A Knowledge Based Framework for Cardiovascular Disease Prediction

Abha Marathe, Dr.Virendra Shete, Dr. Dhananjay Upasani
Department of Electronics and Telecommunication Engineering, MIT School of Engineering and Sciences, Pune, India

Abstract—**Cardiovascular disease has become more concern in the hectic and stressful life of modern era. Machine learning techniques are becoming reliable in medical treatment to help the doctors. But the ML algorithms are sensitive to data sets. Hence a Smart Robust Predictive System is almost essential which can work efficiently on all data sets. The study proposes ensembled classifier validating its performance on five different data sets-Cleveland, Hungarian, Long Beach, Statlog and Combined datasets. The developed model deals with missing values and outliers. Synthetic Minority Oversampling Technique (SMOTE) was used to resolve the class imbalance issue. In this study, performance of five individual classifiers – Support Vector Machine Radial (SVM-R), Logistic Regression (LR), Naïve Bayes (NB), Random Forest (RF) and XGBoost, was compared with five ensembled classifiers on five different data sets. On each data set the top three performers were identified and were combined to give ensemble classifiers. Thus, in all total 25 experimentation were done. The results have shown that out of all classifiers implemented, the proposed system outperforms on all the data sets. The performance was validated by 10-fold cross validation The proposed system gives the highest accuracy and sensitivity of 87% and 86% respectively.**

Keywords—Machine learning; ensemble classifier; cardiovascular disease; performance metrics; classifier techniques

I. Introduction

The urbanization of the population in the world resulted in the increase in the urban population from 37% in 1970 to the projected 61% in 2025[1]. One of the major impacts of such lifestyle is cardiovascular disease-CVD. According to WHO [2] 17.9 million people died in 2019 because of CVD which is nearly 32% of world-wide deaths. For predicting the CVD risk, different traditional risk calculators are used. They assign certain weights to the risk factors and calculate the risk scores. But these calculators have limitation that they are population specific as certain population was considered in the respective cohort study. Also, the risk factors considered are different for different calculators. The decisions given by these calculators differ on the same population [3], hence they are not consistent as well. Therefore, designing a robust system which is applicable for all type of population is the need of the hour. This can be done with the help of ever evolving and reliable sophisticated machine learning and deep learning approaches. Because of manual constraints, many modern age researchers moved towards these approaches. These algorithms are sensitive to data sets.The literature surveyed has revealed that for all different dataset different machine learning techniques were found to be different. The objective of the study is to propose a novel robust algorithm which works on diverse dataset efficiently. This proposed algorithm is a uniquely developed ensembled classifier of three individual classifiers Random Forest, Support Vector Machine Radial and XGBoost. The performance was validated on five different datasets to prove its consistency.

The paper is divided into six sections. Besides Introduction, Section II discusses about the work done till date, Section III focuses on the proposed system, Section IV contains different evaluation parameters used for classifier performance, Section V reveals the results and their discussions and the last Section VI is the conclusion and future scope of the study.

II. Related Work

In past few decades many Machine learning techniques have been used for prediction of heart disease. Many studies involved the comparison of traditional CVD risk calculators with different Machine learning algorithms. Many studies of heart disease prediction used cohort data set. The comparison of traditional CVD risk calculator models with different machine learning models[4] used nearly 30000 subjects from eastern China who were having high risk of CVD for 3-year risk assessment. Random forest was found to be the best with AUC of 0.787. Similarly, in the other study from Korea 4699 subjects were extracted from Korean National Health Insurance Service Health Screening Database. Out of all the 10 ML algorithms applied, XGBoost, Gradient Boost (GB) and RF with AUC nearly 0.81 performed even better than the existing risk models- Framingham and ACC/AHA American College of Cardiology /American Heart Association risk model [5]. Another cohort study based on same population compared the Pooled cohort equations, Framingham risk model and QRISK3 model with different machine learning algorithms. Neural network was found to have highest C-statistics of 0.751[6]. Another study from Athens, Greece which used 10 years of follow up compared the machine learning techniques with statistical approach of Hellenic Score. The Random Forest algorithm gave the best results [7]. One more study where electronically recorded data by UK National Health Service (NHS) was used. The performance was validated by Harrell’s c-statistic. The traditional ACC/AHA model was compared with different ML algorithms like Random Forest, Logistic Regression, Gradient Boosting and Neural Networks. AUC-c statistics was found to be best for Neural network with 0.764 value.[8]. Another cohort study from UK compared the Framingham model, Cox proportional Hazard model and ML algorithms like SVM, RF, NN, AdaBoost, Gradient boosting and Auto prognosis-advanced Bayesian optimization technique to predict the

CVD. The missing values were addressed by Miss Forest algorithm. Auto prognosis performed best out of all the techniques compared [9]. Cohort study for CVD prediction was carried out in Northern California with 32192 patients. Atherosclerotic CVD i.e. ACVD patients were also included in the study. The machine learning algorithms like RF, GBM, XGBoost and logistic regression were compared. XGBoost demonstrated highest AUC 0.70 (95% CI 0.68 to 0.71) in the full CVD cohort and AUC 0.71 (95% CI 0.69 to 0.73) in patients with ASCVD, with comparable performance by GBM, RF and Regression [10]. Apart from cohort studies many researchers used the data sets which are directly provided by the data providers like Cleveland, Hungarian, long Beach, Switzerland, Statlog etc. These data sets are available on UC Irvine Machine learning Repository. There are 76 attributes out of which most relevant 14 attributes are provided by the data providers. Machine learning algorithms are very much sensitive to data sets. To get high efficiency and reliability the data sets need to be properly formed. Before implementing any algorithm, data preprocessing is a must. Data Preprocessing includes steps of data cleaning like addressing the missing values, identifying the outliers, checking for duplicate records etc. In many real-life problems data imbalance is a major challenge in front of the researchers. Specifically, for a medical study like disease detection the data points with one class i.e., normal, and healthy person is more as compared to the patient suffering from a particular disease. This results in class imbalance. Hence before application of any algorithm balancing the data by addressing this data skew becomes a need. Such data preprocessing seen in different studies often leads in better results. In [11] Cleveland data set which is most widely used data set was used for prediction of heart disease. The authors generated artificial records in 5%, 10%, 20% and 50%. They proposed a data duplicate finder algorithm which removes the duplicates in the record. The decision tree C5.0 was used as a classifier which gives the better results for the data set without duplicates as compared to with duplicates. Like the duplicate records, the missing values and outliers are very crucial to handle. These missing values can be either removed with no information loss or can be imputed. In [12], the missing values were imputed by Mean whereas the outliers were identified by Boxplot and were removed. The class imbalance problem was solved by SMOTE technique. Such imbalance in the data set of Framingham data set was balanced by Random Oversampling examples in [13]. The AUC was used as a performance metrics. It reported the maximum achieved AUC by SVM of 0.75.Like the imbalance data set where the number of instances is required to be balanced, the number and the relevance of the attributes is also very important in any predictive system. Addition of irrelevant features in the dataset may misguide the model, hence identification of important attributes and their inclusion in the model is very important. Ample studies are done where different techniques of feature selections and their role in improving the performance of the model are discussed. In [14] different classifiers like Linear Discriminant Analysis-LDA, Decision tree (DT), SVM, GB and RF were used. For feature selection sequential feature selection (SFS) was used. Use of SFS reduces the number of features and hence optimizes computation time. It was found that for Hungary, Switzerland & Long Beach V and Heart Statlog Cleveland Hungary Datasets, Random Forest Classifier SFS and Decision Tree Classifier SFS achieved the highest accuracy ratings of 100%, 99.40% and 100%, 99.76% respectively. There are other feature selection methods like Fast Correlation-Based Filter Solution (FCBF), minimal redundancy maximal relevance (mRMR), Least Absolute Shrinkage and Selection operator (LASSO), and Relief which were used in [15]. It has used 10 ML algorithms and indicated the best algorithm for feature selection method. Extra tree (ET) classifier was found to be superior amongst all. Accuracy of top performer ET and GB found to be 92.09% and 91.34% when all attributes were considered. With relief feature selection algorithm, the accuracy of ET increased from 92.09% to 94.41% whereas for GB the increase was from 91.34% to 93.36% when FCBF feature selection was applied. Another study used the feature selection methods like Relief Feature selection technique and least absolute shrinkage and selection operator algorithm (LASSO) [16]. The data set contained 13 attributes out of which Random Forest bagging method (RFBM) identified most relevant 10 features and accuracy achieved with this was 99.05%. Addition to these traditional feature selection methods, [17] proposed fast conditional mutual information (FCMIM) technique which is based on selection of features on the basis of features mutual information. The combination of SVM-FCMIM gave the highest accuracy of 92.37%. In [18] the heart disease prediction was carried out by dimension reduction method. PCA and Chi-square analysis with Random Forest shown the best performer with 98.7% accuracy. When weak performers are combined, a strong predictive system can be generated. These are called as Ensembled classifiers. Researchers have experimented with different ensemble classifiers and comparison made with individual classifiers. Studies like [19] weighted majority voting ensemble was used. The weights assigned to the individual classifiers' votes were decided as per their AUC values. The results observed were that this ensembled classifier performed best with AUC value of 83.9 for with laboratory parameters and 83.1 without laboratory parameters. There are three different types of ensemble techniques, Bagging, Boosting and Stacking. In [20] all these techniques along with majority voting were used. With bagging technique, the accuracy improved by 6.92%, with boosting this improvement was found to be by 5.94%, with stacking it improved by 6.93% whereas the highest improvement was observed by majority voting which was 7.26%.

III. Proposed System

This study proposes five different ensembled classifiers- E_1, E_2, E_3, E_4 and E_5. The composition of these ensembled classifiers is detailed later. The entire work done was divided into four main phases:

- Phase 1: Five different individual classifiers were trained, tested and validated by five different data sets.
- Phase 2: These classifiers were used to construct the five proposed ensemble classifiers.

- Phase 3: These ensembled classifiers were trained, tested and validated by all data sets.
- Phase 4: The individual and the ensembles were compared with different performance metrics and conclusions were drawn.

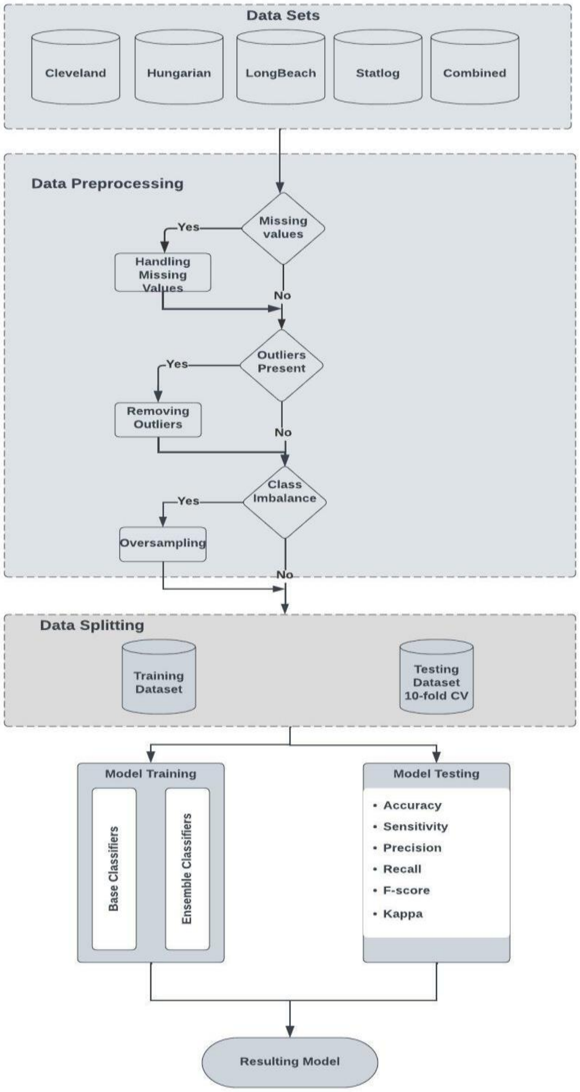

Fig. 1. Proposed system.

The entire system is depicted in Fig. 1. R programming language with R-studio as its IDE is used for this study.

A. Data Set Description

Five different datasets were used for this study-Cleveland, Hungarian, Long Beach, Statlog and Combined datasets taken from UCI Machine Learning repository [21,22]. The Cleveland data set is most used data set by all researchers working on heart diseases. In actual it consists of total 76 attributes, but out of them only 14 are more accurate and are widely used and given by data set provider. The number of instances is 303 with 13 predictors and one response variable. Here the target variable is represented as "num", where 1 stand for presence of heart disease whereas 0 represents absence of heart disease. Similarly, the other data sets are Hungarian dataset with 294 instances, Long Beach Dataset with 200 instances, Statlog Data set contains 270 instances with same attributes and combination of all datasets with 797 instances.

B. Data Preprocessing

Data cleaning is an important task in any machine learning problem. The quality of data becomes more crucial in medicine area [23]. Data cleaning in this study was done by finding missing values, finding outliers and checking and class imbalance problem. The missing values were removed without loss of information. In second step the outliers were checked and were removed from the data set. Third step was for checking the class imbalance in data set. It happens if there are a greater number of samples belonging to one class as compared to other class which may result into biased classifier. Many methods are presented by different researchers to tackle this problem [24]. In this paper, this issue was addressed by Synthetic Minority Oversampling Technique, i.e., SMOTE.

C. Synthetic Minority Oversampling Technique-SMOTE.

In this method minority class is oversampled by creating Synthetic examples unlike oversampling which generates the duplicate data points [25]. It uses the KNN algorithm to generate these synthetic points. A random sample from minority class is selected. Then one sample from K neighbor is selected and the distance vector between this selected point and current data point is calculated and further multiplied by any random number between 0 to 1.This is then added to current point and synthetic data point is created. Class imbalance was found in Hungarian and Long Beach Data sets. SMOTE technique was used to solve this issue in these two data sets.

D. Proposed Ensemble Classifiers

The study proposes five different ensembled classifiers as discussed below:

E_1: It was designed by combining all the base classifiers. Majority voting scheme was used for the prediction. All the five classifiers were considered for voting with same importance. The class label w was predicted by the decision given by each classifier C_i for every feature vector x as given in (1). Mode indicates majority as per usual statistical meaning.

$$w = mode\{decisions\ (C_1(x)\ , C_2(x)\ , \dots .\ C_m(x))\} \tag{1}$$

E_2: This ensemble classifier was based on the baseline accuracy of individual classifier. The classifier with highest accuracy was assigned with more weight. This assigned value of weight was then used as a multiplier for the prediction probability of the respective classifier. For a given feature vector x, depending on the probability of the label class the final decision was taken. The weight of the individual classifier was calculated as given in (2).

$$W_i = \frac{A_i}{\sum_i^m A} \tag{2}$$

TABLE I. TOP THREE PERFORMERS

	Classifiers				
Data Sets	**RF**	**NB**	**SVM-R**	**LR**	**XgBoost**
Cleveland	✓		✓		✓
Hungarian	✓	✓			✓
Long Beach	✓		✓		✓
Statlog		✓	✓		✓
Combined	✓		✓		✓

Where

A_i: Accuracy of individual Classifier

m: Number of classifiers used

W_i: Weight of individual Classifier

The final prediction probability for each class label is calculated as in (3).

$$p = \sum_i^m (W_i * A_i) \quad (3)$$

The class label having highest probability for a given feature vector was assigned to that vector. All the classifiers are considered in this voting scheme.

Total datasets considered for this study are five. All five individual classifiers were applied to all five datasets. Thus total 25 individual experimentations were done. Then for each data set top three performing classifiers based on their accuracies were identified. The top three classifiers for individual datasets are given in Table I.

Thus, three following different unique combinations of classifiers emerge which were considered for further experimentation.

E_3: It was the first combination RF+SVM-R+XgBoost. These three classifiers were used for majority voting. Label class w was assigned for a given feature vector given in (4).

$$w = mode\{Decisions(RF, SVM - R, XgBoost\} \quad (4)$$

e.g., let for any feature vector x, the decisions are as- RF: Class 0, SVM-R: Class 0, XgBoost: Class1. Then the final decision was taken as in (5).

$$w = mode\{0,0,1\} \;=>\; w = 0: Class\ 0 \quad (5)$$

E_4: This was the second unique combination of RF+NB+XgBoost. These three classifiers were used for majority voting. Label class w was assigned for a given feature vector as shown in (6).

$$w = mode\{Decisions(RF, NB, XgBoost\} \quad (6)$$

E_5**:** This was the third unique combination of RF+SVM+XgBoost. These three classifiers were used for majority voting. Label class w was assigned for a given feature vector as in (7).

$$w = mode\{Decisions(NB, SVM, XgBoost\} \quad (7)$$

The different ensembles are depicted in Fig. 2.

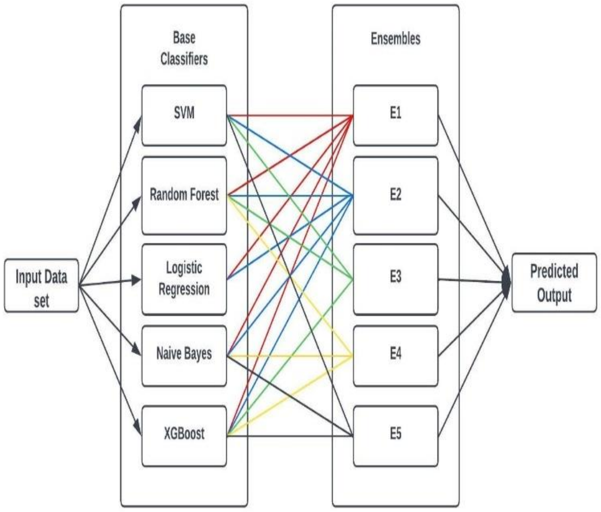

Fig. 2. Construction of ensemble classifier.

IV. EVALUATION PARAMETERS

The evaluation of performance of all the classifiers i.e., individual and ensemble were done by creating an error matrix or Confusion matrix specified in Table II. It shows four different notations True Positive TP these are the patients who are suffering from CVD and the algorithm also predicts the same. The number FN is False negative which shows that these many patients are suffering from disease but they are predicted as they are not suffering. This number needs to be less and costs more in medical studies. False positive FP indicates the number of patients not suffering from disease but identified as they are suffering. True negative refers to the true classification of normal patients. The performance of any classifier is characterized by values of FP and FN. These metrics are discussed below:

TABLE II. CONFUSION MATRIX

		Predicted	
Actual		**Positive**	**Negative**
	Positive	TP	FN
	Negative	FP	TN

Accuracy: It is the ratio of total true predictions to total number. It is given by (8)

$$Accuracy = \frac{(TP+TN)}{(TP+FP+FN+TN)} \quad (8)$$

Sensitivity: It is also called as true positive rate. It is the ratio of positive number classified correctly to the total positive instances. It is given by (9).

$$Sensitivity = \frac{(TP)}{(TP+FN)} \tag{9}$$

Specificity: It is defined as true negative rate. It is the ratio of measures of negative number classified correctly to the total negative instances. It is given by (10).

$$Specificity = \frac{(TN)}{(TN+FP)} \tag{10}$$

Precision: It is also known as positive predictive value. It is the ration of proportion of positive number classified correctly to total predicted positive. It is given by (11).

$$Precision = \frac{(TP)}{(TP+FP)} \tag{11}$$

F-measure: It is a measure of model performance that combines precision and recall into single number. It is given by (12).

$$Fmeasure = \frac{(2*Precsion*Recall)}{(Recall+Precsion)} \tag{12}$$

Kappa**:** It is the measure which accounts for correct classification due to chance

If K >0: Classification is better than chance classification.

K <0: Classification is not better than chance Classification.

K=1: Perfect Classification.

K= 0: Pure chance classification.

V. Results and Discussions

An elaborated discussion of the different results obtained for individual and ensemble classifiers is given below. The various performance metrics are compared and conclusions are drawn. The performance validation of all the models was done by 10-fold cross validation.

A. *With Individual Models*

The five different individual classifiers- RF, SVM, NB, LR and XGBoost were first applied on all the five data sets. For each data set the top performers based on different performance parameters were identified. On Cleveland data set, XGBoost was observed to be the best performer in terms of accuracy with 87.78% value, followed by RF and SVM-R with 84.44% and 83% of accuracy. Same is true for other performance metric like Specificity and Precision. Though the sensitivity of LR is highest amongst all, but it lags in the other parameters. Hence for Cleveland data set, the top three performers are considered as XGBoost, RF and SVM-R. The performance is given in Table III. For Hungarian data set, XGBoost, RF and NB comes out to be the top 3 performers with 91.26%, 90.29% and 86.41% of accuracy. These models also head in the important parameter i.e., Sensitivity with values 94.5%, 89.47% and 90.20%, Table IV shows the comparison. For the third data set which is Long Beach dataset, SVM Radial performs best with accuracy 88.64%. It also leads in the other parameters like Sensitivity, Specificity and Precision as well. The second topper is XGBoost and RF. Therefore, the top three identified classifiers for Long Beach Data Set are XGBoost, RF and SVM-RBF, shown in Table V. The next data set was Statlog data set. The findings depict that, XGBoost, SVM-R, NB and are top three classifiers. Their accuracy values are 86.42%,86.42% and 85.19% respectively. The parameters are compared in Table VI.

The last data set considered was Combined data set. It was found that RF performed best from all models with highest accuracy of 89.06%. The next followers were observed as XGBoost and SVM-R. These three classifiers were toppers in the performance parameters like Sensitivity, Specificity and Precision as well apart from accuracy. This is shown in Table VII. Thus, for Cleveland, Hungarian and Statlog dataset it is XGBoost which is best performer in terms of accuracy whereas for Long Beach it is SVM-R and for Combined data set it is Random Forest. The performance parameters averaged on all dataset for all individual classifier is shown in Table VIII.

TABLE III. Performance Comparison on Cleveland Data Set

Techniques	Accuracy	Sensitivity	Specificity	Precision	F Measure	Kappa
RF	84.44	82.05	86.27	82.05	82.05	0.68
NB	82.22	79.49	84.31	79.49	79.49	0.64
SVM-RBF	83.33	80.0	86.0	82.05	81.01	0.66
LR	82.22	84.85	80.70	71.79	77.78	0.63
Xgboost	87.78	82.05	92.16	88.89	85.33	0.75

TABLE IV. Performance Comparison on Hungarian Data Set

Techniques	Accuracy	Sensitivity	Specificity	Precision	F Measure	Kappa
RF	90.29	89.47	91.3	92.73	91.07	0.81
NB	86.41	90.20	82.69	83.64	86.80	0.73
SVM-RBF	79.61	79.31	80.0	83.64	81.42	0.59
LR	81.55	80.0	83.72	87.27	83.48	0.63
Xgboost	91.26	94.5	87.5	89.66	92.02	0.82

TABLE V. PERFORMANCE COMPARISON ON LONG BEACH DATA SET

Techniques	Accuracy	Sensitivity	Specificity	Precision	F Measure	Kappa
RF	81.82	85.0	79.17	77.27	80.95	0.64
NB	79.55	84.21	76.0	72.73	78.05	0.59
SVM-RBF	88.64	90.48	86.96	86.36	88.37	0.77
LR	79.55	88.24	74.07	68.18	76.92	0.59
Xgboost	84.09	81.82	86.36	85.71	83.72	0.68

TABLE VI. PERFORMANCE COMPARISON ON STATLOG DATA SET

Techniques	Accuracy	Sensitivity	Specificity	Precision	F Measure	Kappa
RF	83.95	81.08	86.36	83.33	82.19	0.67
NB	85.19	78.57	92.31	91.67	84.62	0.70
SVM-RBF	86.42	82.05	90.48	88.89	85.33	0.72
LR	82.72	78.95	86.05	83.33	81.08	0.65
Xgboost	86.42	86.11	86.67	83.78	84.93	0.73

TABLE VII. PERFORMANCE COMPARISON ON COMBINED DATA SET

Techniques	Accuracy	Sensitivity	Specificity	Precision	F Measure	Kappa
RF	89.06	85.81	92.12	91.10	88.38	0.78
NB	82.19	81.12	83.05	79.45	80.28	0.64
SVM-RBF	85.31	83.67	86.71	84.25	83.96	0.70
LR	81.25	80.28	82.02	78.08	79.16	0.62
Xgboost	84.06	82.19	85.63	82.76	82.47	0.68

TABLE VIII. AVERAGE PERFORMANCE COMPARISON ON ALL DATA SETS OF INDIVIDUALS

Techniques	Accuracy	Sensitivity	Specificity	Precision	F Measure	Kappa
RF	85.91	84.68	87.04	85.29	84.92	0.72
NB	83.11	82.71	83.67	81.39	81.84	0.66
SVM-RBF	84.66	83.10	86.03	85.03	84.01	0.69
LR	81.45	82.46	81.31	77.73	79.68	0.62
Xgboost	86.72	85.33	87.66	86.16	85.69	0.73

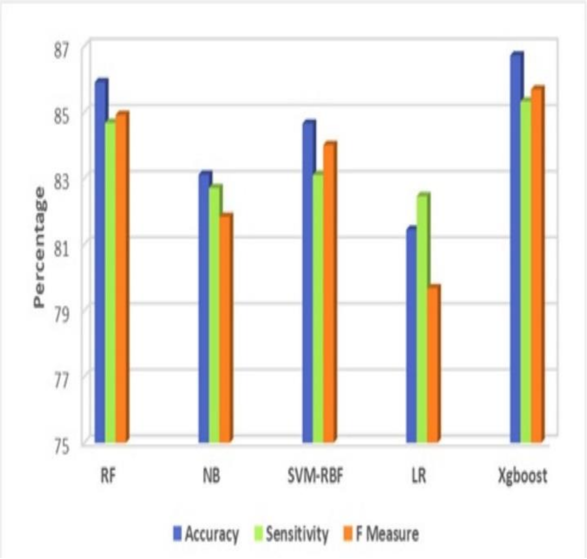

Fig. 3. Performance comparison of individual classifier.

Fig. 3 shows the performance of individual classifiers. The parameters taken for graphical representations are Accuracy, Sensitivity and F measure.

B. With Ensemble Classifiers

After applying individual models for all data sets, ensemble classifiers were designed based on Majority Voting and Weighted Voting. Firstly, all individual classifiers were considered for voting with same importance which forms Ensemble E1. Then the weighted voting was considered with all models with assigned weight as per the accuracy. This gave second ensemble E2. The next three ensemble were E3, E4 and E5, emerged as the top three performers for all five data sets. It was observed that out of all ensembles, E3 performs best with highest accuracy of 88.89% accuracy and 87.18% sensitivity on Cleveland data set and with 90.29% Accuracy and89.47% Sensitivity for Hungarian Data set, refer Table IX and Table X respectively. Similarly, for Statlog data set as well, E3 is best performer with Accuracy 86.42% and Sensitivity 83.78%, as given in Table XI.

For Long Beach data set, E5 leads with 88.64% accuracy and sensitivity 90.48%. shown in Table XII. For Combined data set it is E4 which has highest accuracy of 86.88% and sensitivity of 85.62% amongst all ensembles shown in Table XIII. The values of different performance parameters of all ensembles averaged on all data sets are shown in Table XIV. Out of all ensembled classifier, the proposed ensemble classifier i.e., E3 is observed to have highest accuracy, sensitivity, F measure and Kappa values the best ensemble classifier. The graphical information in Fig 4, shows that E3 has highest Accuracy, Sensitivity and F-measure amongst all Ensembles. When all individual and all ensemble classifiers were compared on their average values of all data sets, proposed ensemble E3 was found to be the best amongst all with highest accuracy, sensitivity, F- measure, and high Kappa values as given in Fig 5.

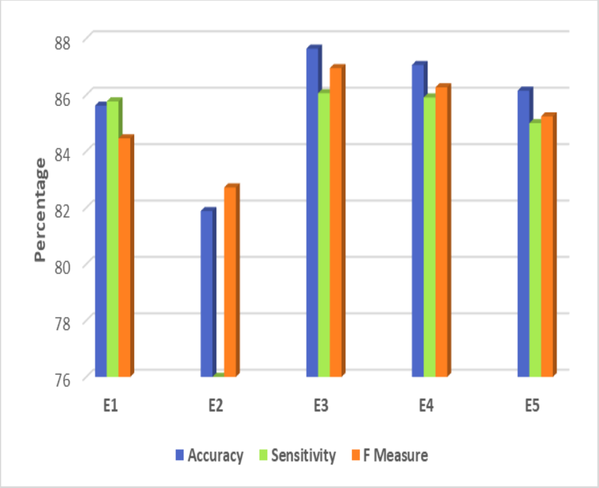

Fig. 4. Performance comparison of ensembles.

TABLE IX. PERFORMANCE COMPARISON OF ENSEMBLES ON CLEVELAND DATA SET

Techniques	Accuracy	Sensitivity	Specificity	Precision	F Measure	Kappa
E1	85.56	84.21	86.54	82.05	83.12	0.71
E2	81.11	72.92	90.48	89.74	80.46	0.63
E3	88.89	87.18	90.20	87.18	87.18	0.77
E4	88.89	87.18	90.20	87.18	87.18	0.77
E5	83.33	80.0	86.0	82.05	81.01	0.66

TABLE X. PERFORMANCE COMPARISON OF ENSEMBLES ON HUNGARIAN DATASET

Techniques	Accuracy	Sensitivity	Specificity	Precision	F Measure	Kappa
E1	87.38	87.50	87.23	89.09	88.29	0.76
E2	82.52	76.81	94.12	96.36	85.48	0.64
E3	90.29	89.47	91.30	92.73	91.07	0.80
E4	90.29	89.47	91.30	92.73	91.07	0.80
E5	87.38	88.89	85.71	87.27	88.07	0.78

TABLE XI. PERFORMANCE COMPARISON OF ENSEMBLES ON STATLOG DATASET

Techniques	Accuracy	Sensitivity	Specificity	Precision	F Measure	Kappa
E1	86.42	83.78	88.64	86.11	84.93	0.73
E2	81.48	71.43	96.88	97.22	82.35	0.64
E3	86.42	83.78	88.64	86.11	84.93	0.73
E4	85.19	81.58	88.37	86.11	83.78	0.70
E5	86.42	82.05	90.48	88.89	85.33	0.73

TABLE XII. PERFORMANCE COMPARISON OF ENSEMBLES ON LONG BEACH DATASET

Techniques	Accuracy	Sensitivity	Specificity	Precision	F Measure	Kappa
E1	84.09	89.47	80.0	77.27	82.92	0.68
E2	81.82	79.17	85.0	86.36	82.61	0.64
E3	86.36	86.36	86.36	86.36	86.36	0.73
E4	84.09	85.71	82.61	81.82	83.72	0.68
E5	88.64	90.48	86.96	86.36	88.37	0.77

TABLE XIII. PERFORMANCE COMPARISON OF ENSEMBLES ON COMBINED DATASET

Techniques	Accuracy	Sensitivity	Specificity	Precision	F Measure	Kappa
E1	84.69	83.92	85.31	82.19	83.05	0.69
E2	82.5	75.28	91.55	91.78	82.72	0.65
E3	86.25	83.55	88.69	86.99	85.24	0.72
E4	86.88	85.62	87.93	85.62	85.62	0.74
E5	85.0	84.03	85.80	82.88	83.45	0.69

TABLE XIV. AVERAGE PERFORMANCE COMPARISON ON ALL DATA SETS OF ENSEMBLES

Techniques	Accuracy	Sensitivity	Specificity	Precision	F Measure	Kappa
E1	85.62	85.77	85.54	83.34	84.46	0.71
E2	81.88	75.12	91.60	92.29	82.72	0.64
E3	87.64	86.06	89.03	87.87	86.95	0.75
E4	87.06	85.91	88.08	86.69	86.27	0.74
E5	86.15	85.0	86.99	85.49	85.24	0.57

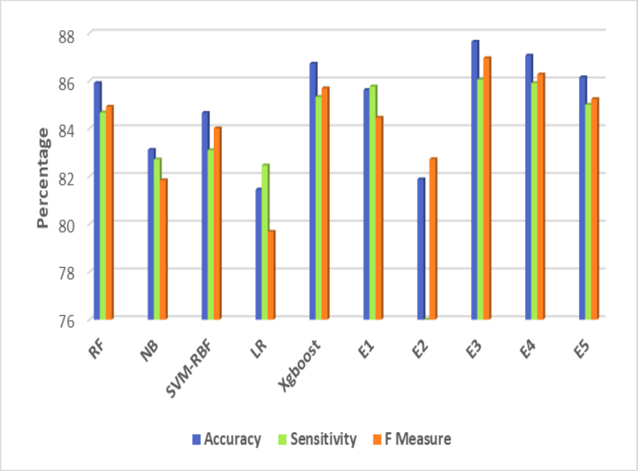

Fig. 5. Performance comparison of all individual and all ensembles.

VI. CONCLUSION AND FUTURE SCOPE

This article presents a reliable framework which can be used for predicting the cardiovascular disease. It deals with data cleaning by removing the noise from the data like outliers and missing values. For the Hungarian and Long Beach data to overcome the class imbalance, Synthetic Minority Oversampling Technique-SMOTE was used. Support Vector Machine Radial (SVM-R), Logistic Regression (LR), Naïve Bayes (NB), Random Forest (RF) and XGBoost were first implemented on all five data sets. Then five different ensembled classifiers were constructed. First ensemble classifier E_1 is designed considering all individual classifiers. Based on majority voting the final decision was taken for prediction. For second ensembled classifier E_2, all independent classifiers were considered but the prediction was done based on weighted majority voting. The weights to the classifiers were assigned depending on the accuracy of the classifier. For third, fourth and fifth ensembles, top three performers were identified on all five data sets. Out of these five different combinations of top performers, three unique combinations were selected. Hence, the third classifier E_3 consists of RF, SVM-R and XGBoost. The fourth ensembled E_4 is made up of RF, NB and XGBoost. The fifth one E_5 is constructed with NB, SVM-R and XGBoost. An exhaustive comparison of all individual classifier along with all ensembled classifiers was done. For any machine learning technique, the true classification rate is very important. Hence accuracy was given first importance. But at the same time for any medical study, the False Negative number is crucial. This number should be as less as possible. Therefore, sensitivity parameter is also focused. 10-fold cross validation was performed for validation. The results have shown that the amongst individual classifier XGBoost has performed the best on all data sets with average 86.7% accuracy, 85.3%sensitivity, 87.6% specificity, 86.1%Precision, 85.6%F-measure and 0.73 as Kappa value. The parameters were found to be improved with E_3 ensembled classifier as 87.6% accuracy, 86.0% Sensitivity,89.0% Specificity,87.8% Precision,86.9% F-measure and 0.75 as Kappa value. Thus, SRPS proves out to be most reliable for CVD prediction amongst all discussed. The future endeavor of the study could be the use of subset of the data set in terms of the attributes. In this study, all features were used for diagnosis. The further improvement can be made by using different feature selection methods like Wrapper method, Correlation based feature selection method, etc. Also, Principal Component Analysis can be used to reduce the dimension. Further, this proposed ensembled classifier can be used for other disease prediction as well.

ACKNOWLEDGMENT

We would like to acknowledge the guidance and the support given by MIT ADT faculty, research colleagues during the development of manuscript and MIT ADT library for provision of various resources.

REFERENCES

[1] I. Mohan, R. Gupta, A. Misra, K. Sharma, A. Agrawal, N. Vikram, et al. “Disparities in Prevalence of Cardio metabolic Risk Factors in Rural, Urban-Poor, and Urban-Middle Class Women in India”, Available: https://pubmed.ncbi.nlm.nih.gov/26881429/ , 2016.

[2] World Health Organization ,"Cardiovascular Disease", who.int,https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1, accessed on 8 January 2022.

[3] G. Allan, F. Nouri, C. Korownyk, M. Kolber, B. Vandermeer, J. McCormack. "Agreement among cardiovascular disease risk calculators. Circulation.".127(19):1948-56. doi: 10.1161/CIRCULATIONAHA.112.000412. Epub 2013 Apr 10. PMID: 23575355, 2013.

[4] L. Yang., H. Wu., X. Jin,P.Zheng,S,Hu,W.Yu,*et al.* "Study of cardiovascular disease prediction model based on random forest in eastern China." *Sci Rep* 10 **,** 5245. doi.org/10.1038/s41598-020-62133-5,2020.

[5] J. Kim., Y. Jeong, J. Kim, J.Lee,D.Park,H.Kim, "Machine Learning-Based Cardiovascular Disease Prediction Model: A Cohort Study on the Korean National Health Insurance Service Health Screening Database". Diagnostics, 11, 943. doi.org/10.3390/diagnostics11060943,2021.

[6] S. Cho, S. Kim, S. Kang, K. Lee, D. Choi, S. Kang, *et al.* "Pre-existing and machine learning-based models for cardiovascular risk prediction. *SciRep* **11,** 88862021. doi.org/10.1038/s41598-021-88257,2021.

[7] A. Dimopoulos, M. Nikolaidou.M, Caballero, W. Engchua ,A.Niubo,H,Arndt,e*tal.* Machine learning methodologies versus cardiovascular risk scores, in predicting disease risk. *BMC Med Res Methodology*.18**,** 179,2018.

[8] S. Weng, J. Reps, J. Kai, J. Garibaldi,N. Qureshi,et al, "Can machine-learning improve cardiovascular risk prediction using routine clinical data?". PLoS ONE12(4): .e0174944. doi.org/ 10.1371/journal.pone.0174944,2017.

[9] A. Alaa, T. Bolton, E. Angelantonio ,J.Rudd,M.Schaar, et.al, "cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants." PLoS ONE 14(5): e0213653. https://doi.org/10.1371/journal.pone.0213653,2021.

[10] A. Sarraju, A. Ward, S. Chung, J.Li, D. Scheinker, F. Rodríguez et al. "Machine learning approaches improve risk stratification for secondary cardiovascular disease prevention in multi-ethnic patients." Open Heart;8: e001802. doi:10.1136/ openhrt-2021-001802,2021.

[11] L. Hafsa, A. Salem, H. Henda, H.Ghezala,et al, "Does data cleaning improve heart disease prediction?" Procedia Computer Science, Volume 176, Pages 1131-1140, ISSN 1877-0509, doi.org/10.1016/j.procs.2020.09.109,2020.

[12] A.Rahim,Y.Rasheed,F.Azam,M.Anwar,M.Rahim,A.Muzaffar, "An Integrated Machine Learning Framework for Effective Prediction of Cardiovascular Diseases", IEEE Access, vol. 9, pp.106575-106588, doi: 10.1109/ACCESS.2021.3098688,2021.

[13] J.Beunza,E.Puertas,E.Ovejero,G.Villalba,E.Condes,G.Koleva,et al, "Comparison of machine learning algorithms for clinical event prediction (risk of coronary heart disease)", Journal of Biomedical Informatics, Volume 97,103257, ISSN 1532-0464, doi.org/10.1016/j.jbi.2019.103257,2019.

[14] G. N. Ahmad, S. Ullah, A. Algethami,H.Fatima,S.Akhter, "Comparative Study of Optimum Medical Diagnosis of Human Heart Disease Using Machine Learning Technique with and Without Sequential Feature Selection", in IEEE Access,vol.10,pp.2380823828,doi:10.1109/ACCESS.2022.3153047, 2022.

[15] Y. Muhammad, M.Tahir., M.Hayat, K.Chong. "Early and accurate detection and diagnosis of heart disease using intelligent computational model". *Sci Rep* 10, 19747. doi.org/10.1038/s41598-020-76635-9,2020.

[16] P.Ghosh,S.Azam,M.Jonkman,A.Karim,F.Shamrat,E.Ignatious,et al., "Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms with Relief and LASSO Feature Selection Techniques," in IEEE Access vol. 9,pp.19304-19326,doi:10.1109/ACCESS.2021.3053759, 2021.

[17] J. P. Li, A. U. Haq, S. U. Din, "Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare", in IEEE Access ,vol. 8,pp. 107562-107582,doi:10.1109/ACCESS.2020.3001149, 2020.

[18] A. Escamilla, A. Hassani, E. Andrès, "Classification models for heart disease prediction using feature selection and PCA," Informatics in Medicine Unlocked, Volume 19,100330,ISSN23529148,doi.org/10.1016/j.imu.2020.100330, 2020.

[19] A. Dinh, S. Miertschin, A. Young ,S.Mohanty, "A data-driven approach to predicting diabetes and cardiovascular disease with machine learning". *BMC Med Inform Decision Making* **,** 211. doi.org/10.1186/s12911-019-0918-5,2019.

[20] C. Latha, S. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques", Informatics in Medicine Unlocked, Volume 16,100203, ISSN 2352-9148, doi.org/10.1016/j.imu.2019.100203,2019.

[21] Heart Disease Data: UCI Machine Learning Repository Center for Machine Learning and Intelligent Systems [online]Available: https://archive.ics.uci.edu/ml/datasets/heart+disease, accessed on 8 January 2022.

[22] Statlog(Heart) Data Set : UCI Machine Learning Repository Center for Machine Learning and Intelligent Systems [online] Available: https://archive.ics.uci.edu/ml/datasets/statlog+(heart), accessed on 8 January 2022.

[23] A. AbuHalimeh, "Improving Data Quality in Clinical Research Informatics Tools". Front. Big Data, 5:871897. doi: 10.3389/fdata.2022.871897,2022.

[24] G. Rekha, A. Tyagi, N. Sreenath, S.Mishra , "Class Imbalanced Data", Open Issues and Future Research Directions,2021 International Conference on Computer Communication and Informatics (ICCCI), pp. 1-6, doi: 10.1109/ICCCI50826.2021.9402272,2021.

[25] N. Chawla, K. Bowyer, L. Hall, W.Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique", Journal of Artificial Intelligence Research, Volume16, pages 321-357, doi.org/10.1613/jair.95,2002.