Hybrid Ensemble and Features Selection for Heart Disease Prediction Using Machine Learning

Sharmila Zope
Department of Computer Engineering
School of computer Science KBCs North Maharashtra University.
Jalgaon, India.
sharmilazope@gmail.com
Nandini Chaudhari
Department of Computer Engineering
DRS Kiran and Pallavi Patel Global University.
Vadodara, India.
nandini113@gmail.com
ORchid Id: 0000-0002-6178-1328
Nita Patil
Department of Computer Science
School of computer Science
KBCs North Maharashtra University. Jalgaon, India.
nitaapatil@gmail.com

Abstract:

Diseases are very harmful, both physically and mentally. Despite of any external injury, diseases can affect the parts of organism. Diseases are always characterized by specific symptoms and signs. Arteria Coronaria Disease id related to blood flow obstruction. It is one of the most common disease in humans. Heart diseases are not predictable or expected. They often occur suddenly. Machine learning techniques can be applied for prediction of heart disease. In this paper we have used the dataset of UCI repository for heart disease prediction using different parameters. After preprocessing and features selection we have used different machine learning classifiers to train the model. We have performed the heart disease prediction using various trained machine learning models. Techniques like Logistic regression, Decision tree, Support Vector Machine, K Nearest Neighbor, Naïve Bayes and Random Forest algorithms are used in the prediction of heart disease and hybrid of these algorithms provides 83.16 % accuracy.

Keywords—heart disease, machine learning, features selection

I. INTRODUCTION

Disease is any harmful changes in human body that leads to deviation from normal functioning of an organism. Diseases are generally predicted based on symptoms. Lots of issues are faced by healthcare industries including prediction of diabetics, prediction of heart diseases etc. High glucose and cholesterol along with damage of blood vessels leads to develop a heart disease. According to survey of the European Society of Cardiology (ESC) 26 million people from the world were affected by heart disease and several nerve diseases. Amongst them 3.6 million were diagnosed per year. [7] Machine learning techniques are able to automatically learn and experience without being explicitly programmed. The reliability, accuracy and efficiency can be improved using machine learning techniques. Decision support system designed using machine learning techniques are able to achieve high accuracy. They deeply understand decisions. Decision makers will also trust machine learning methods. In hybrid machine learning models we combine the strengths and knowledge representation models of various machine learning algorithms. Collection and analysis of data is one of the major challenge in healthcare industry. Machine learning concept can be used for analysis and prediction

of data using different algorithms and techniques. Supervised machine learning algorithms have c\been widely used for disease prediction. Diabetics management involves different issues including routine checking of blood pressure, blood sugar level and other health status. In this paper we have developed supervised machine learning model for prediction of heart disease. During the classification the data is divided into 70% data for training and 30% data for evaluation. Various steps for heart disease prediction are as follows:

- Data preprocessing: During this step we have normalized various features by removing the mean and scaling to unit variance.
- Features Selection: During this step we have evaluated the performance of different features selection algorithms and selected the recursive features elimination method for features selection.
- Supervised machine learning: In this step we have modelled the hart disease prediction using different machine learning algorithms including Logistic regression, Decision tree, Support Vector Machine, K Nearest Neighbor, Naïve Bayes and Random Forest algorithms. We have also devised and implemented the hybrid ensemble of all these algorithms.
- Performance evaluation: We have evaluated the performance of heart disease prediction system using 70% of data for training and 30%of data for testing. The performance is evaluated in terms if accuracy, precision, recall and f measure.

The major contributions of the proposed research work are as follows:

a. We have evaluated the performance of all classifiers under consideration by considering no features selection technique in terms of accuracy, precision, recall and f measure.
b. Performance of various features selection algorithms is evaluated including univariate features selection technique, recursive features elimination, and model based features extraction.
c. This paper aims to predict the suitable features selection model along with hybrid classification model for designing and modeling heart disease prediction system. The system is able to classify healthy and heart diseased people.

The rest of the paper is organized as below. Section 2 deals with the related work done in heart disease prediction. Section 3 describes the dataset used during this work. Section 4 describes the detailed methodology of the proposed heart disease prediction system. Section 5 deals with the experimental results and finally section 6 concludes the paper with future scope.

II. LITERATURE REVIEW

Nowadays changing lifestyles and hereditary is leading to increasing heart diseases. People are at risk due to heart disease. Values of blood pressure, cholesterol and pulse rate varies from person to person. But medically it is proven that the normal values of Blood pressure is 120/90 and pulse rate is 72. Authors in [1] have provided survey of various classification techniques for prediction of risk level. The heart disease risk is predicted based on gender, age, pulse rate, cholesterol, blood pressure, pulse rate etc. Various data mining techniques considered here includes Naïve Bayes, KNN, Decision

Tree Algorithm, Neural Network. etc., Authors have concluded that the accuracy of predicting risk level is high with maximum number of attributes.

Heart disease prediction is always one of the challenging task for healthcare practitioners. Heart diseases are treated in hospitals and other clinics at expensive therapies and operations. Hence heart disease prediction at early stage is required sothat people can take needful actions before getting more serious. One of the main reason of increasing heart disease now a days is lack of exercise, intake of tobacco, alcohol and improper diet. From many years lots of machine learning algorithms have been proposed for heart disease prediction based on data produced by the health care industry. Authors in [2] have used supervised machine learning techniques including neural networks, SVM, KNN, Naïve Bayes and random forest. The performance of these algorithms have been summarized.

Heart disease detection can be confused with other diseases due to its common symptoms including chest pain, breathing symptom and naushea. Hence authors in [3] have proposed machine learning framework for heart disease prediction. The data obtained from patients is weighted according to success. Weight coefficient is determined using the method proposed. Considering 13 different features, the accuracy of 86.90% is obtained.

Authors in [4] have used K means clustering algorithm and SVM for classification and forecasting of heart disease. The combination of back propagation algorithm along with K means clustering improves the results. Dataset from UCI repository is used for evaluating the performance pf the algorithm. 14 features out of 76 features in the dataset are used for heart disease prediction. The performance is evaluated in terms of accuracy, execution tie and error recognition rate.

Heart disease is caused by blood vessels blocking where heart stops functioning. It is one of the major cause of death. But the risk can be minimized if we predict the heart disease at earlier stage. Authors in [5] have developed a data science framework for discovering the chances of heart disease. Along with different classification algorithms, the importance of various parameters and their influence is predicted on Cleveland cardiovascular medical records. The aim of the paper was finding the optimal classification algorithm for heart disease prediction. Various classifiers considered in this study are Random Forest, Vector support, Logistic regression and XG-Boost.

Authors in [6] have used neural networks for heart disease prediction. Various optimizing algorithms are analyzed along with weight initializing techniques. The effect of different parameters on accuracy is analyzed. The dataset used in this research is the coronary heart disease dataset. Results are compared along with different classification algorithms.

Only chest pain is not the symptom for coronary heart disease. Other factors including blood pressure, cholesterol. Blood sugar, ECG, thalach, number of vessels blocked also result in heart disease. Heart disease if predicted earlier can save life. Authors in [7] have used data mining techniques including Naive Bayes, Decision Tree, K-nearest neighbor, etc., are used in the Heart Disease prediction based on the parameters / factors. The work is specific to identification of various factors for heart disease prediction and their importance.
Authors in [8] has explored the use of the Framingham Risk Model for prediction of heart disease risk using a limited set of attributes present in a health risk assessment (HRA) dataset from a digital health company. HRA does not provide information such as LDL

and HDL cholesterol values which are very essential for heart disease prediction. Hence, data from the National Health and Nutrition Examination Survey (NHANES) data from the Centers for Disease Control (CDC), the United States public health agency is used. Authors found that HRA data can be successfully used as input for the Framingham Risk Model for heart disease prediction.
Large amount of heart disease data has been collected by healthcare industries. But the data is not mined for missing values or for finding hidden patterns. Authors in [9] has used PCS for finding minimum number of attributed need for heart disease prediction sothat the performance of supervised classification techniques can be increased. In this paper different data mining approaches are utilized for dietetics disease prediction.

Authors in [10] have used Convolutional Neural Networks (CNNs) and regular Neural Networks (NNs) for heart disease prediction. Series of experiments have been carried out by tuning the parameters. The performance is evaluated for two models using different parameters. The Cleveland database from UCI learning dataset repository is used for diagnosis heart disease. The experimental results shows that NNs performs best as compared with CNNs in most of the cases.[10]

Authors in [11] have proposed a heart disease diagnosis system using ensemble model. The ensemble model is developed based on SVM, decision tree and ANN. The public dataset of UCI for heart disease prediction is used here. The performance is evaluated for three models along with hybrid ensemble model in terms of accuracy, precision, recall and F measure. Th performance of hybrid ensemble model is better.

Authors in [12] have used data mining classification algorithms for stroke dataset. The proposed recommender system is able to predict the risk level of IHD. Different data mining techniques considered here are Logistic Regression, Decision Tree, K nearest Neighbor, Naïve Bayes and SVM on Ischemic Stroke Dataset. The performance of SVM is best with accuracy of 97.91%.

The data created by medical practitioners is converted into useful dataset by preprocessing. It is necessary for the medical experts to predict the CVD at earlier stage. Probability of heart disease increase due to drinking and smoking, lack of physical activities, high level of vital sign, dangerous extent of cholesterol levels, unhealthy and unhygienic diet, damaging use of alcoholic beverage, and high sugar content level food. Pprophecies associated descriptions are principal goals of information mining; in observe Prediction [13] of the processed data involves the attributes or the physiological variables in the data set to find a prediction probability or future state values of an alternative attributes. The Description will emphasize on the discovering a common recognizable patterns that describes that information that can be understood by humans. Authors in [14] have exploited the use of AI devises for order and expectation of heart illness. SVM and Neural networks are used for classification of heart disease. The performance is evaluated in terms if accuracy and training time. A medical choice breaking algorithm is introduced for heart disease prediction faster. The dataset utilized are the Cleveland Heart Database and Statlog Database taken from UCI Machine learning dataset vault.

Authors in [15] has explored the outfit characterization technique by consolidating different classifiers. The system is able to identify the infection at earlier stage. The consequences of the investigation show that group strategies, for example, stowing and boosting, are important for improving the expectation precision of feeble classifiers, and display palatable execution in distinguishing danger of heart disease.

A real time heart disease prediction system is [proposed in [16]. It is based on apache spark which is a strong large scale distributed computed computing platform. The system is divided into two parts streaming processing and data storage and visualization. Streaming processing is performed by MLlib and data events are classified for heart disease prediction.

Authors in [17] have proposed an efficient heart disease prediction system using machine learning algorithms like Naïve Bayes, Random Forest, K-Nearest Neighbour, Support Vector Machine, Xg-Boost. 13 features are used such as age, gender, blood pressure, cholesterol, obesity, cp, etc. The dataset file is uploaded first and algorithm is selected. The performance is evaluated for all algorithms under consideration. Input for each parameter for heart disease prediction is taken form user and the heart disease is predicted.

Recently lots of machine learning algorithms has been implemented for heart disease prediction. Authors in [18] have proposed methodology for finding significant features using machine learning algorithms. The model performance is evaluated for varying combinations of features and machine learning techniques. The accuracy of 88.7% is achieved with the hybrid random forest with a linear model.

III. DATASET DISCRIPTION

The "Cleveland heart disease dataset 2016" is available online from University of California, Irvine [19]. It is used by number of researchers in past decade. During this paper we have used this dataset for design and development of heart disease prediction framework using hybrid ensemble method. The original dataset of Cleveland heart disease dataset contains 76 features with 303 features. But due to missing values 6 records are removed. And final referred dataset contains 297 records with 13 more significant relevant features and one target feature for presence/absence of heart disease.

The description of 14 features of the dataset is as below:

1. Age: indicates the age of the patient.
2. Sex: indicated the gender of the patient using the following format :
   1 = male
   0 = female
3. Chest-pain type: indicated the chest pain type of the patient using the following format :
   1 = typical angina
   2 = atypical angina
   3 = non — anginal pain
   4 = asymptotic
4. Resting Blood Pressure: indicates the resting blood pressure of patient in mmHg (unit)
5. Serum Cholestrol: indicates the serum cholesterol in mg/dl (unit)
6. Fasting Blood Sugar: compares the fasting blood sugar value of patient with 120mg/dl.
   If fasting blood sugar > 120mg/dl then : 1 (true)
   else : 0 (false)

7. Resting ECG : displays resting electrocardiographic results
   0 = normal
   1 = having ST-T wave abnormality
   2 = left ventricular hyperthrophy
8. Max heart rate achieved : displays the max heart rate of the patient.
9. Exercise induced angina :
   1 = yes
   0 = no
10. ST depression induced by exercise relative to rest: displays the value which is an integer or float.
11. Peak exercise ST segment :
   1 = upsloping
   2 = flat
   3 = downsloping
12. Number of major vessels (0–3) colored by flourosopy : displays the value as integer or float.
13. Thal : displays the thalassemia :
   3 = normal
   6 = fixed defect
   7 = reversible defect
14. Diagnosis of heart disease : Predictive variable for heart disease diagnosis
   0 = absence
   1, 2, 3, 4 = present.

IV. METHODOLOGY

The aim of the proposed work is classification of people with heart disease and healthy people. We have used the dataset of UCI machine repository with 297 records and 13 features. Performance of different machine learning algorithms is evaluated without any features selection technique. In next section we have evaluated the performance of various features selection techniques for heart disease prediction. Based on performance of various classifiers, we have finalized the use of recursive features elimination technique of features selection. Different machine learning algorithms considered here includes Logistic regression, Decision tree, Support Vector Machine, K Nearest Neighbor, Naïve Bayes and Random Forest algorithms. We have devised and implemented hybrid ensemble method considering machine learning algorithms. The proposed methodology of heart disease prediction is divided into five stages including 1. Preprocessing of the Dataset 2. Features Selection 3. Cross validation 4. Machine Learning classifiers 5. Performance evaluation.
Figure 1 depicts the framework of the proposed system.

A. Data Preprocessing.

The preprocessing od data improves the representation of data for use in machine learning classifier. Different preprocessing techniques are used including missing values removal, standard scaler. The standard scaler helps to ensure that the feature has the mean 0 and variance 1, bringing all features to the same coefficient. Standardscaler is used for standardizing the features by removal of mean and scaling to unit variance.
The standard score of a sample x is calculated as:

$$z = (x - u) / s$$

where u is the mean of the training samples or zero if with mean=False, and s is the standard deviation of the training samples or one if with_std=False.

B. Feature Selection Algorithms.

Features selection plays importnt role in machine learning process because irrelevant features may degrade the performance of the classifier. Features slection improves the accuracy of the classifier and resuces execution time. The benefits of applying features selection before modelling the data includes

- Reduces Overfitting: Less redundant data means less opportunity to make decisions based on noise.
- Improves Accuracy: Less misleading data means modeling accuracy improves.
- Reduces Training Time: Less data means that algorithms train faster.

We have evaluated the performance of various features selection techniques and found the recursive features ellimination technique to be suitable for heart disease prediction dataset. During this work we have considered following features selection techniques

1) Univariate Selection

In univariate features selection technique the statistical tests re carried out for selecting those variables having strongest relationship with output variable. SelectKBest class from scikit-learn library is used for selecting specific number of features with a suite of different statistical tests.

2) Recursive Feature Elimination

In the recursive features elimination the attributes are removed recursively and model is built on those attributes that remain. The model accuracy is used for identifying useful features for predicting target features. In this work we have used recursive features elimination using logistic regression to select top 6 features.

3) Model-based feature selection (SelectFromModel)

In some machine learning algorithms importance to dataset features is naturally assigned. One of the example is linear regression. In linear regression a coefficient multiplier is applied to each of the feature. The importance of the variable is defined by the value of the coefficient. Feature ranks are generated by such machine learning models and features are pruned based on that ranking.

C. Machine Learning Algorithms

After feature selection, important features are selected form the dataset. Machine learning models are used for prediction of heart disease. In this work we have used supervised machine learning algorithms including Logistic regression, Decision tree, Support Vector Machine, K Nearest Neighbor, Naïve Bayes and Random Forest algorithms. Hybrid ensemble model is devised and implemented using all these classifiers. The performance of the hybrid model is increased.

D. K Fold Cross Validation

K fold cross validation method is used in this paper along with four performance evaluation metrics. In k fold cross validation we divide the dataset into k equal size. Among these k parts k-1 groups are used for training the classifiers and the remaining one is used for evaluating the performance. The validation process is repeated k times and performance is evaluated. The final classifier performance depends on the k results. In this work we have used 10 fold cross validation and 70% data for training and 30% data for evaluation.

E. Performance Metrics

For evaluating the performance of the classifiers we have used four measures, accuracy, precision, recall and f measure. For computing the performance parameters confusion matrix is needed. Confusion matrix is a 2X2 matrix as depicted in table 1.

TABLE I. CONFUSION MATRIX

	Predicted HD patient (1)	***Predicted healthy person (0)***
Actual HD patient (1)	TP	FN
Actual healthy person (0)	FP	TN

From confusion matrix we are able to get four values.

- TP: the output is predicted as true positive (TP) if we have concluded that the HD record is correctly classified and the record has heart disease.
- TN: the output is predicted as true negative (TN) if we have concluded that the healthy record is correctly classified and the record is of healthy person.
- FP : the output is predicted as false positive (TP) if we have concluded that the healthy record is incorrectly classified as HD record.
- FN : the output is predicted as false negative (FN) we have concluded that the HD record is incorrectly classified as healthy.

Accuracy: Classification Accuracy defines the performance of the classifier and is given by

$$\text{Accuracy} = (TP+TN) / N$$

Where N is the total number of cases.

Precision: Precision defines the fraction of correctly identified heart disease patients and healthy individuals. Precision is defined as

$$\text{Precision} = TP/ (TP+FP)$$

Recall: Recall is defined as the fraction of the HD patients that are successfully predicted.

Recall is defined as

$$\text{Recall} = TP / (TP+FN)$$

F1 score: F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account.

F1 Score = 2*(Recall * Precision) / (Recall + Precision)

V. EXPERIMENTAL RESULTS AND DISCUSSIONS

This section deals with the empirical results obtained in this work. In first section we have evaluated the performance of different supervised machine learning algorithms by considering all features in the dataset. Different machine learning algorithms used here are logistic regression, k-nearest Neighbor, artificial neural network, support vector machine, Naive Bayes, and decision tree on Cleveland heart disease dataset. We have devised and implemented the hybrid ensemble classifier based on these classification techniques. The performance of hybrid method is better as compared with other machine learning approaches. In second section we have evaluated the performance of different features selection algorithms. We found the recursive features elimination technique most suitable for out dataset. Third section deals with evaluating the performance based on features selection using recursive features elimination. K fold cross validation is used here. All computations were performed in Python on an Intel(R) Core™ i5 -2400CPU @3.10 GHz PC.

A. Analysing the performance of various classifiersWithout Featuees Selection

In this section we have evaluated the performance of the proposed heart disease prediction system without features selection. Performance of different classifiers including logistic regression, k-nearest Neighbor, artificial neural network, support vector machine, Naive Bayes, and decision tree is evaluated in terms of precision, recall , F measure and accuracy. Table 2 below depicts the performance. As shown in the table 2 the performance of logistic regression is more as compared with other models under consideration.

Table 2. Model performance for the used classifiers without feature selection.

	Accuracy	Precision	Recall	*F Score*
Logistic Regression	**0.792079**	0.792731	0.792079	0.792327
Decision Tree	0.683168	0.686487	0.683168	0.670524
SVM	0.772277	0.773536	0.772277	0.769024
KNN	0.782178	0.782873	0.782178	0.779655
Naive Bayes	0.782178	0.785055	0.782178	0.778429
Random Forest	0.762376	0.761553	0.762376	0.761662
Hybrid Ensemble	0.762376	0.762451	0.762376	0.759623

Figure 2 depicts the performance comparison of the proposed heart disease prediction system without features selection.

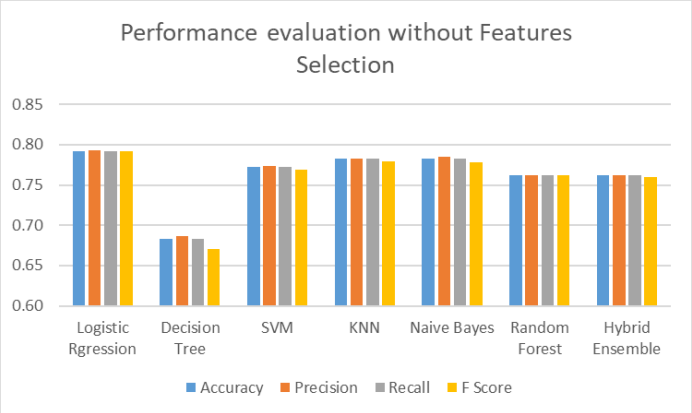

Figure 2: Performance Comparison of heart disease prediction system without features selection.

As depicted in figure 2 logistic regression achieves maximum accuracy of 79% as compared to other models under consideration.

B. Analysing the performance of various Features Selection Techniques

The classifier performance can be increased by eliminating the unrelated features from the dataset and selecting the most appropriate features. In this study we have evaluated the performance of three features election techniques including univariate features selection (UFS), recursive features elimination (RFE) and model based features selection (MSE) . We have evaluated the performance of these features selection techniques using different classification techniques. Table 3 below depicts the accuracy of various classifiers using different features selection methods.

Table 3 Accuracy of Various Classifiers Using Different Feature Selection Techniques

	No Features Selection	UFS	RFE	*MFS*
Logistic Regression	0.792079	0.722772	**0.80198**	0.782178
Decision Tree	0.683168	0.683168	**0.742574**	0.742574
SVM	0.772277	0.693069	**0.792079**	0.772277
KNN	0.782178	0.70297	**0.811881**	0.782178
Naive Bayes	0.782178	0.722772	**0.792079**	0.752475
Random Forest	0.762376	0.70297	0.782178	**0.80198**
Hybrid Ensemble	0.762376	0.722772	**0.811881**	0.792079

As depicted in table 3 , the accuracy of logistic regression without features selection is 79% which is more as compared to all other classifiers. But the performance is increased with recursive features elimination with 3 features consideration, we achieve the maximum accuracy of 81.18%.

Figure 3 depicts the accuracy comparison of all features selection techniques.

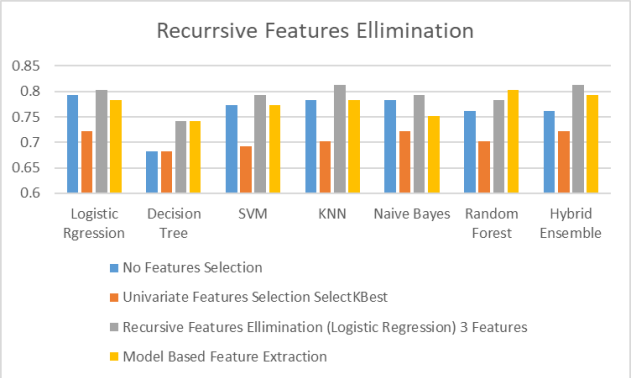

Figure 3: Accuracy of features selection techniques

As depicted in figure 3 the accuracy of almost all classifiers under consideration is improved when we use recursive features elimination.
Table 4 depicts the performance of all features selection methods in terms of precision. Precision is the fraction of predicted records that are relevant.

Table 4 Precision of Various Classifiers Using Different Feature Selection Techniques

	No Features Selection	UFS	RFE	*MFS*
Logistic Regression	0.792731	0.724549	**0.80185**	0.781491
Decision Tree	0.686487	0.686487	**0.758725**	0.749936
SVM	0.773536	0.692347	**0.794128**	0.771468
KNN	0.782873	0.70297	**0.81172**	0.781708
Naive Bayes	0.785055	0.721279	**0.797055**	0.75506
Random Forest	0.761553	0.704817	0.781491	**0.803295**
Hybrid Ensemble	0.762451	0.722772	**0.812577**	0.792304

As depicted in table 4 the precision of recursive features elimination is more for all classifiers under consideration. The hybrid ensemble method devised and implemented here achieves higher precision of 81.25%.
Table 5 depicts the performance of all features selection methods in terms of recall. Recall is the fraction of relevant records that are successfully predicted.

Table 5 Recall of Various Classifiers Using Different Feature Selection Techniques

	No Features Selection	UFS	RFE	*MFS*
Logistic Regression	0.792079	0.722772	**0.80198**	0.782178

Decision Tree	0.683168	0.683168	**0.742574**	0.742574
SVM	0.772277	0.693069	**0.792079**	0.772277
KNN	0.782178	0.70297	**0.811881**	0.782178
Naive Bayes	0.782178	0.722772	**0.792079**	0.752475
Random Forest	0.762376	0.70297	0.782178	**0.80198**
Hybrid Ensemble	0.762376	0.722772	**0.811881**	0.792079

As depicted in table 5 the recall of recursive features elimination is more for all classifiers under consideration. The hybrid ensemble method devised and implemented here achieves higher precision of 81.18%.

Table 6 depicts the performance of all features selection methods in terms of F measure. F measure is the measure that combines precision and recall. It is the harmonic mean of precision and recall.

Table 6 F measure of Various Classifiers Using Different Feature Selection Techniques

	No Features Selection	UFS	RFE	*MFS*
Logistic Regression	0.792327	0.723377	**0.800622**	0.781523
Decision Tree	0.670524	0.670524	**0.743332**	0.743639
SVM	0.769024	0.69264	**0.789109**	0.771178
KNN	0.779655	0.70297	**0.809194**	0.780684
Naive Bayes	0.778429	0.72087	**0.787842**	0.747432
Random Forest	0.761662	0.703619	0.781523	**0.799686**
Hybrid Ensemble	0.759623	0.722772	**0.810167**	0.790185

As depicted in table 6 the recall of recursive features elimination is more for all classifiers under consideration. The hybrid ensemble method devised and implemented here achieves higher precision of 81.01%.

C. *Features Selection Using Recursive Features Ellimination (Logistic Regression)*

In earlier section we have observed the recursive feature elimination technique is the best for out dataset. RFE is an efficient approach for eliminating features from a training dataset for feature selection. RFE works in two stages

- searching for a subset of features by starting with all features in the training dataset : This is achieved by fitting the given machine learning algorithm used in the core of the model and
- successfully removing features until the desired number remains : This is achieved by ranking features by importance, discarding the least important features, and re-fitting the model

This process is repeated until a specified number of features remains.
There are two important configuration options when using RFE:

- the choice in the number of features to select and
- the choice of the algorithm used to help choose features.

In this work we have used logistic regression for features selection using logistic regression. In this study we compare the performance of classifiers by varying number of features sothat we can determine the optimal number of features giving highest accuracy.
Table 7 below depicts the accuracy of various classifiers using recursive features selection method by varying number of features. We have evaluated the performance by varying number of features from 3 to 13. As depicted in the table we found that the maximum accuracy is achieved with 6 important features selected using RFE. The hybrid ensemble model achieves the accuracy of 83.16%.
Table 8 below depicts the precision of various classifiers using recursive features elimination (RFE) method. By considering the varying number of features we found the best performance in terms of precision is achieved with 6 features selected by RFE. The hybrid ensemble model achieves the best precision of 83.53%.
Table 9 below depicts the recall of various classifiers using recursive features elimination (RFE) method. By considering the varying number of features we found the best performance in terms of recall is achieved with 6 features selected by RFE. The hybrid ensemble model achieves the best recall of 83.16%.
Table 8 below depicts the F measure of various classifiers using recursive features elimination (RFE) method. By considering the varying number of features we found the best performance in terms of F measure is achieved with 6 features selected by RFE. The hybrid ensemble model achieves the best F measure of 82.92%.

VI. CONCLUSION AND FUTURE SCOPE

In this paper we have proposed a hybrid ensemble machine learning based heart disease prediction system. The performance of the proposed system was evaluated on Cleveland heart disease dataset with 297 records having 13 attributes. During preprocessing we have used standardscaler and missing values removal techniques. We have evaluated the performance by considering six well known supervised machine learning classifiers including logistic regression, decision tree, SVM, KNN, Naïve Bayes and Random Forest. We have also devised and implemented the hybrid ensemble model considering all these classifiers. We have also evaluated the performance of three features selection methods, univariate features selection (UFS), recursive features elimination (RFE) and model based features selection (MFS).
The performance is evaluated in terms of accuracy, precision, recall and f measure. When no features selection method is applied , logistic regression achieves the maximum accuracy of 79.20%. When using features selection with RFE method the maximum accuracy of 83.16% is achieved by hybrid ensemble model proposed here. The performance in terms of precision, recall and f measure is also improved.
Theis works concludes that the hybrid ensemble method proposed here is more suitable for heart decision support system. Features selection methods also plays important role in increasing the accuracy of prediction. Hence RFE features elimination with logistic regression is most suitable for heart disease dataset under consideration.
In future we will try to increase the performance of the system by considering more features selection techniques. We will also increase the dataset by considering more

records. The system can be extended to real time application of heart disease prediction with the use of AI devises.

ACKNOWLEDGMENT

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g”. Avoid the stilted expression “one of us (R. B. G.) thanks ...”. Instead, try “R. B. G. thanks...”. Put sponsor acknowledgments in the unnumbered footnote on the first page.

REFERENCES

[1] J. Thomas and R. T. Princy, "Human heart disease prediction system using data mining techniques," 2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT), Nagercoil, India, 2016, pp. 1-5. doi: 10.1109/ICCPCT.2016.7530265

[2] R. Katarya and P. Srinivas, "Predicting Heart Disease at Early Stages using Machine Learning: A Survey," 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2020, pp. 302-305. doi: 10.1109/ICESC48915.2020.9155586

[3] A. Erdoğan and S. Güney, "Heart Disease Prediction by Using Machine Learning Algorithms," 2020 28th Signal Processing and Communications Applications Conference (SIU), Gaziantep, Turkey, 2020, pp. 1-4. doi: 10.1109/SIU49456.2020.9302468

[4] M. Chakarverti, S. Yadav and R. Rajan, "Classification Technique for Heart Disease Prediction in Data Mining," 2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT), Kannur, India, 2019, pp. 1578-1582. doi: 10.1109/ICICICT46008.2019.8993191

[5] C. S. Prakash, M. Madhu Bala and A. Rudra, "Data Science Framework - Heart Disease Predictions, Variant Models and Visualizations," 2020 International Conference on Computer Science, Engineering and Applications (ICCSEA), Gunupur, India, 2020, pp. 1-4.doi: 10.1109/ICCSEA49143.2020.9132920

[6] V. Sharma, A. Rasool and G. Hajela, "Prediction of Heart disease using DNN," 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, 2020, pp. 554-562. doi: 10.1109/ICIRCA48905.2020.9182991

[7] G. Shanmugasundaram, V. M. Selvam, R. Saravanan and S. Balaji, "An Investigation of Heart Disease Prediction Techniques," 2018 IEEE International Conference on System, Computation, Automation and Networking (ICSCA), Pondicherry, India, 2018, pp. 1-6.doi: 10.1109/ICSCAN.2018.8541165

[8] A. Mohawish, R. Rathi, V. Abhishek, T. Lauritzen and R. Padman, "Predicting Coronary Heart Disease risk using health risk assessment data," 2015 17th International Conference on E-health Networking, Application & Services (HealthCom), Boston, MA, USA, 2015, pp. 91-96 doi: 10.1109/HealthCom.2015.7454479

[9] B. D. Kanchan and M. M. Kishor, "Study of machine learning algorithms for special disease prediction using principal of component analysis," 2016 International Conference on Global Trends in Signal Processing, Information Computing and

Communication (ICGTSPICC), Jalgaon, 2016, pp. 5-10. doi: 10.1109/ICGTSPICC.2016.7955260

[10] C. -H. Lin, P. -K. Yang, Y. -C. Lin and P. -K. Fu, "On Machine Learning Models for Heart Disease Diagnosis," 2020 IEEE 2nd Eurasia Conference on Biomedical Engineering, Healthcare and Sustainability (ECBIOS), Tainan, Taiwan, 2020, pp. 158-161. doi: 10.1109/ECBIOS50299.2020.9203614

[11] X. Wenxin, "Heart Disease Prediction Model Based on Model Ensemble," 2020 3rd International Conference on Artificial Intelligence and Big Data (ICAIBD), Chengdu, China, 2020, pp. 195-199. doi: 10.1109/ICAIBD49809.2020.9137483

[12] D. B. Mehta and N. C. Varnagar, "Newfangled Approach for Early Detection and Prevention of Ischemic Heart Disease using Data Mining," 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2019, pp. 1158-1162. doi: 10.1109/ICOEI.2019.8862544

[13] S. Kaura, A. Chandel and N. K. Pal, "Heart disease-Sinus arrhythmia prediction system by neural network using ECG analysis," 2019 2nd International Conference on Power Energy, Environment and Intelligent Control (PEEIC), Greater Noida, India, 2019, pp. 466-471. doi: 10.1109/PEEIC47157.2019.8976829

[14] S. Radhimeenakshi, "Classification and prediction of heart disease risk using data mining techniques of Support Vector Machine and Artificial Neural Network," 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 2016, pp. 3107-3111.

[15] B. Keerthi Samhitha, M. R. Sarika Priya., C. Sanjana., S. C. Mana and J. Jose, "Improving the Accuracy in Prediction of Heart Disease using Machine Learning Algorithms," 2020 International Conference on Communication and Signal Processing (ICCSP), Chennai, India, 2020, pp. 1326-1330. doi: 10.1109/ICCSP48568.2020.9182303

[16] A. Ed-Daoudy and K. Maalmi, "Real-time machine learning for early detection of heart disease using big data approach," 2019 International Conference on Wireless Technologies, Embedded and Intelligent Systems (WITS), Fez, Morocco, 2019, pp. 1-5. doi: 10.1109/WITS.2019.8723839

[17] S. Farzana and D. Veeraiah, "Dynamic Heart Disease Prediction using Multi-Machine Learning Techniques," 2020 5th International Conference on Computing, Communication and Security (ICCCS), Patna, India, 2020, pp. 1-5. doi: 10.1109/ICCCS49678.2020.9277165

[18] S. Mohan, C. Thirumalai and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," in IEEE Access, vol. 7, pp. 81542-81554, 2019. doi: 10.1109/ACCESS.2019.2923707

[19] Dua, D., and C. Graff. 2019. "UCI Machine Learning Repository." In. Irvine, CA: University of California, School of Information and Computer Science.

[20] Khairandish, Mohammad Omid, Ruchika Gupta, and Meenakshi Sharma. "A hybrid model of faster R-CNN and SVM for tumor detection and classification of MRI brain images." Int. J. Mech. Prod. Eng. Res. Dev 10.3 (2020): 6863-6876.

[21] Pethuraj, M., M. Uthayakmar, and S. Rajakarunakaran. "Study on ultrasonic assisted drilling of aluminium sillimanite reinforced composites." Int. J. Mech. Prod. Eng. Res. Dev.(IJMPERD) 9.2: 923-932.

[22] Gunjigavi, Sanjeev Kumar S., T. Anil Kumar, and Ashwin Kulkarni. "Study Of Ischemic Heart Disease Among Patients With Asymptomatic Type-2 Diabetes Mellitus In A Tertiary Hospital In South India Using Computed Tomographic Angiography." International Journal of Medicine and Pharmaceutical Sciences (IJMPS) 10 (2020): 9-18.

[23] Sundharavadivel, G., and B. Zipporah Matilda. "A Study On Occupational Stress Among Working Womens." International Journal Of Human Resource Management And Research (Ijhrmr) 8.6 (2018): 113-120.

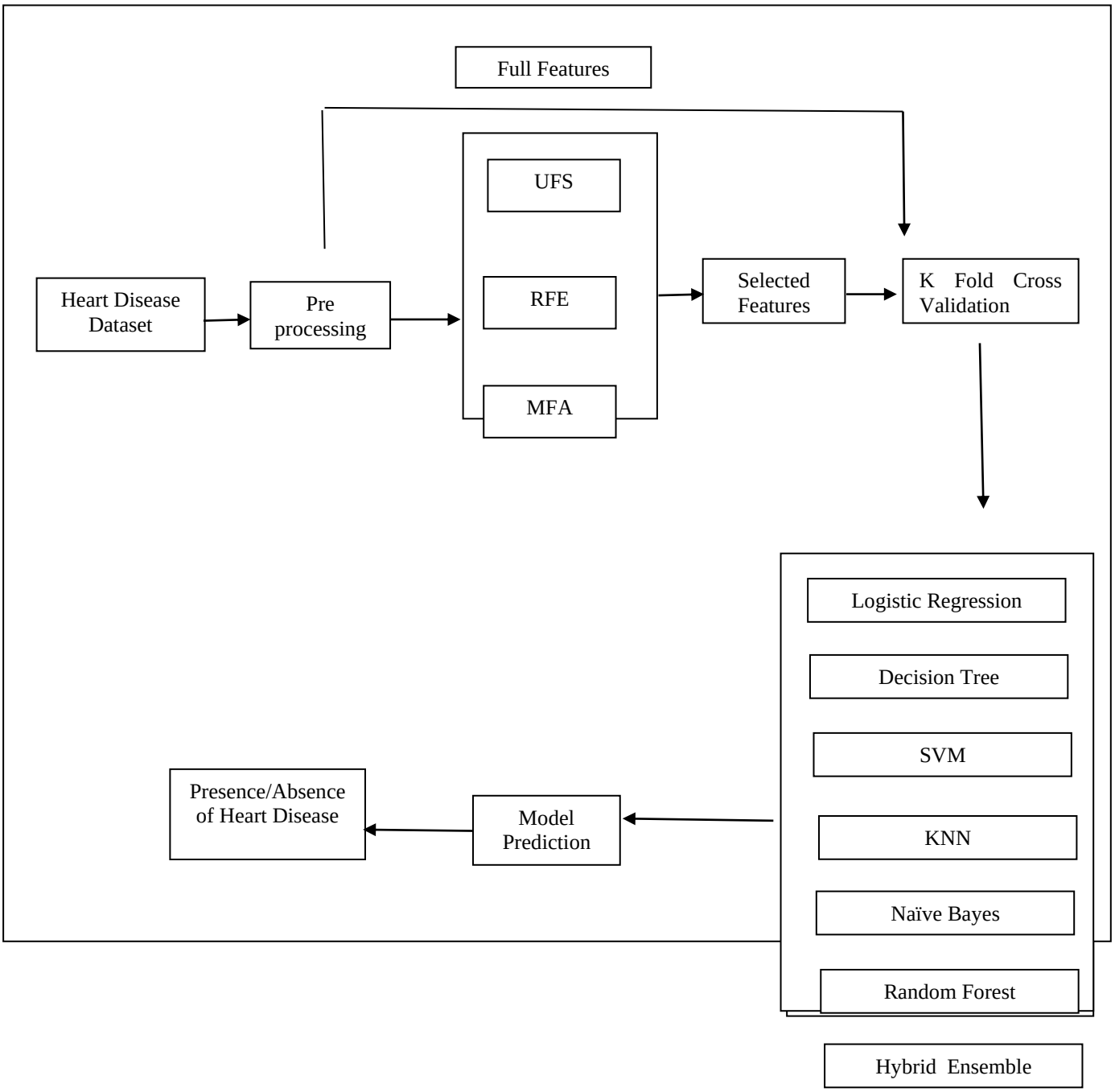

Figure 1 : Architecture of Heart Disease Prediction Syste

Table 7 Accuracy of Features Selection Using Recursive Features Ellimination (Logistic Regression)

	Logistic Regression	Decision Tree	SVM	KNN	Naive Bayes	Random Forest	Hybrid Ensemble
3 features	0.80198	0.742574	0.792079	0.811881	0.792079	0.782178	0.811881
4 Features	0.80198	0.742574	0.811881	0.80198	0.80198	0.811881	0.821782
5 Features	0.792079	0.742574	0.811881	0.792079	0.792079	0.772277	0.80198
6 features	**0.821782**	**0.742574**	**0.80198**	**0.811881**	**0.80198**	**0.811881**	**0.831683**
7 features	0.80198	0.683168	0.742574	0.772277	0.782178	0.772277	0.762376
8 features	0.80198	0.683168	0.792079	0.821782	0.811881	0.80198	0.792079
9 features	0.782178	0.683168	0.762376	0.752475	0.792079	0.772277	0.762376
10 features	0.811881	0.683168	0.762376	0.782178	0.80198	0.782178	0.762376
11 features	0.772277	0.683168	0.782178	0.782178	0.80198	0.782178	0.792079
12 features	0.772277	0.683168	0.752475	0.742574	0.772277	0.80198	0.772277
All features	0.792079	0.683168	0.772277	0.782178	0.782178	0.762376	0.762376

Table 8 Precision of Features Selection Using Recursive Features Ellimination (Logistic Regression)

	Logistic Regression	Decision Tree	SVM	KNN	Naive Bayes	Random Forest	Hybrid Ensemble
3 features	0.80185	0.758725	0.794128	0.81472	0.797055	0.781491	0.812577
4 Features	0.80185	0.758725	0.818053	0.805839	0.805839	0.812577	0.826623
5 Features	0.7915	0.749936	0.812577	0.794128	0.797055	0.771793	0.80185
6 features	**0.821367**	**0.749936**	**0.803295**	**0.81472**	**0.805839**	**0.81472**	**0.835312**
7 features	0.803295	0.686487	0.741615	0.771468	0.78835	0.771793	0.761553
8 features	0.803295	0.686487	0.792304	0.821367	0.822704	0.803295	0.792304
9 features	0.781491	0.686487	0.762451	0.751759	0.797055	0.771793	0.761565

10 features	0.811516	0.686487	0.761565	0.78835	0.809586	0.781491	0.761565
11 features	0.771793	0.686487	0.781491	0.781708	0.809586	0.781491	0.7915
12 features	0.771793	0.686487	0.751437	0.742028	0.771468	0.803295	0.772032
All features	0.792731	0.686487	0.773536	0.782873	0.785055	0.761553	0.762451

Table 9 Recall of Features Selection Using Recursive Features Ellimination (Logistic Regression)

	Logistic Regression	Decision Tree	SVM	KNN	Naive Bayes	Random Forest	Hybrid Ensemble
3 features	0.80198	0.742574	0.792079	0.811881	0.792079	0.782178	0.811881
4 Features	0.80198	0.742574	0.811881	0.80198	0.80198	0.811881	0.821782
5 Features	0.792079	0.742574	0.811881	0.792079	0.792079	0.772277	0.80198
6 features	**0.821782**	**0.742574**	**0.80198**	**0.811881**	**0.80198**	**0.811881**	**0.831683**
7 features	0.80198	0.683168	0.742574	0.772277	0.782178	0.772277	0.762376
8 features	0.80198	0.683168	0.792079	0.821782	0.811881	0.80198	0.792079
9 features	0.782178	0.683168	0.762376	0.752475	0.792079	0.772277	0.762376
10 features	0.811881	0.683168	0.762376	0.782178	0.80198	0.782178	0.762376
11 features	0.772277	0.683168	0.782178	0.782178	0.80198	0.782178	0.792079
12 features	0.772277	0.683168	0.752475	0.742574	0.772277	0.80198	0.772277
All features	0.792079	0.683168	0.772277	0.782178	0.782178	0.762376	0.762376

Table 10 F Measure of Features Selection Using Recursive Features Ellimination (Logistic Regression)

	Logistic Regression	Decision Tree	SVM	KNN	Naive Bayes	Random Forest	Hybrid Ensemble
3 features	0.800622	0.743332	0.789109	0.809194	0.787842	0.781523	0.810167
4 Features	0.800622	0.743332	0.808048	0.798572	0.798572	0.810167	0.818714
5 Features	0.791076	0.743639	0.810167	0.789109	0.787842	0.771959	0.800622

6 features	**0.821246**	**0.743639**	**0.799686**	**0.809194**	**0.798572**	**0.809194**	**0.829279**
7 features	0.799686	0.670524	0.7418	0.771178	0.776998	0.771959	0.761662
8 features	0.799686	0.670524	0.790185	0.821246	0.806721	0.799686	0.790185
9 features	0.781523	0.670524	0.759623	0.75022	0.787842	0.771959	0.760746
10 features	0.811618	0.670524	0.760746	0.776998	0.797271	0.781523	0.760746
11 features	0.771959	0.670524	0.781523	0.780684	0.797271	0.781523	0.791076
12 features	0.771959	0.670524	0.751281	0.739592	0.771178	0.799686	0.770202
All features	0.792327	0.670524	0.769024	0.779655	0.778429	0.761662	0.759623