International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Combined Hybrid Feature Selection and Classification for Heart Disease Prediction in The Cloud-Based Iot Health Care System Using Machine Learning

N. Keerthika[1] Dr. S. Nithyanandam[2]

[1]Ph.D. Research Scholar, Department Of Computer Science And Engineering, Ponnaiyah Ramajayam Institute Of Science And Technology (Prist) Deemed To Be University, Thanjavur
[2]Professor, Department Of Computer Science And Engineering, Ponnaiyah Ramajayam Institute Of Science And Technology (Prist) Deemed To Be University, Thanjavur

Abstract
Health care Management System (HMS) is a key to successful management of any health care industry. Health care management system has so many research dimensions such as identifying disease and diagnostic, drug discovery manufacturing, Bioinformatics' problem, personalized treatments, Patient image analysis and so on. Heart Disease Prediction (HDP) is a process of identifying heart disease in advance and recognizes patient health condition by applying techniques on patient heart related symptoms. Now a day's the problem of identifying heart diseases are solved by machine learning techniques. In this paper we are constructed heart disease prediction method using combined feature selection and classification machine learning techniques. According to the existing study the one of the main difficult in heart disease prediction system is that the available data in open sources are not properly recorded the necessary characteristics and also there is some lagging in finding the useful features from the available features. The process of removing inappropriate features from an available feature set while preserving sufficient classification accuracy is known as feature selection. A methodology is proposed in this paper that consists of two phases: Phase one employs two broad categories of feature selection techniques to identify the efficient feature sets and it is given to the input of our second phase such as classification. In this work we will concentrated on filter based method for feature selection such as Chi-square, Fast Correlation Based Filter (FCBF), Gini Index (GI), RelifeF, and wrapper based method for feature selection such as Backward Feature Elimination (BFE), Exhaustive Feature Selection (EFS), Forward Feature Selection (FFS), and Recursive Feature Elimination (RFE). The UCI heart disease data set is used to evaluate the output in this study. Finally, the proposed system's performance is validated by various experiments setups.

Keywords: Health care Management System, Heart Disease Prediction, machine learning techniques, feature selection techniques, classification, Filter FS, Wrapper FS, FCBF, EFS, FFS, RFE, BFE, Chi-square**,** GI, RelifeF.

Introduction

Health care Management System (HMS) is a key to successful management of any health care industry. It is used to build a long term relationship between a patient and health care service providers. Health care service providers help a patient through hospitals infrastructure, emergency medical service including ambulance service, pharmacy, and stack holders in hospital system such as doctors of various specialization, nurses and other working professionals in the hospital. Health care management system has so many research dimensions such as identifying disease and diagnostic, drug discovery manufacturing, Bioinformatics' problem, personalized treatments, Patient image analysis and so on [1]. Disease prediction is a process of identifying disease in advance and recognizes patient health condition by applying techniques on patient health symptoms. The goal of disease prediction system is to save a life of humans in advance. This predictive disease modeling is applied in all human parts such as lungs, heart, eyes, kidney, brain, digestive systems and so on. Based on the research outcome the brain tumors, mental disorder, blood cancer, breast cancer, cardiac arrest, coronary heart problem, lung infection, asthma are the high probability dangerous diseases found in all over the world [2]. In earlier days this disease prediction system was done in manual way for a countable number of patients, but now days it is not possible due to the large volume of patient records. The globalization of health care system supports the international movement of health care specialties, professionals, patient records and sharing knowledge and information's. The collected large of volume of globalized patient data is used to identify the disease with higher accuracy compared to small dataset.

Any illness that affects the heart is referred to as heart disease. In worldwide every year, 17.9 million people dying due to cardiovascular disease based on the conducted survey of World Health Organization (WHO). Heart Disease Prediction (HDP) is a process of identifying heart disease in advance and recognizes patient health condition by applying techniques on patient heart related symptoms such as chest pain, breathing difficult, stomach pain, fatigue, an irregular heartbeat, sweating and so on. Heart disease prediction is not a easy task because of it may affected a patient by several reasons like smoking, cholesterol, uncontrolled blood pressure, obesity, uncontrolled diabetes, uncontrolled stress, depression and anger. And also it has different types such as Coronary Artery Disease (CAD), Mitral valve regurgitation, Congenital heart defects (Abnormal heart valves, Septal defects, Atresia), Dilated cardiomyopathy, Heart Arrhythmias (Tachycardia, Bradcardia, Premature contractions, Atrial fibrillation) , Myocardial infarction, Hypertrophic cardiomyopathy, Mitral valve prolaps, Aortic stenosis, Pericardial disease, and Heart muscle disease [3]. So it is very important to find a solution through available artificial intelligent techniques, soft computing techniques, machine learning, and optimization techniques and so on. For predicting heart disease so many base classification algorithms are applied such as Decision tree (DT), K-Nearest neighbor (KNN), Regression analysis (linear (SLR), Multiple (MLR), and logistic (LR)), Naïve Bayes (NB), Support Vector Machine (SVM), Random Forest (RF), Artificial Neural Networks (ANN) in advance [4-5]. Most of research studies in heart disease prediction system show that single classification model does not produce a satisfactory result to health care management system. Now a day's researchers used an ensemble classification techniques and hybrid models which is combination of two or more classification or clustering followed by classification to improve the quality of the model. Kavitha et al. (2021) proposed a novel combined random forest classification with decision tree enhanced model for heart disease prediction applied on UCI Cleveland heart disease dataset. Normally hybrid model is used to combine the befit of the *N* number of models in a hierarchal manner to enhance the prediction accuracy [6]. The main intend of this work is to create a best fitting model for predicting heart disease via

prediction accuracy, precision, recall and reduce the error value in the health care industry. The first difficult in patient disease prediction system is that the total number of patient belong to one class is higher than the number of patient belong to another classes that is called Class imbalance problem. Heart disease prediction is comes under binary and multi classification based on the total number of distinct values in prediction process. Class imbalance problem degrade the performance of the prediction system and also misclassify the new data which are comes under the improperly trained minor class [7]. The second difficult in prediction system is the necessary characteristics for predicting the particular heart disease which are not properly recorded in data set creation process. And also there is some lagging in finding the useful features from the available features. The process of removing inappropriate features from an available feature set while preserving sufficient classification accuracy is known as feature selection. The features chosen are critical because they can have a direct correlated to the outcomes of all application oriented data sets. A methodology is proposed in this paper that consists of two phases: Phase one employs two broad categories of feature selection techniques to identify the efficient feature sets and it is given to the input of our second phase such as classification. The UCI heart disease data set is used to evaluate the output in this study. Finally, the proposed system's performance is validated by various experiments setups.

Literature Survey

Health care management system has so many research dimensions such as identifying disease and diagnostic, drug discovery manufacturing, bioinformatics' problem, personalized treatments, patient image analysis and so on which are solved by artificial intelligence techniques are tabulated in Table1. Uddin et al. (2019) surveyed the performance of base classification algorithms (LR, SVM, DT, RF, NB, KNN and ANN) in the field of disease prediction via Scopus and PubMed databases papers which are published in the year from 1999 to 2018. Then they identified 48 unique papers based on various diseases such as asthma, breast cancer, cerebral, diabetes, heart, hemoglobin, hypertension, kidney disease, liver disease, lung, micro RNA, Parkinson's disease, prostate cancer, and stroke. They identified the usage percentage for this seven classification algorithms (LR, SVM, DT, RF, NB, KNN and ANN) and all these base classification algorithms are evaluated based on confusion matrix and ROC curve. Finally they concluded which algorithm gives maximum accuracy to each and every identified 14 diseases [8]. Sliwoski et al. (2104) studied the drug discovery system such as structure based and ligand based methods corresponding to the real world problems [9].

Table 1: Literature survey for health care management system (HMS)

S. No	Dimension of HMS	Techniques	References
1	Disease prediction	Base classification models	Uddin et al. (2019)-[8]
2	Drug discovery	Structure & ligand models	Sliwoski et al. (2014)-[9]
3	Bioinformatics problems	Machine learning models	De Heredia et al.(2016)-[10]
4	Personalized treatments	Dynamic models	Saez et al.(2020)-[11]
5	Patient image analysis	Deep learning models	Gozes et al. (2020)-[12]

De Heredia et al. (2016) surveyed the fundamental difficulties faced in the research of gene expression in RNA sequence data. The first difficult in patient bioinformatics system is that the massive amount records found in the dataset. Also they list out the various solutions to overcome the problems using various

machine learning data processing techniques [10]. Saez-Rodriguez et al. (2020) proposed dynamic models for personalized treatment for a patient instead of static data. They proved that the dynamic model yield a higher accuracy compared to static data, because in dynamic models include the patient specific data [11]. Gozes at al. (2020) build an automatic artificial intelligent tool for prediction corona virus infection through the patient thoracic feature included CT image. The training dataset is modeled by 2D and 3D deep learning techniques and the learned training model is used to predict the accuracy of the collected 157 international patients from US and china and evaluated by increased sensitivity and reduced specificity [12].

Disease prediction is a process of identifying disease in advance and recognizes patient health condition by applying techniques on patient health symptoms. This predictive disease modeling is applied in all human parts such as lungs, heart, eyes, kidney, brain, digestive systems and so on which are solved by data mining techniques are tabulated in Table 2. Monsi et al. (2019) discovered the performance of Conventional Neural Network (CNN) in the field of Lung disease prediction via chest X-rays (1024 X 1024 pixels). They collected 112,110 samples of chest X-ray images from NIH- X-ray data source for 30000 patients who have 14 different diseases such as Atelectasis, Edema, Mass, No finding and so on. They applied normal pre-processing steps like resize the image and normalized the color of the image. The training data modeled by two rotations (base model and retrain model) to boost the accuracy of the training processes [13]. Patel et al. (2015) predicted the performance of base classification algorithms (DT, J48, LR, and RF) in the field of heart disease prediction through Cleveland data set (303 samples and 76 attributes). The drawback of the system is that they consider only 12 features out of 75 features and they concluded J48 yield the better accuracy compared to other three algorithms based on train error and test error measures [14].

Table 2: Literature survey for different disease prediction systems

S.No	Dimension of HMS	Techniques	References
1	Lung Disease prediction	CNN	Monsi et al. (2019) –[13]
2	Heart Disease prediction	DT, J48, LR, RF HRLFM OFBAT-RBFL RF+RS	Patel et al. (2015)-[14] Mohan et al.(2019)- [1] Reddy et al.(2017)-[17] Yekkala et.al (2018)- [18]
3	Kidney Disease prediction	DT, NB	Sathya et al.(2018)-[15]
4	Brain Disease prediction	KNN+MLP	Mathur et al.(2019)-[16]

Sathya priya et al. (2018) studied the performance of base classification algorithms (DT, NB) in the field of kidney disease prediction through chronic kidney disease data set which consist 400 samples and 25 attributes like age, pc, sod, dm, and etc. They used hold-out method for data splitting (training and test data set) and they generated the model according to decision tree and naïve bayes algorithms then they concluded DT yield the better classification accuracy compared to naïve bayes algorithm based on accuracy, sensitivity and specificity [15]. Mathur et al.(2019) studied the performance of base classification algorithms (Combined KNN and Bagging, Ada-boosting.M1, MLP (Multilayer Preceptron)) in the field of Parkinson disease prediction (comes under brain disease) through UCI tumor disease data set which consist 195 samples and 24 attributes like Fo, NHR, D2 and etc. They used 10 fold cross validation method for data splitting (training and test data set) and they generated the model according to

Combined KNN and Bagging, Ada-boosting.M1, MLP (Multilayer Preceptron)) algorithms then they concluded KNN-MLP yield the better classification accuracy compared to other hybrid algorithms based on accuracy, error, time taken by build model, precision, Recall and F-measure [16]. According to Literature survey most of the researchers used single classification algorithms, ensemble classification algorithms, clustering based algorithm, Neural network based algorithms, Deep learning concepts, Optimization techniques are used to find the heart disease prediction. Mohan et al. (2019) suggested a hybrid approach (HRLFM) which collaborate the Linear Method (LM) and Random Forest (RF) for prediction of heart disease [1]. Reddy, G et al. (2017) proposed a novel method for heart disease prediction (Cleveland, Hungarian and Switzerland datasets) system based on OFBAT-RBFL (Oppositional firefly BAT- rule-based fuzzy logic) which obtains the maximum accuracy of 78% [17]. Yekkala et.al (2018) proposed a novel heart disease prediction by using three different classifications: KNN, RF, NB and one common Rough Set (RS) feature selection. [18]. But still there is lagging in that prediction process that is no one is identified the best features for each and every disease prediction models. A methodology is proposed in this paper that includes filter and wrapper based feature selection techniques before classification.

3. Heart Disease Prediction models

Figure 1 show the step by step process involved in our proposed system. Every data analytics problem starts with the data collection so first we collected the data for our application such as heart diseases prediction from UCI machine learning repository [19]. The second step is preprocessing our considered cleveland.data by simple preprocessing steps which used to increase the consistency and accuracy of the system. Then the preprocessed data is given to feature selection techniques used to find the effective features for model building. Then we generated model using base classification algorithms like DT, RF, SVM, KNN, and NB. Finally, the proposed system's performance is validated by various experiments setups.

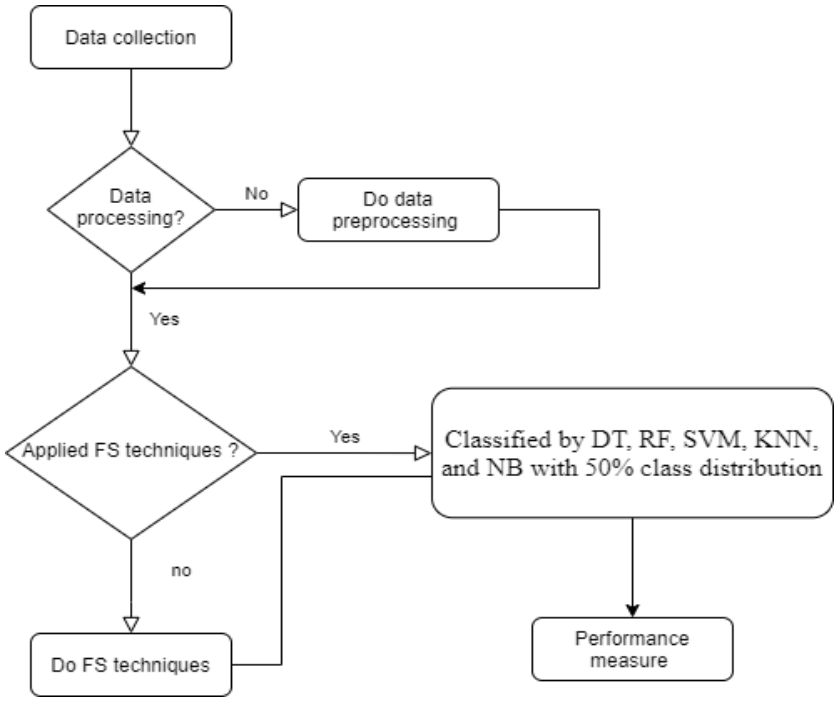

Fig 1: Feature selection based heart disease prediction system framework.

3.1 Dataset description and preprocessing

So many bench mark dataset are available in HDP system like cleveland.data provided by Cleveland clinic foundation, hungarian.data provided by Hungarian institute of cardiology, long-beach-va.data provided by V.A.medical center located in long beach and Switzerland.data provided by university zurich located in Switzerland. We considered Cleveland heart diseases data found in UCI machine learning repository which consists of 303 patients and 13 features (age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal) with 1 class labels (num). The class label has 5 distinct entries such as 0 to 4, here 0 represent the number of healthy patients and 1 to 4 represents the level of the disease affected by a patient. Out of 303 patients 164 patients are healthy patients and remaining 139 patients are affected by different level of heart disease like 1, 2, 3 and 4. Here 55 patients are comes under the level1, 36 patients are comes under the level2, 35 patients are comes under the level 3 and 13 students are comes under the level 4. All 13 features are in discrete and continuous form [19]. In this data set some of the attributes has less no of missing values these missing values are replaced by mean of that attribute.

3.2 Feature selection Techniques

The process of removing inappropriate features from an available feature set while preserving sufficient classification accuracy is known as feature selection. The features chosen are very important process because it can have a direct correlation to the outcomes of all application. Feature selection techniques are classified in to filter based, wrapper based, embedded based and hybrid based feature selection. Information gain, ReliefF, Gain ratio, Fast correlated based filter, Chi-Square test, Fishers score, Correlation coefficient, Variance threshold, Mean absolute difference, Dispersion ratio, Interact are some of the filter based feature selection techniques. Forward feature selection, backward feature elimination, Exhaustive feature selection, Recursive feature selection, Best first feature selection, Hill climbing feature selection are some of the wrapper based feature selection. Lasso regularization, Random forest importance are some of the embedded based feature selection techniques. Particle Swarm based feature selection, Fuzzy based feature selection, Rough set theory based feature selection are some of the hybrid based feature selection [20]. In this work we will concentrated on filter based method for feature selection such as Chi-square, Fast Correlation Based Filter (FCBF), Gini Index (GI), RelifeF and wrapper based method for feature selection such as Backward Feature Elimination (BFE), Exhaustive Feature Selection (EFS), Forward Feature Selection (FFS), and Recursive Feature Elimination (RFE). The filter based feature selection method is based on applying some of the statistical operation to each and every feature which is correlated to outcome of the dataset and the best features set are generated based on maximum score which are represented in Fig 2 (a) [21]. The wrapper based feature selection is totally opposite to filter based techniques, here the subset of features are randomly chosen and given for module building. Based on the model outcome in the next iteration the process will add some more features which are represented in Fig 2 (b) [22].

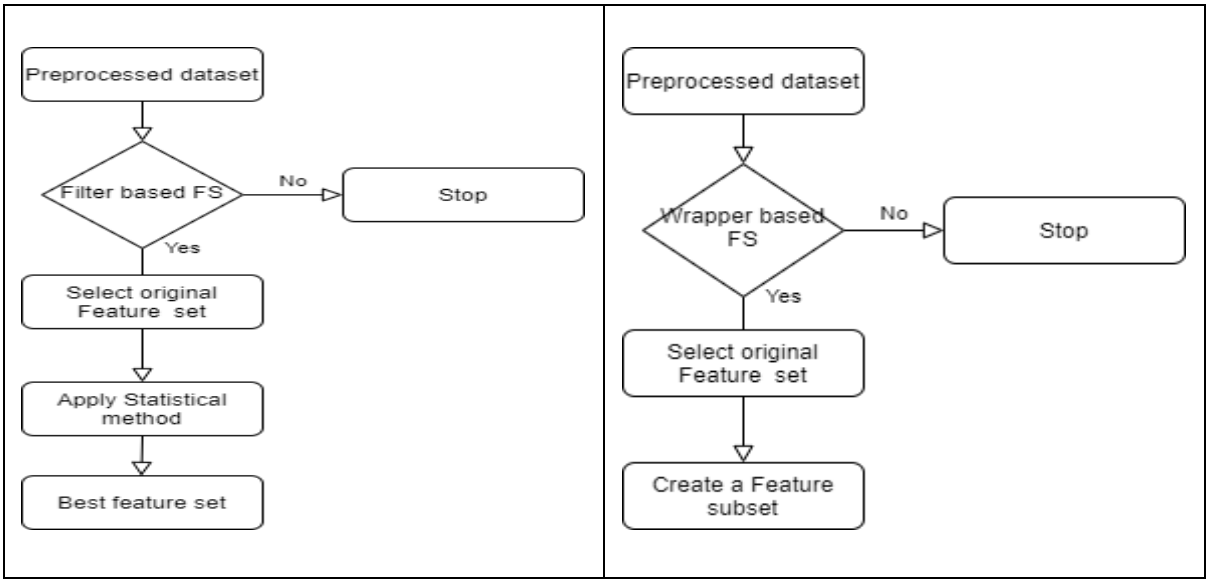

Fig 2: Frame work for filter (a) and wrapper (b) based feature selection

3.2.1. Chi-square

Chi-square method is an extended version of correlation based feature selection and it suits for both nominal and numerical attributes. Normally chi-square method has 2 hypotheses set by the user like H_0 represents no association between one attributes to class label and H_1 represents there is some association between selected attribute to class label. For this process we have to calculate the chi-square value (X^2) based on expected value (e) and the original similar value (o) to the following Eq. (1).

$$X2 = \sum \frac{(o-e)2}{e} \qquad \text{Eq. (1)}$$

The calculated chi-square value is greater than the significant chi-square value (based on degrees of freedom and significant value) means the calculated value is found in the colored region it tell to us to reject our null hypothesis which means we will accept our alternate hypothesis such as there is some association between selected attribute to class label. We are selected all the efficient features based on the same procedure [23].

3.2.2. Fast Correlation Based Filter (FCBF)

The Fast Correlation Based Filter (FCBF) is a filter based feature selection technique proposed by Yu et al. (2003) for high dimensional data with and a threshold value (I). [24]. It identified an efficient feature sub set (FSe) based on correlation between the input dataset with N features and class label. First step ,the algorithm calculate the similarity value (S) for each and every attribute to class label then the calculated similarity value for a feature is greater than the threshold means ($S > I$) the particular feature is added to the efficient feature set. In next step the identified feature set is ordered from left to right based on maximum similarity function. In the last step it verifies any redundancy variables are available in the obtained feature or not. Any redundant attributes are found in the efficient feature set means it eliminate that redundant attribute then gives the final feature set to the user.

3.2.3. Gini Index (GI)

Gini Index method is an extended version of information gain and gain ratio which utilize the entropy concept proposed by Claude Shannon [25]. Gini index measures the impurity of data set D and a data partition or set of training tuples based on the following Eq. (2) - Eq. (4). Here D is represent the dataset consist of training tuples and class labels for each and every tuples and m is the number of class label in

considered dataset and *p* represent the probability values with respect to class labels and D_i represent the number distinct values found in a particular attribute *A*. Finally the informative features are selected based on maximum Gini index value.

$$Gini(D) = 1 - \sum_{i=1}^{m} p_i^2 \quad \text{Eq. (2)}$$

Gini of particular attribute

$$Gini_A(D) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2) \quad \text{Eq. (3)}$$

$$\Delta Gini(A) = Gini(D) - Gini_A(D) \quad \text{Eq. (4)}$$

3.2.4. ReliefF

ReliefF is one of the advanced version relief filter based feature selection techniques based k-nearest neighbor concept which are utilized by both binary classification problems and multi classification applications [26]. It identified an efficient feature sub set based on nearest miss (*m*) and hit (*h*) instance value between the randomly chosen data sample (R_X) to other sample in the dataset. Initially all the feature weights (*W*) are assumed as 0 then it is updated with the following Eq. (5) for each and every attribute (*A*) with *v* training instances.

W_U = *W*- difference ($R_{X,,}$ *A, h)/v*+ difference ($R_{X,,}$ *A, m)/v* Eq. (5)

Here difference *($R_{X,,}$ A, h)* is calculated based on Euclidian distance between the parameters such as $R_{X,,}$ A, h.

3.2.5. Backward Feature Elimination (BFE)

Backward feature elimination is a greedy based wrapper feature selection technique which is completely opposite to forward feature selection [27]. It follows the top down approach so in first iteration it includes all the features for model building and check the efficiency of the system, Then in next iteration onwards it eliminate some of irrelevant features from the features set and proceed the same step until it will reach the maximum accuracy is listed here.

1. Start with *N* number of features
2. Eliminate some of the *i* irrelevant features from *N (each iteration)*
3. Terminate the process with efficient features

3.2.5 Exhaustive Feature Selection (EFS)

Exhaustive Feature Selection is a greedy based wrapper feature selection technique which is a combination of forward feature selection and backward feature elimination [28]. The upside EFS follows the top down approach so in first iteration it includes all the features for model building and check the efficiency of the system, Then in next iteration onwards it eliminate some of irrelevant features from the features set and proceed the same step until it will reach the maximum accuracy is listed here.

1. Start with *N* number of features
2. Eliminate some of the *i* irrelevant features from *N (each iteration)*
3. Terminate the process with efficient features

The down side EFS follows the bottom down approach so in first iteration it has null feature set, then in next iteration onwards it added some of most efficient relevant features from features set for model building and check the efficiency of the system and proceed the same step until it will reach the maximum accuracy is listed here.

1. Start with *N*ull feature set

2. Add some of the *i* relevant feature from *N (each iteration)*
3. Terminate the process with efficient features

3.2.7. Forward Feature Selection (FFS)

Forward feature selection is a greedy based wrapper feature selection technique which is completely opposite to backward feature elimination [29]. It follows the bottom down approach so in first iteration it has null feature set, then in next iteration onwards it added some of most efficient relevant features from features set for model building and check the efficiency of the system and proceed the same step until it will reach the maximum accuracy is listed here.

1. Start with *N*ull feature set
2. Add some of the *i* relevant feature from *N (each iteration)*
3. Terminate the process with efficient features

3.2.8. Recursive Feature Elimination (RFE)

Recursive feature elimination is a wrapper based feature selection technique that created a model based on coefficient function and cross validation for each and every attributes and ranked them. The model is used to eliminate the weakest features from each feature set based on dependencies and co-linearity [30].

3.3. Classification Algorithms

The classification algorithms are used to find the class label of or (classify) any test data according to the model generated by them using application oriented real world historical data. The application data is collected and preprocessed first using available preprocessing concept in technical study then it divided into training and testing data using some framework. Generally training data used to generate the model and test data is predicted based on the generated model it is purely dependent on the particular data set and the classification algorithm which is used in the model building process is highlighted in Fig. 3.

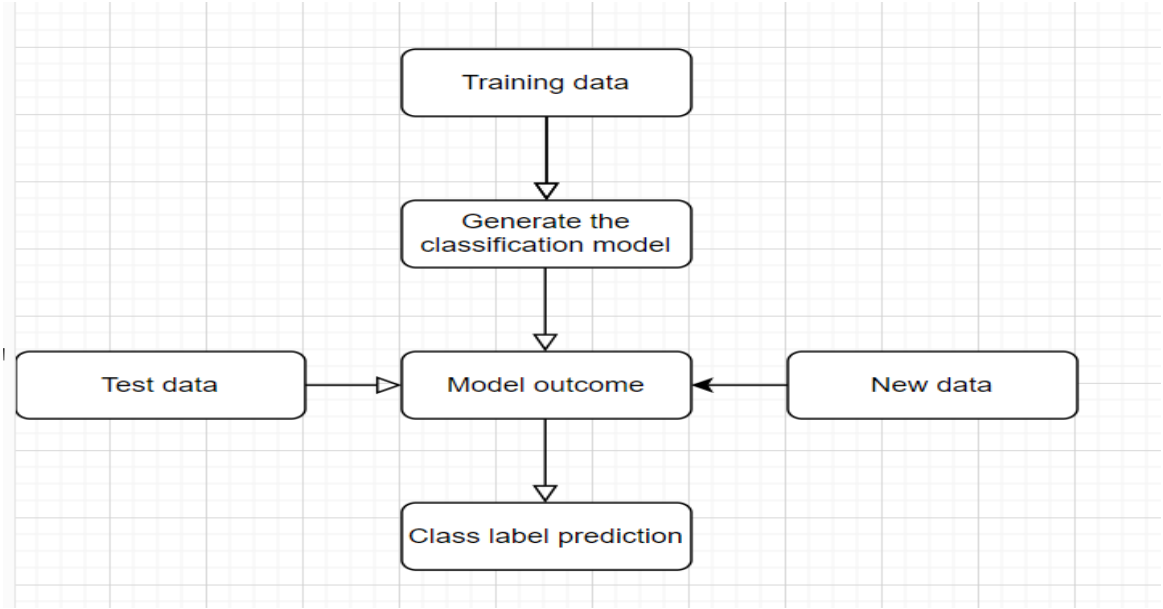

Fig 3: General Frame work for any classification algorithm

3.3.1. Decision Tree (DT)

Decision tree is one of the old classification algorithm proposed by Quinlan in the year of 1970. Here we are going to generate the decision tree depending up on the training data then the decision tree used to

derive the *N* number of if –then classification rules from root node to leaf node which cover all the class label of your dataset. The constructed rules are used find the class label of the upcoming test data and also find the class label of the existing data. The decision tree has three components such as root, internal nodes and leaf nodes. Root and internal nodes represent the testing of each attribute with respect to their ranking. The number of branches in each and every level is depending on the number of distinct values in that attribute. Finally the leaf node holds the class label (i.e. outcome) of the dataset [14].

3.3.2. Random Forest (RF)

Random forest is the extended version of decision tree algorithm with added benefits of ensemble based techniques. Usually embedded based techniques are used to increase the model accuracy based on *N* number model generated from *N* number of training dataset instead of a single model which are used in base classification algorithms. For generating *N* numbers of models we need to build *N* number of training dataset from original dataset. The RF algorithm build *N* number of training dataset based on varying the feature size for each and every set. Then the test data is predicted majority voting provided by each and every model. So many ensemble classification models are available in the data analytics field, but most of the researchers concluded RF is the efficient ensemble model based on decision tree background [31].

3.3.3. Support Vector Machine (SVM)

Support vector Machine is one of most important classification algorithm proposed by Vapnik et al in 1992. SVM algorithm is better than other classification algorithm depends up on the accuracy and also it suits for linearly separable data points and non-linearly separable data points. Most of the researchers used SVM classification algorithm for their application models such as business intelligence, time series analysis, educational sector, image based prediction, health care system, pattern recognition, and test mining. But SVM algorithm takes much time to training process compared to other classification algorithms such as DT, NB, KNN and NB. The SVM classification algorithm has three components such as hyper plane, support vector, and margin. The hyper plain is used to separate the data samples from one class to other classes and it similar to a line equation shown in Eq. (1) which has one independent attribute and dependent attribute. The Equation is expanded based on the number of attributes in the data set. The Eq. (6) is used to find the intercept and coefficient value for each and every independent variables [31]. The next step is to find the support vectors for each and every class separately based on the distance measure between the hyper plain to each and every data sample in each class. The data point getting minimum distance is act as a support vector for that class. Then we can draw the marginal lines which touch the support vectors and also it is parallel to hyper plan. After this model building the test data is classified depending on the $d(X^T)$ based on the Eq. (7). It is depend on the numbers of support vectors in the dataset (l), the class label (y_u) of the support vectors (Y_u), attribute values of the test data (X^T), and constant values (α_u, b_0).

$$u.X+v=0 \qquad \text{Eq. (6)}$$

$$d(X^T)= \sum y_u \alpha_u X_u X^T + b_0 \qquad \text{Eq. (7)}$$

3.3.4. Naïve Bayes (NB)

Naïve bayes algorithm is a probabilistic based classification algorithm which depends on the bayes theorem. Here each and every test data probability is calculated depend on each and every class label in our dataset. That parameter *P (C|S)* is called posterior probability of a test data *(S)* depend on each and every class label *(C_i)* calculated by Eq. (8). *P(S)* is always constant value for each and every class label *(C_i)* so the posterior probability is depend on the following two component such as *P (S|C)* and *P(C_i)*,

where $P(S|C_i)$ is prior probability of a test data (S) depend on each and every class label (C_i) and $P(C_i)$ is the prior probability of each and every class label (C_i). Finally the test data is predicted similar to maximum probability class label [31].

$$P(C_i|S) = P(S|C_i) \text{ X } P(C_i)/ P(S) \quad \text{Eq. (8)}$$

3.3.4. K-Nearest Neighbor (KNN)

K-Nearest Neighbor algorithm is one of the simple lazy classification algorithms based on the distance measure. The application data is collected and preprocessed first using available preprocessing concept in technical study then it is divided into training and testing data using some framework. Then find the Euclidian distance measure between each and every training data to all test data. Each and every test data is classified based on the maximum no of voting provided by *N*-Nearest neighbor's samples. The KNN algorithm takes less time and giving moderated accuracy values for all the application but it is purely dependent of the *N* value (Number of nearest neighbors) , this *N* value is chosen based on trial and error method [31].

4. Experimental setups and Results

The proposed system's performance is validated by various experiments setups such as in setup I the performance of the chosen classification algorithms (DT, RF, SVM, KNN, and NB) without feature selection is evaluated and in setup II the performance of the chosen classification algorithms mapped with each and every chosen feature selection techniques Chi-square, Fast Correlation Based Filter (FCBF), Gini Index (GI), RelifeF, Backward Feature Elimination (BFE), Exhaustive Feature Selection (EFS), Forward Feature Selection (FFS), and Recursive Feature Elimination (RFE) is evaluated and in setup III the proposed system's performance is validated with existing systems like HRLFM, RBFL+OFBAT, and RF+RS proposed by (Mohan et al.(2019)- [1], Reddy et al.(2017)-[17], Yekkala et.al (2018)-[18]) in this study.

4.1 Performance measures

The heart disease prediction application is comes under the multi-classification problem which has more than 2 class labels that is already we discussed under the data set description section. The class label has 5 distinct entries such as 0 to 4, here 0 represent the number of healthy patients and 1 to 4 represents the level of the disease affected by a patient. The proposed model classification involving attributes selection and evaluated by the following evaluation parameters like average accuracy, error rate, precision, recall, F-score for micro and macro averaging. Table 3 represents the confusion matrix for multi-classification with 5 class labels. Here Z_{11} represent the number of healthy patients are predicted as healthy, Z_{22} represents the number stage 1 patients are predicted as stage1, Z_{33} represents the number stage 2 patients are predicted as stage2, Z_{44} represents the number stage 3 patients are predicted as stage3, Z_{55} represents the number stage 4 patients are predicted as stage 4. Accuracy in this problem is the ratio between the correctly predicted patients from each and every group compared to over all patients in this application. It is depend on four parameters like tp_i (True Positive), tn_i (True Negative), fp_i (False Positive) and fn_i (False Negative) represented in the table 4 for each and every class labels ($i = 1$ *to* l). Similarly we have to calculate the following parameters average accuracy, error rate, precision, recall, F-score for micro and macro averaging for each and every class ($l = 1$ *to* 5) in our 5 classification problem through the following equations Eq. (9) - Eq. (16) with $\beta=1$.

The aim the proposed model is to increase the average accuracy, reduce the error rate, and increase the precision, recall and F-score value for micro and macro averaging.

Table 3: Confusion matrix for heart disease 5 label classification problem

Actual	Predicted				
	0	1	2	3	4
0	Z_{11}	Z_{12}	Z_{13}	Z_{14}	Z_{15}
1	Z_{21}	Z_{22}	Z_{23}	Z_{24}	Z_{25}
2	Z_{31}	Z_{32}	Z_{33}	Z_{34}	Z_{35}
3	Z_{41}	Z_{42}	Z_{43}	Z_{44}	Z_{45}
4	Z_{51}	Z_{52}	Z_{53}	Z_{54}	Z_{55}

Table 4: Confusion matrix for each end every class label (i)

Actual	Predicted	
	positive	negative
positive	tp_i	fn_i
neagative	fp_i	tn_i

$$\frac{\sum_{i=1}^{l}\frac{tp_i+tn_i}{tp_i+fn_i+fp_i+tn_i}}{l}$$ Average accuracy Eq. (9)

$$\frac{\sum_{i=1}^{l}\frac{fp_i+fn_i}{tp_i+fn_i+fp_i+tn_i}}{l}$$ Error rate Eq. (10)

$$\frac{\sum_{i=1}^{l}\frac{tp_i}{tp_i+fp_i}}{l}$$ Precision (Macro) Eq. (11)

$$\frac{\sum_{i=1}^{l}\frac{tp_i}{tp_i+fn_i}}{l}$$ Recall (Macro) Eq. (12)

$$\frac{(\beta^2+1)Precision_{macro}Recall_{macro})}{\beta^2.Precision_{macro}+Recall_{macro})}$$ F-score (Macro) Eq. (13)

$$\frac{\sum_{i=1}^{l}tp_i}{\sum_{i=1}^{l}tp_i+fp_i}$$ Precision (Micro) Eq. (14)

$$\frac{\sum_{i=1}^{l}tp_i}{\sum_{i=1}^{l}tp_i+fn_i}$$ Recall (Micro) Eq. (15)

$$\frac{(\beta^2+1)Precision_{micro}Recall_{micro})}{\beta^2.Precision_{micro}+Recall_{micro})}$$ F-score(Micro) Eq. (16)

4.2 Results for Setup-1

In set up- I, we considered Cleveland heart diseases data found in UCI machine learning repository which consists of 303 patients and 13 features (age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal) with 1 class labels (num). The class label has 5 distinct entries such as 0 to 4, here 0 represent the number of healthy patients and 1 to 4 represents the level of the disease affected by a patient. Out of 303 patients 164 patients are healthy patients and remaining 139 patients are affected by different level of heart disease like 1, 2, 3 and 4. Here 55 patients are comes under the level 1, 36 patients are comes under the level 2, 35 patients are comes under the level 3 and 13 students are comes under the level 4. All 13 features are in discrete and continuous form [19]. In this data set some of the attributes has less no of missing values these missing values are replaced by mean of that attribute. After that we divided the data set into training and test dataset based on 50% data distribution. The training data consist of 153 patients which are distributed as 82 patients under 0, 28 patients under 1, 18 patients under 2, and 18 patients under

3 and 7 patients under 4. The test data consist of 150 patients which are distributed as 82 patients under 0, 27 patients under 1, 18 patients under 2, and 17 patients under 3 and 6 patients under 4. Table 5 represents the performance of the chosen classification algorithms (DT, RF, SVM, KNN, and NB) with respect to average accuracy, error rate, precision, recall, F-score for micro and macro averaging. Among all the classifier SVM classifier perform better than the other classifier in terms of average accuracy is shown in Fig 4.

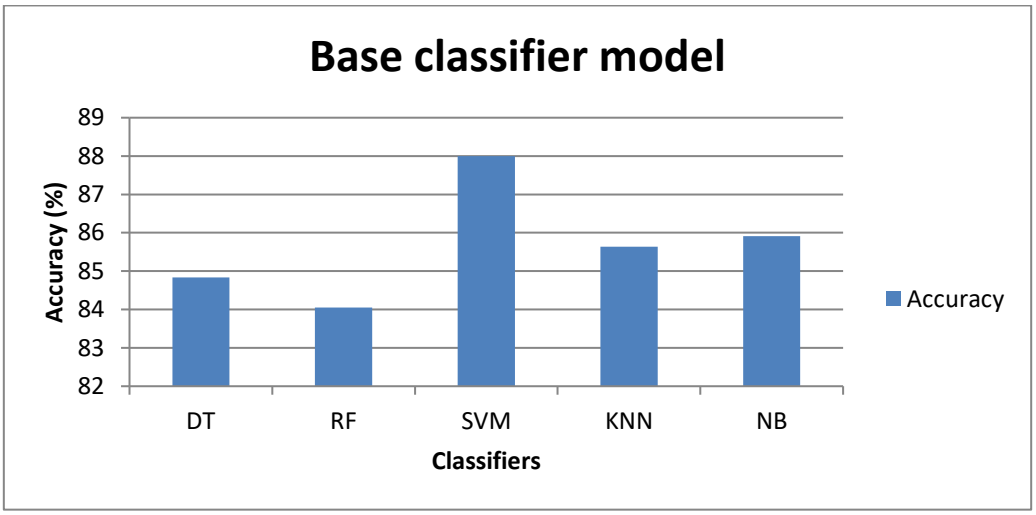

Fig 4: Accuracy of multi classification techniques

Table 5: Performance of multi classification techniques

Classifier	DT	RF	SVM	KNN	NB
Average accuracy	84.84	84.05	88.00	85.64	85.91
Error rate	15.16	15.95	12.00	14.36	14.09
Precision ($_{Macro}$)	44.31	40.35	52.66	46.92	46.11
Recall ($_{Macro}$)	45.27	41.74	52.91	47.49	44.65
F-score ($_{Macro}$)	44.78	41.03	52.79	47.21	45.37
Precision ($_{Micro}$)	62.75	60.78	70.00	64.71	65.36
Recall ($_{Micro}$)	62.75	60.78	70.00	64.71	65.36
F-score ($_{Micro}$)	62.75	60.78	70.00	64.71	65.36

4.3 Results for Setup-2

In set up II, we considered Cleveland heart diseases data found in UCI machine learning repository which consists of 303 patients and 13 features (age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal) with 1 class labels (num). The class label has 5 distinct entries such as 0 to 4, here 0 represent the number of healthy patients and 1 to 4 represents the level of the disease affected by a patient. Out of 303 patients 164 patients are healthy patients and remaining 139 patients are affected by different level of heart disease like 1, 2, 3 and 4. Here 55 patients are comes under the level 1, 36 patients are comes under the level 2, 35 patients are comes under the level 3 and 13 students are comes under the level 4. All

13 features are in discrete and continuous form. In this data set some of the attributes has less no of missing values these missing values are replaced by mean of that attribute.

Filter based feature selection process: After the preprocessing the data matrix (303 X 14 including class label) is given to filter based features selection techniques such as Chi-square, FCBF, GI, RelifeF. The filter based feature selection method is based on applying some of the statistical operation to each and every feature which is correlated to outcome of the dataset and the best features set are generated based on maximum score. R language F-Selector package used for identification of top 5 features (It is varied depend on chosen feature selection techniques). So our data matrix size is remapped based on top 5 features, 303 samples with 1 class label and it is given to chosen classification algorithms (DT, RF, SVM, KNN, and NB). Table 6-9 represents the performance of the each and every chosen classification algorithms (DT, RF, SVM, KNN, and NB) mapped with the chosen filter based feature selection techniques and compared the efficient performance measures like average accuracy, error rate, precision, recall, F-score for micro and macro averaging. Among all the combined model RelifeF+ SVM classifier perform better than the other combined model classifiers in terms of average accuracy is shown in Fig 5.

Table 6: Performance of Chi-square with multi classification techniques

Classifier	DT	RF	SVM	KNN	NB
Average accuracy	91.82	92.28	94.47	92.81	87.25
Error rate	8.18	7.72	5.53	7.19	12.75
Precision ($_{Macro}$)	72.04	74.50	77.44	76.37	52.57
Recall ($_{Macro}$)	80.09	81.01	82.50	83.36	63.59
F-score ($_{Macro}$)	75.85	77.62	79.89	79.71	57.56
Precision ($_{Micro}$)	79.87	80.92	86.36	82.24	68.63
Recall ($_{Micro}$)	79.87	80.92	86.36	82.24	68.63
F-score ($_{Micro}$)	79.87	80.92	86.36	82.24	68.63

Table 7: Performance of FCBF with multi classification techniques

Classifier	DT	RF	SVM	KNN	NB
Average accuracy	93.89	93.36	94.15	92.82	88.58
Error rate	6.11	6.64	5.85	7.18	11.42
Precision ($_{Macro}$)	81.72	77.67	81.98	75.70	56.04
Recall ($_{Macro}$)	86.35	79.68	87.06	79.19	65.81
F-score ($_{Macro}$)	83.97	78.66	84.45	77.41	60.53
Precision ($_{Micro}$)	84.97	83.66	85.53	82.35	71.90
Recall ($_{Micro}$)	84.97	83.66	85.53	82.35	71.90
F-score ($_{Micro}$)	84.97	83.66	85.53	82.35	71.90

Table 8: Performance of GI with multi classification techniques

Classifier	DT	RF	SVM	KNN	NB
Average accuracy	89.64	90.71	92.83	93.93	88.05
Error rate	10.36	9.29	7.17	6.07	11.95
Precision (Macro)	57.26	60.69	73.04	76.66	54.65
Recall (Macro)	66.78	69.75	79.16	80.14	64.33
F-score (Macro)	61.65	64.90	75.97	78.36	59.09
Precision (Micro)	74.51	77.12	82.35	85.06	70.59
Recall (Micro)	74.51	77.12	82.35	85.06	70.59
F-score (Micro)	74.51	77.12	82.35	85.06	70.59

Table 9: Performance of RelifeF multi classification techniques

Classifier	DT	RF	SVM	KNN	NB
Average accuracy	92.00	93.33	**95.20**	91.47	90.67
Error rate	8.00	6.67	4.80	8.53	9.33
Precision (Macro)	67.63	71.50	76.73	65.24	57.05
Recall (Macro)	72.67	76.38	78.08	70.32	60.32
F-score (Macro)	70.06	73.86	77.40	67.68	58.64
Precision (Micro)	80.00	83.33	88.00	78.67	76.67
Recall (Micro)	80.00	83.33	88.00	78.67	76.67
F-score (Micro)	80.00	83.33	88.00	78.67	76.67

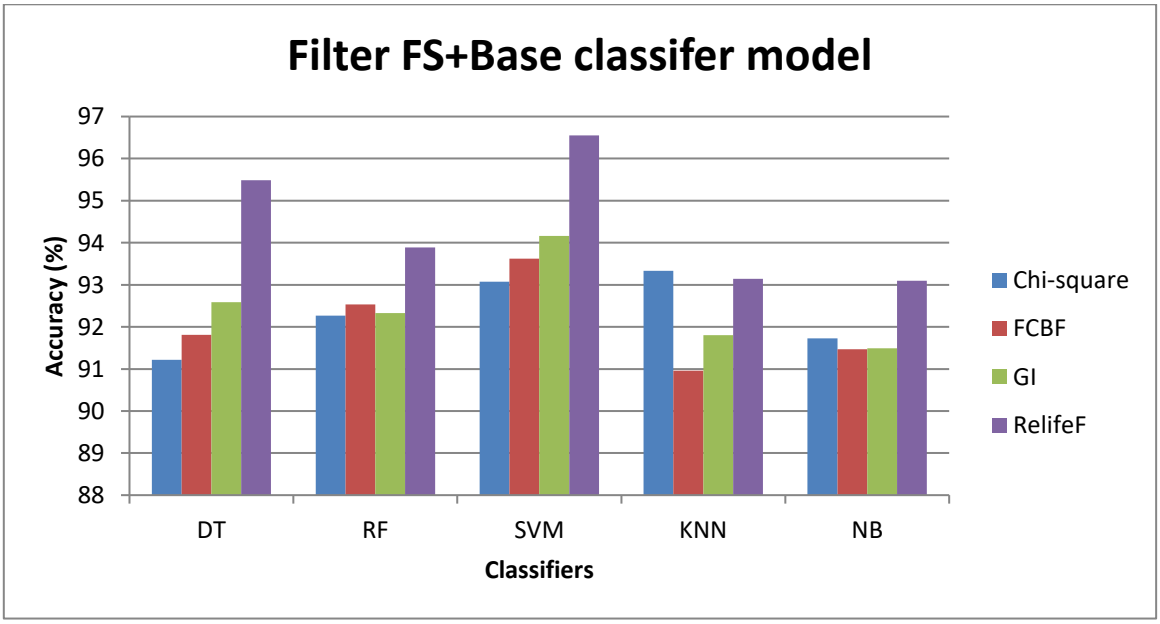

Fig 5: Accuracy of filter feature selection techniques with multi classification techniques

Wrapper based feature selection process: After the preprocessing the data matrix (303 X 14 including class label) is given to wrapper based features selection techniques such as Backward Feature Elimination (BFE), Exhaustive Feature Selection (EFS), Forward Feature Selection (FFS), and Recursive Feature Elimination (RFE). The wrapper based feature selection is totally opposite to filter based techniques, here the subset of features are randomly chosen and given for module building. Based on the model outcome in the next iteration the process will add some more features. R language F-Selector package used for identification of top feature sets (It is varied depend on chosen feature selection techniques). So our data matrix size is remapped based on top features, 303 samples with 1 class label and it is given to chosen classification algorithms (DT, RF, SVM, KNN, and NB). Table 10-13 represents the performance of the each and every chosen classification algorithms (DT, RF, SVM, KNN, and NB) mapped with the chosen wrapper based feature selection techniques and compared the efficient performance measures like average accuracy, error rate, precision, recall, F-score for micro and macro averaging. Among all the combined model RFS+ SVM classifier perform better than the other combined model classifiers in terms of average accuracy is shown in Fig 6.

Table 10: Performance of BFE with multi classification techniques

Classifier	DT	RF	SVM	KNN	NB
Average accuracy	91.22	92.27	93.07	93.33	91.73
Error rate	8.78	7.73	6.93	6.67	8.27
Precision ($_{Macro}$)	68.00	70.34	72.72	73.07	69.23
Recall ($_{Macro}$)	78.70	80.85	81.58	82.32	80.36
F-score ($_{Macro}$)	72.96	75.23	76.90	77.42	74.38
Precision ($_{Micro}$)	78.29	80.67	82.67	83.33	79.33
Recall ($_{Micro}$)	78.29	80.67	82.67	83.33	79.33
F-score ($_{Micro}$)	78.29	80.67	82.67	83.33	79.33

Table 11: Performance of EFS with multi classification techniques

Classifier	DT	RF	SVM	KNN	NB
Average accuracy	91.81	92.53	93.62	90.96	91.47
Error rate	8.19	7.47	6.38	9.04	8.53
Precision ($_{Macro}$)	71.32	70.96	74.37	66.27	67.66
Recall ($_{Macro}$)	78.57	81.10	77.91	75.47	77.14
F-score ($_{Macro}$)	74.77	75.69	76.10	70.57	72.09
Precision ($_{Micro}$)	80.00	81.33	84.21	77.63	78.67
Recall ($_{Micro}$)	80.00	81.33	84.21	77.63	78.67
F-score ($_{Micro}$)	80.00	81.33	84.21	77.63	78.67

Table 12: Performance of FFS with multi classification techniques

Classifier	DT	RF	SVM	KNN	NB
Average accuracy	92.59	92.33	94.16	91.80	91.49
Error rate	7.41	7.67	5.84	8.20	8.51
Precision ($_{Macro}$)	74.42	73.42	76.98	71.87	69.30
Recall ($_{Macro}$)	81.22	80.05	85.43	78.57	82.03
F-score ($_{Macro}$)	77.67	76.59	80.99	75.07	75.13
Precision ($_{Micro}$)	81.94	81.29	85.62	80.00	78.95
Recall ($_{Micro}$)	81.94	81.29	85.62	80.00	78.95
F-score ($_{Micro}$)	81.94	81.29	85.62	80.00	78.95

Table 13: Performance of RFE with multi classification techniques

Classifier	DT	RF	SVM	KNN	NB
Average accuracy	95.49	93.89	**96.55**	93.14	93.10
Error rate	4.51	6.11	3.45	6.86	6.90
Precision ($_{Macro}$)	80.84	75.52	85.29	73.84	75.59
Recall ($_{Macro}$)	81.63	84.23	87.62	82.27	82.47
F-score ($_{Macro}$)	81.23	79.64	86.44	77.83	78.88
Precision ($_{Micro}$)	88.89	84.87	91.50	83.23	83.01
Recall ($_{Micro}$)	88.89	84.87	91.50	83.23	83.01
F-score ($_{Micro}$)	88.89	84.87	91.50	83.23	83.01

Table 14 represent the overall performance of the combined classifier with respect to the filter based techniques (Chi-square, Fast Correlation Based Filter (FCBF), Gini Index (GI), RelifeF) and wrapper based techniques (Backward Feature Elimination (BFE), Exhaustive Feature Selection (EFS), Forward Feature Selection (FFS), and Recursive Feature Elimination (RFE)) based on average accuracy, error rate, precision, recall, F-score for micro and macro averaging. Table shows that feature selection plays a vital role to increases the prediction accuracy.

Table 14: Over all accuracy comparison of combined classifiers

Wrapper based FS	**REF**	95.49	93.89	**96.55**	93.14	93.10	4.51	6.11	3.45	6.86	6.90
	FFS	92.59	92.33	94.16	91.80	91.49	7.41	7.67	5.84	8.20	8.51
	EFS	91.81	92.53	93.62	90.96	91.47	8.19	7.47	6.38	9.04	8.53

	BFE	91.22	92.27	93.07	93.33	91.73	8.78	7.73	6.93	6.67	8.27
Filter based FS	**ReliefF**	92.00	93.33	**95.20**	91.47	90.67	8.00	6.67	4.80	8.53	9.33
	GI	89.64	90.71	92.83	93.93	88.05	10.36	9.29	7.17	6.07	11.95
	FCBF	93.89	93.36	94.15	92.82	88.58	6.11	6.64	5.85	7.18	11.42
	Chi-Square	91.82	92.28	94.47	92.81	87.25	8.18	7.72	5.53	7.19	12.75
classifier	**Base classifier**	84.84	84.05	88.00	85.64	85.91	15.16	15.95	12.00	14.36	14.09
Performance measure		DT	RF	SVM	KNN	NB	DT	RF	SVM	KNN	NB
		Average Accuracy					**Error rate**				

And also the RelifeF+ SVM classifier model achieves the highest accuracy 95.2% than the other combined classifiers in filter based approaches and the RFS+ SVM classifier model achieves the highest accuracy 96.5% than the other combined model classifiers in wrapper based approaches.

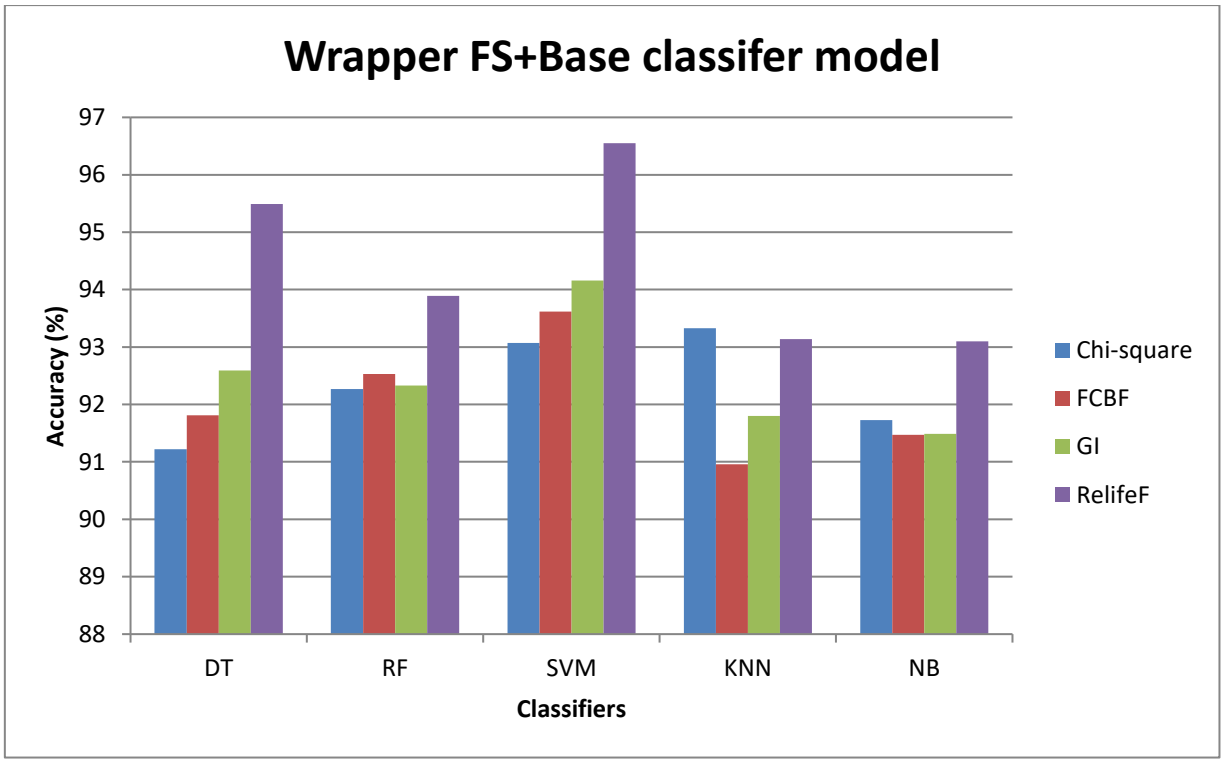

Fig 6: Accuracy of wrapper feature selection techniques with multi classification techniques

4.4 Results for Setup-3

Table 15 represent the overall performance of the combined classifier with respect to the filter based techniques (Chi-square, Fast Correlation Based Filter (FCBF), Gini Index (GI), RelifeF) and wrapper based techniques (Backward Feature Elimination (BFE), Exhaustive Feature Selection (EFS), Forward Feature Selection (FFS), and Recursive Feature Elimination (RFE)) based on average accuracy, error rate. In set up III, the processed data set on top 5 features, 303 samples with 1 class label from setup 2 is given to some of the existing models like HRLFM, RBFL+OFBAT, and RF+RS proposed by (Mohan et al.(2019)- [1], Reddy et al.(2017)-[17], Yekkala et.al (2018)-[18] and find the average accuracy, error rate, precision, recall, F-score for micro and macro averaging which is presented in Table 15. Mohan et al. (2019) suggested a hybrid approach (HRLFM) which collaborate the Linear Method (LM) and Random Forest (RF) for prediction of heart disease [1]. Reddy, G et al. (2017) proposed a novel method for heart disease prediction (Cleveland, Hungarian and Switzerland datasets) system that is a combination of rule-based fuzzy logic (RBFL) and oppositional firefly with BAT (OFBAT) which obtains the maximum accuracy of 78% [17].

Yekkala et.al (2018) proposed a novel heart disease prediction by using three different classifications: KNN, RF, NB and one common Rough Set (RS) feature selection. [18]. The table shows that the proposed systems like RelifeF+ SVM and RFS+ SVM yields a better accuracy compared other existing models.

.

Table 15: Accuracy comparison with existing models

Methods	Accuracy
RelifeF+ SVM (proposed)	95.20
RFS+ SVM (proposed)	96.55
HRLFM proposed by Mohan et al.(2019)- [1]	89.25

RBFL+OFBAT proposed by Reddy et al.(2017)-[17]	86.45
RF+RS proposed by Yekkala et.al (2018)-[18]	93.67

5. Conclusion

In this paper we are constructed a model for heart disease prediction using combined feature selection and classification machine learning techniques. According to the existing study the one of the main difficult in heart disease prediction system is that the available data in open sources are not properly recorded the necessary characteristics and also there is some lagging in finding the useful features from the available features. To overcome the flaws in the existing system a methodology is proposed in this paper that consists of two phases: Phase one employs two broad categories feature selection techniques to identify the efficient feature sets and it is given to input of our second phase such as classification. First, the considered data set (303 samples, 13 attribute and 1 class label) some of the attributes has less no of missing values these missing values are replaced by mean of that attribute. After preprocessing the preprocessed data set is given to on filter based method for feature selection such as Chi-square, Fast Correlation Based Filter (FCBF), Gini Index (GI), RelifeF and wrapper based method for feature selection such as Backward Feature Elimination (BFE), Exhaustive Feature Selection (EFS), Forward Feature Selection (FFS), and Recursive Feature Elimination (RFE) to indentify a best fitting features. The UCI heart disease data set is used to evaluate the output in this study. After that we divided the data set into training and test dataset based on 50% data distribution. Finally, the proposed system's performance is validated by various experiments setups. The experiment result shows that the model RelifeF+ SVM classifier achieves the highest accuracy 95.20% than the other combined model classifiers in filter based approaches and the model RFS+ SVM classifier achieves the highest accuracy 96.55% than the other combined model classifiers in wrapper based approaches. In future we are going to construct a heart disease prediction system with combination feature selection and ensemble classification techniques to increase the efficiency of classification performance.

References

1. Mohan, Senthilkumar, Chandrasegar Thirumalai, and Gautam Srivastava. "Effective heart disease prediction using hybrid machine learning techniques." *IEEE Access* 7 (2019): 81542-81554.
2. Jacob, Joseph, et al. "Predicting outcomes in rheumatoid arthritis related interstitial lung disease." *European Respiratory Journal* 53.1 (2019).
3. Chen, Austin H., et al. "HDPS: Heart disease prediction system." *2011 computing in cardiology*. IEEE, 2011.
4. Bhatla, Nidhi, and Kiran Jyoti. "An analysis of heart disease prediction using different data mining techniques." *International Journal of Engineering* 1.8 (2012): 1-4.
5. Palaniappan, Sellappan, and Rafiah Awang. "Intelligent heart disease prediction system using data mining techniques." *2008 IEEE/ACS international conference on computer systems and applications*. IEEE, 2008.
6. Kavitha, M., et al. "Heart Disease Prediction using Hybrid machine Learning Model." *2021 6th International Conference on Inventive Computation Technologies (ICICT)*. IEEE, 2021.
7. Turabieh, Hamza. "A hybrid ann-gwo algorithm for prediction of heart disease." *American Journal of Operations Research* 6.2 (2016): 136-146.

8. Uddin, Shahadat, et al. "Comparing different supervised machine learning algorithms for disease prediction." *BMC medical informatics and decision making* 19.1 (2019): 1-16.
9. Sliwoski, Gregory, et al. "Computational methods in drug discovery." *Pharmacological reviews* 66.1 (2014): 334-395.
10. De Heredia, Unai López, and José Luis Vázquez-Poletti. "RNA-seq analysis in forest tree species: bioinformatic problems and solutions." *Tree Genetics & Genomes* 12.2 (2016): 30.
11. Saez-Rodriguez, Julio, and Nils Blüthgen. "Personalized signaling models for personalized treatments." *Molecular systems biology* 16.1 (2020): e9042.
12. Gozes, Ophir, et al. "Rapid ai development cycle for the coronavirus (covid-19) pandemic: Initial results for automated detection & patient monitoring using deep learning ct image analysis." *arXiv preprint arXiv:2003.05037* (2020).
13. Monsi, Justin, et al. "XRAY AI: Lung Disease Prediction Using Machine Learning." *International Journal of Information* 8.2 (2019).
14. Patel, Jaymin, Dr TejalUpadhyay, and Samir Patel. "Heart disease prediction using machine learning and data mining technique." *Heart Disease* 7.1 (2015): 129-137.
15. Sathya priya "Chronic Kidney Disease Prediction Using Machine Learning." *International Journal of Computer Science and Information Security (IJCSIS)* 16.4 (2018).
16. Mathur, Richa, Vibhakar Pathak, and Devesh Bandil. "Parkinson disease prediction using machine learning algorithm." *Emerging Trends in Expert Applications and Security*. Springer, Singapore, 2019. 357-363.
17. Reddy, G. Thippa, and Neelu Khare. "An efficient system for heart disease prediction using hybrid OFBAT with rule-based fuzzy logic model." *Journal of Circuits, Systems and Computers* 26.04 (2017): 1750061.
18. Yekkala, Indu, and Sunanda Dixit. "Prediction of heart disease using random forest and rough set based feature selection." *International Journal of Big Data and Analytics in Healthcare (IJBDAH)* 3.1 (2018): 1-12.
19. https://archive.ics.uci.edu/ml/datasets/heart+disease
20. Bashir, Saba, et al. "Improving heart disease prediction using feature selection approaches." *2019 16th international bhurban conference on applied sciences and technology (IBCAST)*. IEEE, 2019.
21. Peter, T. John, and K. Somasundaram. "Study and development of novel feature selection framework for heart disease prediction." *International Journal of Scientific and Research Publications* 2.10 (2012): 1-7.
22. Usman, Ali Muhammad, Umi Kalsom Yusof, and Syibrah Naim. "Cuckoo inspired algorithms for feature selection in heart disease prediction." *International Journal of Advances in Intelligent Informatics* 4.2 (2018): 95-106.
23. Jin, Xin, et al. "Machine learning techniques and chi-square feature selection for cancer classification using SAGE gene expression profiles." *International Workshop on Data Mining for Biomedical Applications*. Springer, Berlin, Heidelberg, 2006.
24. Yu, Lei, and Huan Liu. "Feature selection for high-dimensional data: A fast correlation-based filter solution." *Proceedings of the 20th international conference on machine learning (ICML-03)*. 2003.
25. Reddy, N. Satish Chandra, et al. "Classification and feature selection approaches by machine learning techniques: Heart disease prediction." *International Journal of Innovative Computing* 9.1 (2019).

26. Spolaôr, Newton, et al. "ReliefF for multi-label feature selection." *2013 Brazilian Conference on Intelligent Systems*. IEEE, 2013.
27. Kostrzewa, Daniel, and Robert Brzeski. "The data dimensionality reduction in the classification process through greedy backward feature elimination." *International Conference on Man–Machine Interactions*. Springer, Cham, 2017.
28. Ren, Jiangtao, et al. "Forward semi-supervised feature selection." *Pacific-Asia conference on knowledge discovery and data mining*. Springer, Berlin, Heidelberg, 2008.
29. Lee, Chia-Yen, and Bo-Syun Chen. "Mutually-exclusive-and-collectively-exhaustive feature selection scheme." *Applied Soft Computing* 68 (2018): 961-971.
30. Yan, Ke, and David Zhang. "Feature selection and analysis on correlated gas sensor data with recursive feature elimination." *Sensors and Actuators B: Chemical* 212 (2015): 353-363.
31. Shah, D., Patel, S.B., & Bharti, S. (2020). Heart Disease Prediction using Machine Learning Techniques. *SN Comput. Sci., 1*, 345.