department of ECE, MMEC, Maharishi Markandeshwar (Deemed to be University), Mullana, Ambala, Haryana, India
Rohit Lamba	INDIA	Department of ECE, MMEC, Maharishi Markandeshwar (Deemed to be University), Mullana, Ambala, Haryana, 133207, India
Dr. Renu Sharma	INDIA	Department of Physics, MMEC, Maharishi Markandeshwar (Deemed to be University), Mullana, Ambala, Haryana, 133207, India

**3. PREAMBLE OF THE DESCRIPTION**

The following specification particularly describes the invention and the manner in which it is to be performed

## **FIELD OF THE INVENTION**

This invention relates to a hybrid decision support system which is used in remote and isolated areas where heart specialist and modern diagnosis tools are not available. The present invention can play a vital role in early stage detection of heart disease. It can decrease the mortality rate. Further, It can assist doctor in taking decisions through clinical data collected from non invasive tests.

## **BACKGROUND OF THE INVENTION**

The heart is the most important part of the human body 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. In heart disease, the heart is not able to supply enough oxygen-rich blood to the organs of the body, which causes heart attack. According to data from the world health organization, around 17 million people die every year worldwide due to heart disease (Latha & Jeeva, 2019). There are various types of heart diseases such as Congenital Heart Disease, Coronary Artery Disease, Arrhythmia etc. Patient suffering from heart disease has various symptoms such as chest pain, dizzy sensations, and deep sweating (Subhadra & Vikas, 2019). Invasive methods of diagnosing the disease are expensive and painful. Therefore, there is a need of technique that can diagnose heart disease in a less expensive and less painful manner (Jain et al., 2019).

Different researchers had proposed different decision support systems to predict heart disease. Bashir et al. (Bashir et al., 2014) 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. Olaniyi and Oyedotun (Olaniyi & Oyedotun, 2015) developed a heart disease diagnosis system with a multilayer perceptron neural network and support vector machine algorithms. Verma et al. (Verma et al., 2016) 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 combined with Particle Swarm Optimization. Verma and Srivastava (Verma & Srivastava, 2016) proposed a system for diagnosing coronary artery disease using a neural

network model. Miranda et al. (Miranda et al. , 2016) developed a decision support system to detect the risk of cardiovascular disease using the naive bayes Classifier. Wiharto et al. (Wiharto et al. , 2016) performed the classification of heart disease using the C4.5 algorithm on the data of UCI (University of California, Irvine) repository. The problem of imbalanced data was handled by SMOTE (Synthetic Minority Oversampling Technique). Dimensionality reduction was done by selecting the relevant features using Information Gain (IG) method. Jabbar et al. (Jabbar et al., 2016) in their proposed system used the random forest to perform heart disease prediction. Chi-square method was used for feature selection to select relevant attributes. Kim and Kang (Kim and Kang, 2017) 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. Arabasadi et al. (Arabasadi et al., 2017) developed a system for predicting heart disease using neural network optimized by genetic algorithm. Liu et al. (Liu et al., 2017) 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 (Verma et al., 2018). David and Belcy (David & Belcy, 2018) used random forest, decision tree and naive bayes classifiers to predict heart disease. Haq et al. (Haq et al., 2018) proposed heart disease prediction system using six machine learning algorithms logistic regression, support vector machine, naive bayes, artificial neural network, decision tree, and K-nearest neighbor. Malav and Kadam (Malav & Kadam, 2018) 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 (Poornima & Gladis, 2018) 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 by using OLPP (Orthogonal Local Preserving Projection). The classification was performed using a neural network. Khourdifi and Bahaj (Khourdifi & Bahaj, 2019) 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. Ali et al. (Ali et al., 2019) 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. (Mohan et al., 2019)

developed a heart disease prediction system using random forest and linear method. Ali et al. (Ali et al., 2019) developed a system for 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.

The major motivation behind proposing this decision support system is to develop a system, which can be used by any educated person in the absence of a doctor to diagnose heart disease at an early stage. In developing countries, many people lost their lives due to the non-detection of heart disease timely. They usually approach to a heart specialist at a very late stage. This kind of system may assist the doctor while taking the decision.

## **SUMMARY OF THE INVENTION**

The aim of this research was to propose a decision-making system for predicting heart disease that can assist doctors in early stage prediction of heart disease.

Non detection of heart disease at early stage can become the cause of death. In developing countries, where heart specialist doctors are not available in remote, semi-urban and rural areas, there is need of decision support system which can help people in absence of doctor to diagnose heart disease at early stage.

Inventors have used Multivariate Imputation by Chained Equations algorithm to handle the missing value, and hybridized Genetic and Recursive Feature Elimination algorithm for selection of suitable features from dataset. Further for pre-processing of data, SMOTE(Synthetic Minority Oversampling Technique) and standard scalar methods are used. In the last step of development of system, inventors have used naive bayes, support vector machine, random forest ,logistic regression and adaboost classifiers. It was tested on Cleveland heart disease dataset available in UCI (University of California, Irvine) machine learning repository. The present invention has given the highest accuracy of 86.6% with random forest . Accuracy given by proposed system is superior to existing systems in the literature. Proposed decision support system can be used for early detection of heart disease and can reduce mortality rate.

## **BRIEF DESCRIPTION OF THE DRAWINGS**

Figure 1: Proposed hybrid heart disease prediction system

Figure 2: Multiple Imputation Chained Equations algorithm

Figure 3: Genetic algorithm

Figure 4: Recursive Feature Elimination algorithm

Figure 5: Hyperplane separating two classes in SVM

Figure 6: Random Forest algorithm

Figure 7: Adaboost algorithm

Figure 8: Increase in accuracy of classifiers using class balancing

Figure 9: Increase in sensitivity of classifiers using class balancing

Figure 10: Increase in specificity of classifiers using class balancing

Figure 11: Increase in precision of classifiers using class balancing

Figure 12: Increase in F-Measure of classifiers using class balancing

Figure 13: Increase in accuracy of classifiers using feature selection

Figure 14: Increase in sensitivity of classifiers using feature selection

Figure 15: Increase in specificity of classifiers using feature selection

Figure 16: Increase in precision of classifiers using feature selection

Figure 17: Increase in F-Measure of classifiers using feature selection

Figure 18: Improvement in accuracy of proposed system

## **DETAILED DESCRIPTION OF THE INVENTION**

### **Methods**

#### **Methodology**

In this research, the inventors 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 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 by 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. It creates synthetic samples of minor class resulting in an equal number of samples of both classes (Saez et al., 2016). 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. Proposed hybrid system was tested in the simulation environment developed using Python. Methodology of proposed hybrid system for heart disease prediction is shown in Figure 1.

### 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 8 are categorical features and 6 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 4 types of chest pain: typical angina, atypical angina, non-angina pain and asymptomatic. 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 120mg/dl then 1 is stored in this feature, otherwise 0 is stored. RECR has electrocardiographic results and maximum heart rate of patient is stored in MHR. EIGA has a value of 1 if 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 0 for the absence of heart disease and 1 to 4 for different levels of the disease. Levels 1 to 4 are merged to indicate the presence of disease. Dataset has 6 instances having missing values. There are 4 missing values in NMVCF feature and 2 missing values in THALM feature (Detrano, 1989).

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

### **Multivariate Imputation by Chained Equations algorithm (MICE)**

Cleveland dataset has 6 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 (Azur et al., 2011). Steps of this algorithm are shown in Figure 2.

### **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 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 in the population. This diversity is introduced by using

mutation operator which makes random changes in the individuals (Arabasadi et al., 2017). Steps of genetic algorithm are shown in Figure 3.

Recursive Feature Elimination (RFE) algorithm recursively removes irrelevant features. A classifier is trained on the training data and the 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 (Priscila & Hemalatha, 2017). Working of RFE algorithm is shown in Figure 4.

Figure 4: Recursive Feature Elimination algorithm

### **Classification algorithms**

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

#### **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 (Rani et al., 2020). It calculates the posterior probability of the class using equation 1:

$$P(c|I) = P(I|c)P(c) \div P(I) \quad (1)$$

Posterior probability denotes the probability of occurring class  $c$  with input  $I$ . In equation 1,  $P(c)$  is prior class probability and  $P(I)$  is prior feature probability.  $P(I|c)$  is likelihood which indicates the probability of occurring feature  $I$  with class  $c$ . This algorithm can also be used for multi classification problems (Dulhare, 2018).

#### **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 hyperpalne 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 (Rani et al., in press). Hyperplane can be created as given in equation 2:

$$H_0: w^T x + b = 0 \quad (2)$$

Two more hyperplanes are created in parallel to constructed hyperplane as given in equations 3 and 4.

$$H_1: w^T x + b = -1 \quad (3)$$

$$H_2: w^T x + b = 1 \quad (4)$$

Hyperplane should satisfy the constraints given by equation 5 and 7 for each input vector  $I_j$

$$wI_j + b \geq +1 \text{ for } I_j \text{ having class 1} \quad (5)$$

and

$$wI_j + b \geq -1 \text{ for } I_j \text{ having class 0} \quad (6)$$

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

Figure 5: Hyperplane separating two classes in SVM

### **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 by equation 7 calculates the probability by which input  $X$  should be classified as class 1:

$$P(X) = \frac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)} \quad (7)$$

Here  $\beta_0$  is bias and  $\beta_1$  is the weight which is multiplied by input  $X$  (Rani et al., in press).

### **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 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 (Jabbar et al., 2016). The working of random forest is shown in Figure 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, inventors had done hyperparameter tuning of Random Forest using RandomizedSearchCV() method.

Figure 6: Random Forest algorithm

### **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. It 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 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 final prediction. Weak classifier to which boosting is to be applied can be passed as a parameter to Adaboost algorithm. Default classifier used for boosting in Adaboost is decision tree (Latha & Jeeva, 2019). Working of AdaBoost algorithm is shown in Figure 7.

Figure 7: Adaboost algorithm

### **Evaluation Parameters**

The performance of classifiers was evaluated on the scale of accuracy, sensitivity, specificity, precision, and recall (Mienye et al., 2020). If a person suffering from the disease is predicted to be a heart disease patient by the system then it is true positive, otherwise, it is false negative. Similarly, if a healthy person is predicted to be disease free then it is true negative, otherwise it is false positive.

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

$$\text{Accuracy} = \left( \frac{\text{Correct Predictions}}{\text{Total Predictions}} \right) * 100$$

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

$$\text{Sensitivity} = \left( \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \right) * 100$$

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

$$\text{Specificity} = \left( \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}} \right) * 100$$

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

$$\text{Precision} = \left( \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \right) * 100$$

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

$$\text{F - Measure} = 2 * \frac{\text{Sensitivity} * \text{Precision}}{\text{Sensitivity} + \text{Precision}}$$

## Results

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 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.

### 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. Performance of classifiers on full feature set is shown in Table 2. NB classifier provided the highest performance on full feature set, whereas SVM provided the lowest performance.

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

### 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 performance of classifiers. Performance of some classifiers increased, whereas performance of some classifiers decreased. NB, RF, and Adaboost resulted in no change in performance.

Accuracy of SVM increased by 6.66%, whereas 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 doesn't have a positive impact on the performance of other classifiers. Impact of scaling on the performance of classifiers is shown in Table 3.

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

### 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. 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 Figures 8, 9, 10, 11 and 12 respectively.

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

---

### Performance improvement using feature selection

Accuracy of classifiers was further improved using feature selection. Hybrid GA and RFE 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 Figures 13,14,15,16 and 17 respectively.

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

---

### Discussions

Proposed system is a hybrid system that can be used to predict heart disease using clinical parameters. Beauty of this system is that it will depend entirely on clinical data that does not require a heart specialist doctor. However, various heart disease decision support systems have proposed with varying degrees of accuracy, most of the researchers have not considered the issue of missing values and feature selection approach collectively. In the proposed system, inventors have not only handled the issue of missing value but also handled the feature selection issue efficiently.

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. Preprocessing techniques of scaling and class balancing also contributed to performance enhancement. This hybrid approach resulted in system with increased accuracy than some existing systems in Literature.

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 Figure 18.

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

	Random Forest and Chi-square (Jabbar, M.A. et al. 2016)	Majority Vote with NB, BN, RF, and MP (Latha & Jeeva, 2019)	Proposed System (RF with MICE, Standard Scalar, SMOTE, and GARFE )
Handling of Missing values	Missing value imputation mechanism was not used in the system.	Missing value imputation mechanism was not used in the system.	Missing values are imputed using MICE algorithm.
Class balancing	Class balancing was not performed in the system.	Class balancing was not performed in the system.	Class balancing is done using SMOTE algorithm.
Scaling	Scaling of values was not done.	Scaling of values was not done.	Values of features are scaled using standard scalar algorithm
Feature Selection	Feature Selection was done using Chi-Square algorithm.	Features were selected by creating different feature subsets randomly.	Features are selected by using a hybrid mechanism combining GA and RFE algorithms.
Classifiers	Random Forest Classifier.	Results of Naive Bayes, Bayes Network, Multilayer Perceptron and	Random Forest Classifier.



		Random Forest classifiers were combined using voting mechanism.	
Accuracy	83.60	85.48	86.60

As the system improved accuracy as compared to existing systems, it can be very helpful for doctors.

### Conclusion & 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 inventors have proposed a hybrid heart disease decision support system. Either it is the stage of missing value, feature selection or classifier selection; inventors have identified the best algorithms through simulation and used them while proposing the hybrid decision support system. To test and compare the proposed system, inventors have used the Cleveland dataset in the simulated environment developed using python. It has shown better performance relative to other hybrid decision support system found in the literature. Proposed system is not a replacement for a doctor, it can be used in remote and rural areas where heart specialist doctor or other modern medical facilities are not available. Moreover, it can also assist doctor in taking quick decision.

In future, inventors have 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, inventors have planned to use IOT devices to club with system so that real time clinical data can be collected easily. Further, by deploying the system on cloud, it can be accessed remotely also.

### References

- [1] Ali, L., Niamat, A., Khan, J.A., Golilarz, N.A., Xingzhong, X., Noor, A., Nour, R., Bukhari, S.A.C.(2019). An optimized stacked support vector machines based expert system for the effective prediction of heart failure. *IEEE Access*, 7, 54007-54014..
- [2] Ali, L., Rahman, A., Khan, A., Zhou, M., Javeed, A., & Khan, J. A. (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.
- [3] Arabasadi, Z., Alizadehsani, R., Roshanzamir, M., Moosaei, H., & Yarifard, A. A. (2017). Computer aided decision making for heart disease detection using hybrid neural network-Genetic algorithm. *Computer methods and programs in biomedicine*, 141, 19-26.


- [4] Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: what is it and how does it work?. *International journal of methods in psychiatric research*, 20(1), 40-49.
- [5] Bashir, S., Qamar, U., Khan, F. H., & Javed, M. Y. (2014). MV5: a clinical decision support framework for heart disease prediction using majority vote based classifier ensemble. *Arabian Journal for Science and Engineering*, 39(11), 7771-7783..
- [6] David, H., & Belcy, S. A. (2018). Heart disease prediction using data mining techniques. *ICTACT Journal on Soft Computing*, 9(1), 1824-1830.
- [7] Detrano, R. Long Beach and Cleveland Clinic Foundation. VA Medical Center: <http://archive.ics.uci.edu/ml/datasets/Heart+Disease>, 1989.
- [8] Dulhare, U. N. (2018). Prediction system for heart disease using Naive Bayes and particle swarm optimization, *Biomedical Research*, 29(12), 2646-2649.
- [9] Haq, A. U., Li, J. P., Memon, M. H., Nazir, S., & Sun, R. (2018). A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms. *Mobile Information Systems*, 2018, 1-21.
- [10] Jabbar, M. A., Deekshatulu, B. L., & Chandra, P. (2016). Prediction of heart disease using random forest and feature subset selection. In *Innovations in bio-inspired computing and applications* (pp. 187-196). Springer, Cham.
- [11] Jain, A., Tiwari, S., & Sapra, V. (2019). Two-Phase Heart Disease Diagnosis System using Deep Learning. *International Journal of Control and Automation*, 12 (5), 558-573.
- [12] Khouridifi, Y., & Bahaj, M. (2019). Heart Disease Prediction and Classification Using Machine Learning Algorithms Optimized by Particle Swarm Optimization and Ant Colony Optimization. *International Journal of Intelligent Engineering & Systems*, 12(1), 242-252.
- [13] Kim, J. K., & Kang, S. (2017). Neural network-based coronary heart disease risk prediction using feature correlation analysis. *Journal of healthcare engineering*, 2017, 1-13.
- [14] Latha, C. B. C., & Jeeva, S. C. (2019). Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. *Informatics in Medicine Unlocked*, 16, 100203.
- [15] 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. *Computational and mathematical methods in medicine*, 2017, 1-11.
- [16] Malav, A., & Kadam, K. (2018). A hybrid approach for heart disease prediction using artificial neural network and K-means. *International Journal of Pure and Applied Mathematics*, 118(8), 103-110.
- [17] Mienye, I. D., Sun, Y., & Wang, Z. (2020). Improved sparse autoencoder based artificial neural network approach for prediction of heart disease. *Informatics in Medicine Unlocked*, 18, 1-5.
- [18] Miranda, E., Irwansyah, E., Amelga, A. Y., Maribondang, M. M., & Salim, M. (2016). Detection of cardiovascular disease risk's level for adults using naive Bayes classifier. *Healthcare informatics research*, 22(3), 196-205.
- [19] Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. *IEEE Access*, 7, 81542-81554.
- [20] Olaniyi, E. O., Oyedotun, O. K., & Adnan, K. (2015). Heart diseases diagnosis using neural networks arbitration. *International Journal of Intelligent Systems and*

- Applications, 7(12), 75-82.
- [21] Poornima, V., & Gladis, D. (2018). A novel approach for diagnosing heart disease with hybrid classifier. *Biomedical Research*, 29, 2274-2280.
- [22] Priscila, S.S, & Hemalatha, M. (2017). Improving the Performance of Entropy Ensembles of Neural Networks (EENNS) on Classification of Heart Disease Prediction. *International Journal of Pure and Applied Mathematics*, 117(7), 371-386.
- [23] Rani, P., Kumar, R., & Jain, A. (2020). Multistage Model for Accurate Prediction of Missing Values in Heart Disease Dataset. *Proceedings of International Conference on Sentimental Analysis and Deep Learning*, 147-158.
- [24] Rani, P., Kumar, R., Jain, A., & Lamba, R. (in press). Taxonomy of Machine Learning Algorithms and its Applications. *Journal of Computational and Theroretical Nanoscience*.
- [25] Saez, J. A., Krawczyk, B., & Woźniak, M. (2016). Analyzing the oversampling of different classes and types of examples in multi-class imbalanced datasets. *Pattern Recognition*, 57, 164-178.
- [26] Subhadra, K., & Vikas, B. (2019). Neural Network Based Intelligent System for Predicting Heart Disease. *International Journal of Innovative Technology and Exploring Engineering* , 8(5), 484-487.
- [27] Verma, L., & Srivastava, S. (2016). A data mining model for coronary artery disease detection using noninvasive clinical parameters. *Indian Journal of Science and Technology*, 9(48), 1-6.
- [28] Verma, L., Srivastava, S., & Negi, P. C. (2016). A hybrid data mining model to predict coronary artery disease cases using non-invasive clinical data. *Journal of medical systems*, 40(7), 1-7.
- [29] Verma, L., Srivastava, S., & Negi, P. C. (2018). An intelligent noninvasive model for coronary artery disease detection. *Complex & Intelligent Systems*, 4(1), 11-18.
- [30] Wiharto, W., Kusnanto, H., & Herianto, H. (2016). Interpretation of clinical data based on C4. 5 algorithm for the diagnosis of coronary heart disease. *Healthcare informatics research*, 22(3), 186-195.

**We Claim:**

1. A Hybrid Decision Support System for Heart Disease Prediction comprises three stages: data collection, data pre-processing, and model construction which is used to diagnose heart disease in patients.
2. The system as claimed in claim 1, used Naive Bayes, Support Vector Machine, Logistic Regression, Random Forest and Adaboost classifiers; wherein Random Forest provided highest performance.
3. The system as claimed in claim 1, used Cleveland dataset from UCI to perform experiment; wherein heart disease is diagnosed using several medical parameters available in dataset;  
wherein said parameters are used to perform classification with class 1 indicating that the person has a disease and class 0 indicating that person is disease-free.
4. The system as claimed in claim 1, wherein said Dataset is having missing values in 6 instances; and said values are imputed using MICE algorithm.
5. The system as claimed in claim 1, wherein said application of this algorithm resulted in a complete dataset with no instance having missing value; and system performance is measured on the scale of accuracy, sensitivity, specificity, precision, and F-measure.

Dated this June 18, 2020

  
(Ashish Sharma)  
Authorized Agent for the Applicant  
Indian Patent Agent Regn No. IN/PA-3021

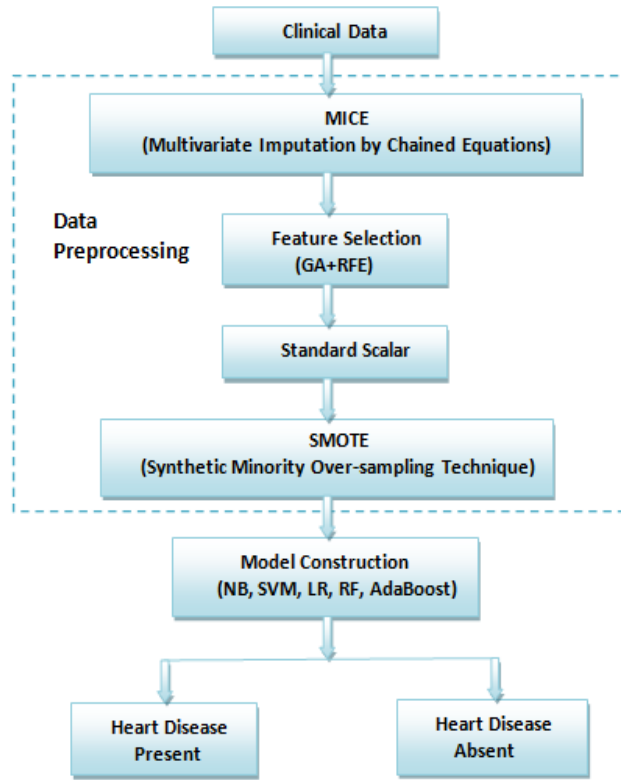


Figure 1: Proposed hybrid heart disease prediction system

*Ashish*

(Ashish Sharma)  
Authorized Agent for the Applicant  
Indian Patent Agent Regn No. IN/PA-3021

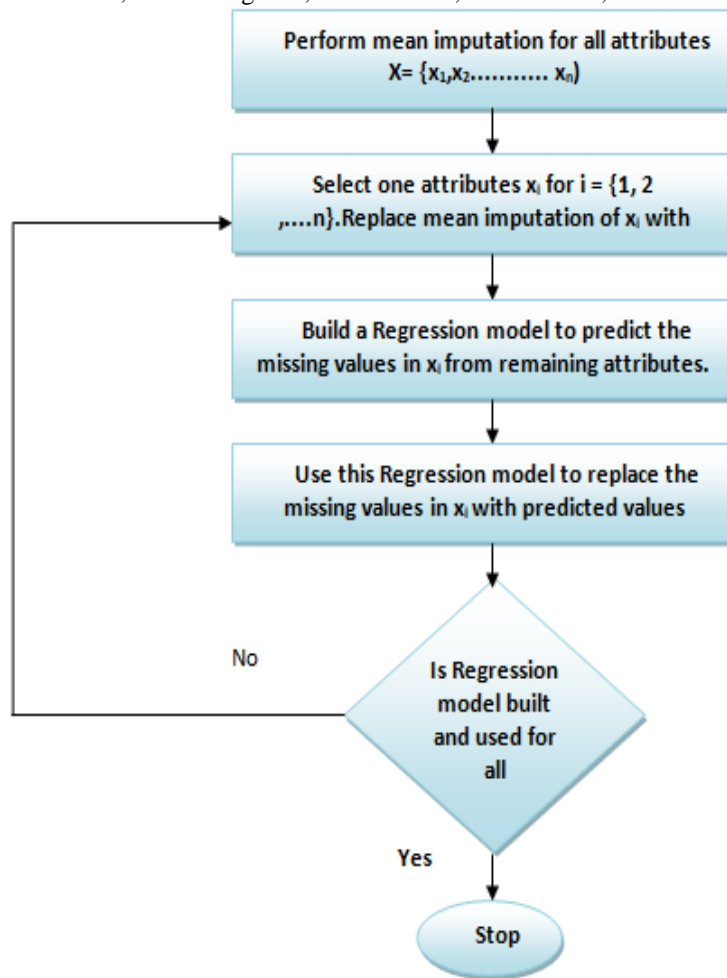


Figure 2: Multiple Imputation Chained Equations algorithm

*Ashish*

(Ashish Sharma)  
Authorized Agent for the Applicant  
Indian Patent Agent Regn No. IN/PA-3021

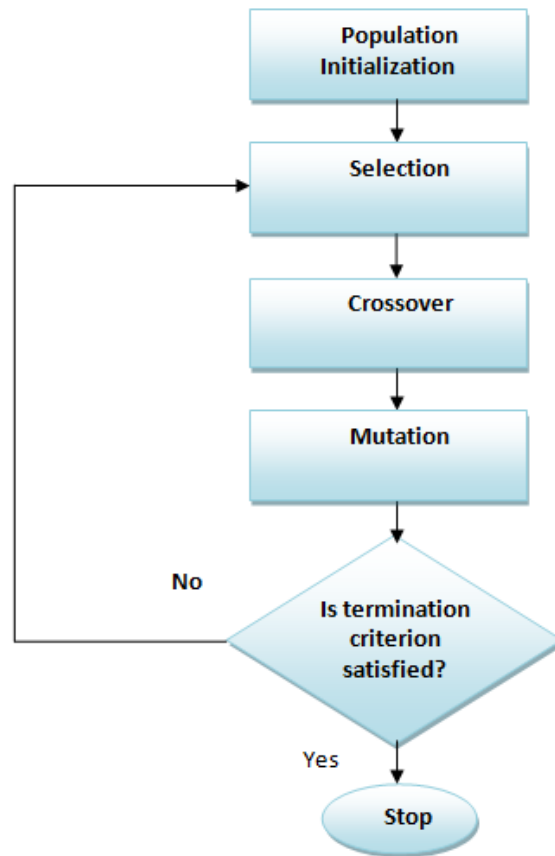


Figure 3: Genetic algorithm

*Ashish*

(Ashish Sharma)  
Authorized Agent for the Applicant  
Indian Patent Agent Regn No. IN/PA-3021

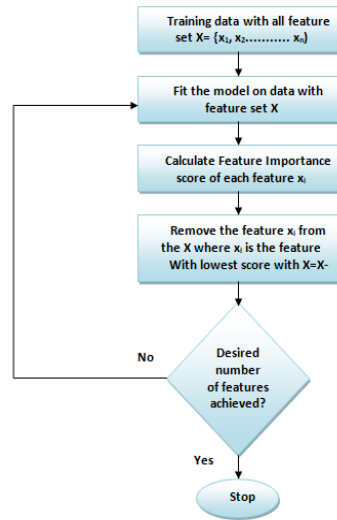


Figure 4: Recursive Feature Elimination algorithm

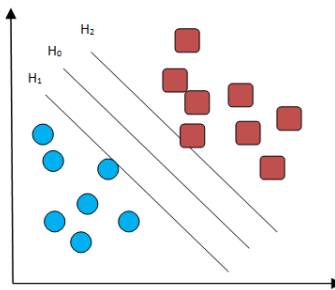


Figure 5: Hyperplane separating two classes in SVM

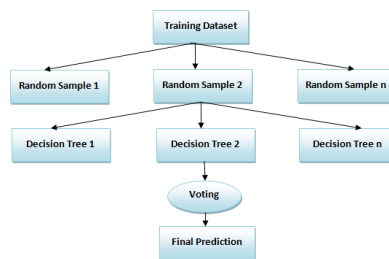


Figure 6: Random Forest algorithm

*Ashish*

(Ashish Sharma)  
 Authorized Agent for the Applicant  
 Indian Patent Agent Regn No. IN/PA-3021



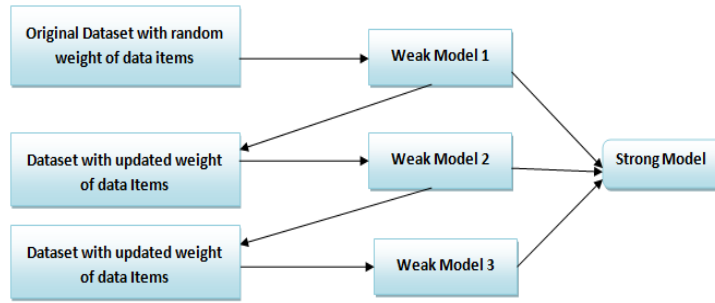


Figure 7: Adaboost algorithm

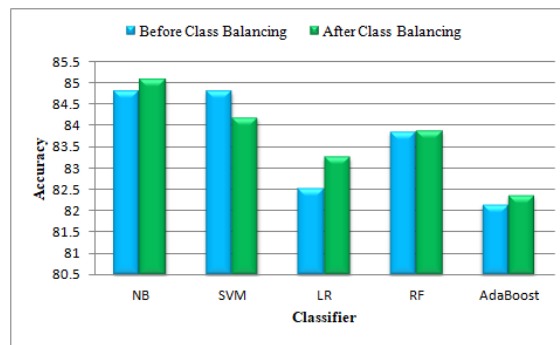


Figure 8: Increase in accuracy of classifiers using class balancing

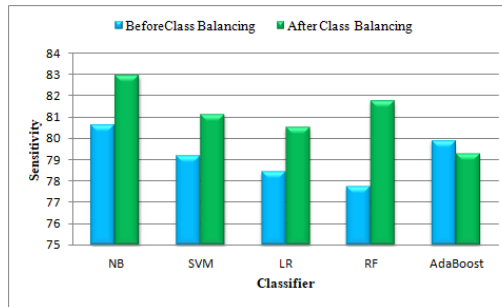


Figure 9: Increase in sensitivity of classifiers using class balancing

*Ashish*

(Ashish Sharma)  
 Authorized Agent for the Applicant  
 Indian Patent Agent Regn No. IN/PA-3021

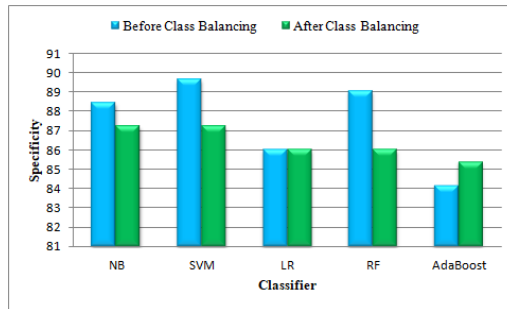


Figure 10: Increase in specificity of classifiers using class balancing

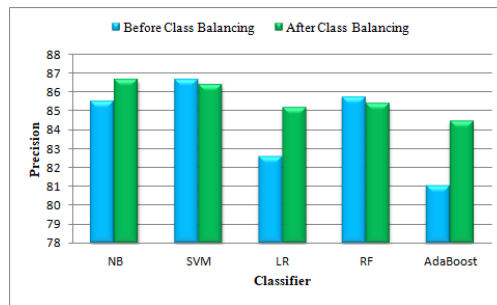


Figure 11: Increase in precision of classifiers using class balancing

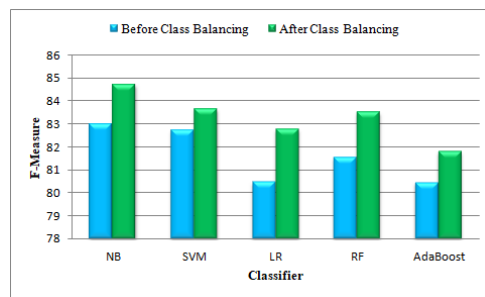


Figure 12: Increase in F-Measure of classifiers using class balancing

*Ashish*

(Ashish Sharma)  
Authorized Agent for the Applicant  
Indian Patent Agent Regn No. IN/PA-3021

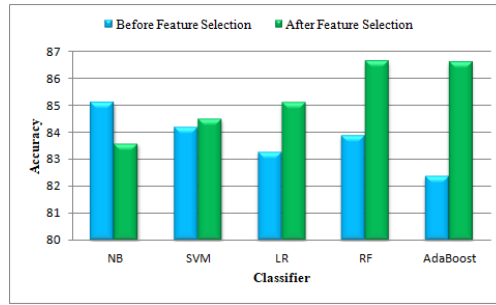


Figure 13: Increase in accuracy of classifiers using feature selection

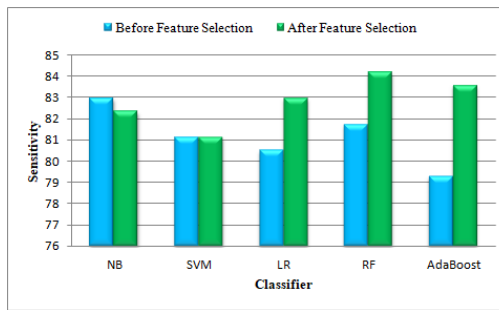


Figure 14: Increase in sensitivity of classifiers using feature selection

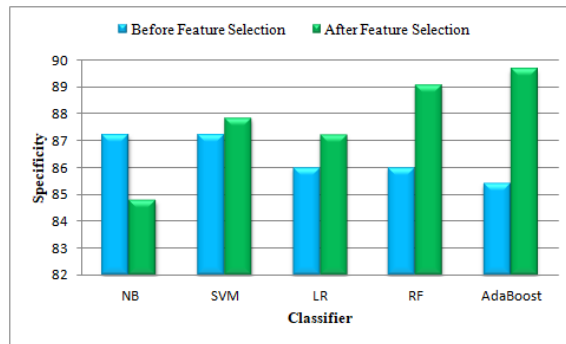


Figure 15: Increase in specificity of classifiers using feature selection

(Ashish Sharma)  
Authorized Agent for the Applicant  
Indian Patent Agent Regn No. IN/PA-3021

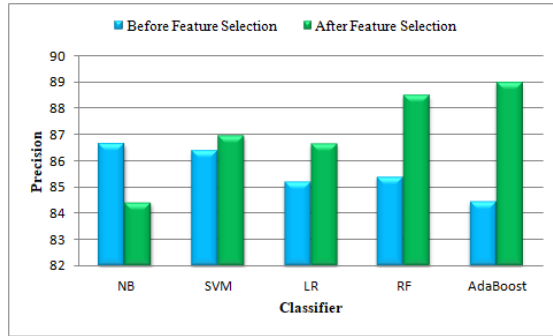


Figure 16: Increase in precision of classifiers using feature selection

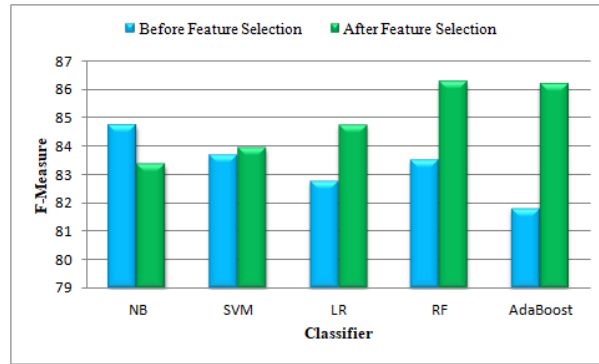


Figure 17: Increase in F-Measure of classifiers using feature selection

*Ashish*

(Ashish Sharma)  
Authorized Agent for the Applicant  
Indian Patent Agent Regn No. IN/PA-3021

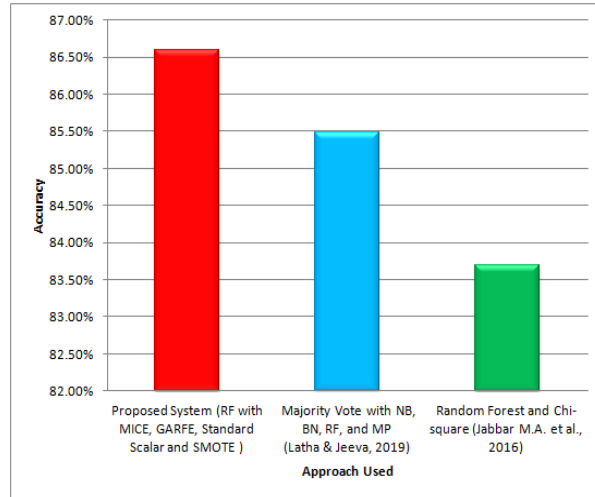


Figure 18: Improvement in accuracy of proposed system

*Ashish*

(Ashish Sharma)  
Authorized Agent for the Applicant  
Indian Patent Agent Regn No. IN/PA-3021

## **ABSTRACT**

### **Hybrid Decision Support System for Heart Disease Prediction**

Discloses herein a Hybrid Decision Support System for Heart Disease Prediction comprises Naive Bayes, Support Vector Machine, Logistic Regression, Random Forest and Adaboost which are used to diagnose heart disease in patients. Non detection of heart disease at early stage can become the cause of death. In developing countries, where heart specialist doctors are not available in remote, semi-urban and rural areas, there is need of decision support system which can help people in absence of doctor to diagnose heart disease at early stage.

Inventors have used Multivariate Imputation by Chained Equations algorithm to handle the missing value, and hybridized Genetic and Recursive Feature Elimination algorithm for selection of suitable features from dataset. Further for pre-processing of data, SMOTE(Synthetic Minority Oversampling Technique) and standard scalar methods are used. In the last step of development of system, inventors have used naive bayes, support vector machine, random forest ,logistic regression and adaboost classifiers. It was tested on Cleveland heart disease dataset available in UCI (University of California, Irvine) machine learning repository. System has given the highest accuracy of 86.6% with random forest . Accuracy given by proposed system is superior to existing systems in the literature. Present invention can be used for early detection of heart disease and can reduce mortality rate.