An Effective Heart Disease Prediction Framework based on Ensemble Techniques in Machine Learning

Deepali Yewale[1], S. P. Vijayaragavan[2], V. K. Bairagi[3]
Department of Electronics and Communication, Bharath Institute of Higher Education and Research, Chennai-611026, India[1]
Department of Electrical and Electronics Engineering, Bharath Institute of Higher Education and Research, Chennai-611026, India[2]
Department of Electronics and Telecommunication, AISSMS Institute of Information Technology, Pune-411001, India[1, 3]

***Abstract*—To design a framework for effective prediction of heart disease based on ensemble techniques, without the need of feature selection, incorporating data balancing, outlier detection and removal techniques, with results that are still at par with cutting-edge research. In this study, the Cleveland dataset, which has 303 occurrences, is used from the UCI repository. The dataset comprises 76 raw attributes, however only 14 of them are listed by the UCI repository as significant risk factors for heart disease when the dataset is uploaded as an open source dataset. Data balancing strategies, such as random over sampling, are used to address the issue of unbalanced data. Additionally, an isolation forest is used to find outliers in multivariate data, which has not been explored in previous research. After eliminating anomalies from the data, ensemble techniques such as bagging, boosting, voting, stacking are employed to create the prediction model. The potential of the proposed model is assessed for accuracy, sensitivity, and specificity, positive prediction value (PPV), negative prediction value (NPV), F1 score, ROC-AUC and model training time. For the Cleveland dataset, the performance of the suggested methodology is superior, with 98.73% accuracy, 98% sensitivity, 100% specificity, 100% PPV, 97% NPV, 1 as F score, and AUC as 1 with comparatively very less training time. The results of this study demonstrate that our proposed approach significantly outperforms the existing scholarly work in terms of accuracy and all the stated performance metrics. No earlier research has focused on these many performance parameters.**

Keywords*—Machine learning; heart disease; ensemble techniques; random over sampling, isolation forest

I. INTRODUCTION

Cardiovascular disease (CVD) is considered to be the foremost reason of death in the world. It is estimated that half of all CVD cases occur in Asia. Besides, near about three-quarters of all the global mortality are anticipated to happen because of persistent diseases by 2021, with 75% of deaths due to heart disease. These circumstances create an imperative need to design decision support system (DSS) for early prediction of heart disease. About 80 % of CVD are preventable if predicted at an early stage.

Traditional method of heart disease diagnosis includes extensive examination of patient. In this case the diagnosis of disease totally depends upon domain experts' knowledge and exactness of collected clinical data. This invasive method of diagnosis is not reliable and efficient. So there is need for cost effective and highly accurate Non-invasive approach such as Machine Learning (ML) to design prediction model.

ML is able to filter out pertinent connection between tremendous measures of data. Advancement in Artificial intelligence has played very significant role in the new era of computing. Various researchers have tried and tested data mining techniques in the healthcare domain and found it to be outperforming. The limitation of the prevailing analysis was the utilization of tedious task of feature engineering within the classification process. The principal inspiration of this research is to makeover data to usable form so that researchers can use it to design DSS to enhance safety of the patients.

In this work, proposed a model with data pre-processing, data balancing and outlier detecting followed by ensemble classifier. Standard scalar is used in the data pre-processing stage where each attributes with standard deviation of one is achieved. Imbalanced data is explored and oversampling technique was proposed to improve the reliability of the model. Anomalies or outliers are considered as noise in the data and may lead to misclassification. Furthermore proposed anomalies detection and removal using Isolation forest algorithm. Besides, Ensemble techniques bagging, boosting, Voting and stacking are applied to measure the effectiveness of the proposed methodology. This article focuses on the prediction of cardiac disease employing ensemble techniques of ML without feature engineering.

The main contributions of the proposed work include:

- A novel combination of data standardization, data balancing, and outlier detection to transform the data into usable form.
- The study involves the isolation forest for outlier detection of multivariate data, which has not been extensively explored in the previous research.
- The study assesses the performance of heart disease prediction system using ensemble techniques without feature selection designed, which have not been studied in depth in previous research.

The rest of the paper is structured as follows: In Section II literature analysis is provided, research gap is highlighted in Section III; Methodology is projected in Section IV. The detailed results of the proposed approach are presented in Section V followed by comparative study with existing research work in Section VI, and finally conclusion and future scope in Section VII.

II. Literature Analysis

According to WHO, heart disease represent predominant reason of death in developing countries. One of the reasons to fail in the treatment of heart disease is unidentified pattern with cardiac data. Machine learning has been proved the remedy for that, as it is able to extract the pattern in cardiac data to predict the heart disease.

Researchers have explored and evaluated several methodologies, including single-base classifiers, ensemble approaches, and hybrid techniques as the prediction model. Furthermore, data pre-processing techniques, feature selection techniques, and optimization approaches were employed to improve the performance of the prediction system.

Many researchers implemented basic ML classifiers on a cardiac dataset and achieved good results. Authors suggested a modified random forest [1] to boost the prediction ability of the classifier. The proposed work achieved the highest accuracy of 86.84% with the UCI Cleveland dataset. Author [2] implemented Logistic regression (LR), K nearest neighbour (KNN) and Random Forest (RF) classifiers on a medical dataset from the UCI repository. The highest accuracy achieved with KNN is 87.5%, when implemented on the Python platform.

The above mentioned, state of the art research used conventional algorithms to design decision support system for heart disease prediction and it has been observed that the average accuracy is below 90%.

Instead of relying on a conventional model to provide an exceptional solution, the ensemble method leverages the strengths of numerous models to mitigate the limitations of a single model. In [3], proposed homogeneous ensemble learning using an accuracy-based weighted ageing classifier. The proposed model achieved an accuracy of 93% on the Cleveland dataset. Instead of utilizing conventional single model, majority vote Ensemble model can be used [4] in the prediction system of heart disease. This approach has produced 90% accuracy for Cleveland dataset. According to this literature survey, an ensemble method has shown to be more successful than a single model strategy.

Many feature selection methods have been proposed by researchers for obtaining more relevant features from a given dataset. Javeed et al. [5] proposed a randomized search algorithm (RSA) to get the optimal subset of features, and grid search optimized RF was used as a classifier. The experimental results have achieved 93.33% accuracy while improving the training accuracy as well. Muhammad et al. [6] proposed a model where four feature selection methods, namely fast correlation-based feature selection (FCBF) , minimum redundancy, maximal relevance (mRMR), least absolute and selection operator (LASSO), and Relief, were tested on 10 different machine learning classifiers. It was found that, for the features chosen by FCBF, ETC's accuracy increased from 92.09 % to 94.41 % compared to complete features. Dissanayake et al. [7] performed research where filter, wrapper, and embedded feature selection approaches were implemented. The test findings show that DT with backward elimination wrapper feature selection outperforms with an accuracy of 88.52 %.

III. Research Gap

The majority of the scholarly work is focused on improving accuracy via feature selection techniques. However, to eliminate the data cleaning operations while yielding high disease prediction accuracy, a computationally effective feature selection approach is required [8].There is need to investigate a new intelligent technique to generate a meaningful concise set of features. Noise and outliers present in the data make it difficult to select exact features [9]. As feature selection is a tiresome activity and only some of the existing work discussed in the literature is able to predict heart disease with good accuracy, there is a dire need to test machine learning framework without feature selection for the effective prediction of cardiovascular disease [10].

The dataset discussed in the existing work is imbalanced with an uneven contribution of the majority and minority classes. Class imbalanced has not been amply focused in the previous research. The problem of class imbalance must be taken care of before implementing any classification mechanism. On the other hand; only a few researchers have worked on outlier removal from dataset. There is a need for an outlier detection method that can differentiate between normal data and outliers [11].Only few Researchers have experimented unsupervised outlier detection techniques such as DBSCAN, isolation forest, K–means clustering. There is need to experiment and analyse Isolation forest for outlier detection. The utilization of ensemble based algorithms needs to be experimented rigorously and analysed for effective prediction of heart disease.

The unique aspect of the proposed research work is to design a framework for heart disease prediction that can handle the problems of imbalanced class, outlier detection and still conveys comparable performance index without any feature engineering.

IV. Materials and Methods

In this research, we present a new paradigm for predicting cardiac disease, which can predict the presence and absence of heart disease reliably, as shown in Fig. 1.

A. Dataset Description

The Cleveland dataset used for this proposed work has 303 instances with 76 clinical and physical parameters. Most of the research work has chosen just 14 features in their scholarly work as these attributes are the most significant in the prediction of heart disease. Other attributes such as exercise protocol and time when ST measure depression was performed, had minor effects on heart disease and so 62 attributes are omitted by researchers.

As seen in Table I, the UCI repository specifically mentions these 14 attributes when uploading the dataset for open access.

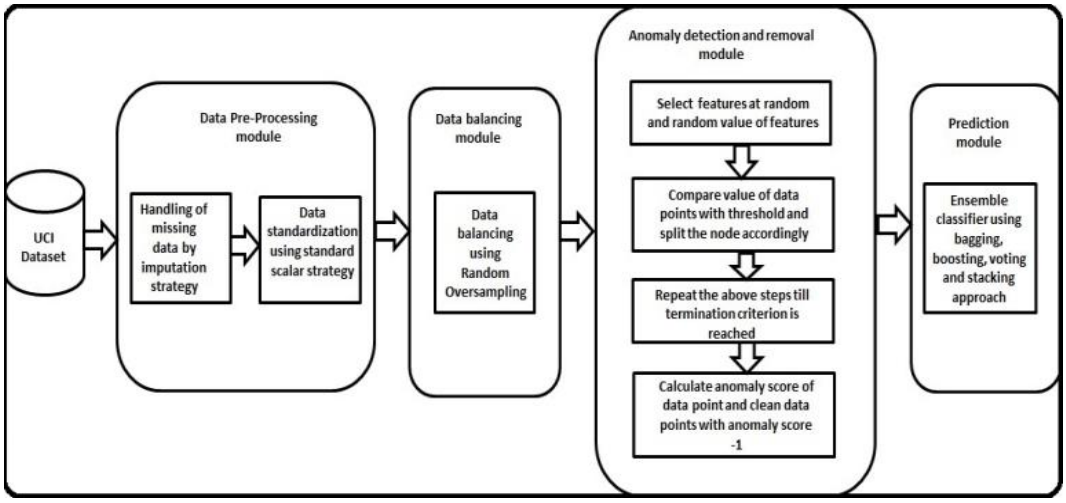


Fig. 1. Proposed heart disease prediciton system using ramdom oversampling, isoalation forest and ensemble prediction model.

TABLE I. IMPORTANT 14 ATTRIBUTES FROM 76 ATTRIBUTES OF UCI DATASET

Sr. No.	Name of Attribute	Position in the dataset	% of the data Complete
1	Age in years	3	100
2	Sex	4	100
3	Chest pain type	9	100
4	Resting Blood Pressure	10	100
5	Serum Cholesterol	12	100
6	Fasting blood sugar	16	100
7	Resting ECG	19	100
8	Maximum Heart Rate	32	100
9	Exercise-induced angina	38	100
10	ST depression	40	100
11	The slope of the ST segment	41	100
12	Number of containers colored by fluoroscopy	44	98.67
13	Thalassemia	51	99.33
14	Diagnosis value	58	100

Among these, 13 are independent variables and 1 is a dependent target variable for the diagnosis of heart disease, where 0 represents the absence of heart disease and 1 represents the presence of heart disease.

B. Data Preprocessing

Before catering data into the machine learning classifier, it is important to analyse and pre-process the data to improve its quality. A few attributes, as shown in Table I, have missing values. Missing values are replaced by the mean value of those attributes [12].

Creating a data-analysis-based decision support system necessitates standard data, which frequently necessitates pre-processing activities such as data cleansing, pruning, and scaling. Data standardisation is performed to scale each feature to unit variance. Attributes assessed at different scales do not contribute equally to the model fitting and may result in bias. To address this possible issue, feature-wise standardisation is utilised prior to model fitting [13]. The feature in each column of x is normalized independently, so that each feature has a mean μ =0 and a standard deviation σ =1.

A value is standardized as (1).

$$z = \frac{x-\mu}{\sigma} \quad (1)$$

Where mean μ is defined in (2).

$$\mu = \frac{1}{N}\sum_{i=1}^{N} Xi \quad (2)$$

And standard deviation σ is as in (3).

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(xi-\mu)^2} \quad (3)$$

Where N is the number of instances in each column (N=303).

C. Data Balancing

As per the exploratory data analysis of the target variable, the given Cleveland dataset has 138 instances of healthy people and 165 instances of people with heart disease. The dataset has unequal distribution of negative (0) and positive (1) instances.

Because of this unequal number of positive and negative classes, it will be difficult for machine learning models to learn the pattern of the dataset and it may hamper the performance of the model [14]. To address the problem of imbalanced data, the random oversampling technique is proposed. It randomly duplicates examples from the minority class with substitution and adds them into the training dataset. The superiority of random oversampling is that all individuals from the minority and majority classes are maintained; therefore, no information from the original training set is lost. Synthetic Minority Oversampling Technique (SMOTE) is another method of data for oversampling where samples are created synthetically but it may create noise with high dimensional data. Data augmentation using SMOTE may provide diverse results and may not always be beneficial for medical data [15].

D. Anomaly Detection

Anomaly detection is the technique of identifying outliers in data. Researchers have preferred to use unsupervised anomaly detection models. Isolation forest is a unique method based on this isolation property of outliers and is fundamentally different from other density-based and cluster-based outlier detection methods [16]. In this research paper, it is proposed to use an isolation forest (iForest) to isolate the anomalies from the data samples.

Here is the algorithm to compute an isolation tree:

1) Choose a feature at random from data and refer it as f.

2) Choose a value at random from the feature f and utilize as threshold‘t’.

3) Data points with f < t are saved in Node 1 whereas the data points with f ≥ t kept in Node 2.

4) Steps 1–3 repeated for Node 1 and Node 2.

5) Stop the process when the tree has reached full maturity or when a termination requirement is fulfilled.

An isolation tree can be extended to an isolation forest—an ensemble of multiple isolation trees.

The isolation forest in sklearn has 2 important inputs:

n_estimators: Number of Isolation trees to be trained

Contamination: Fraction of anomalous data points.

In our case we suspect 5% of the data to be anomalous and set contamination to 0.05.

Steps in building an Isolation forests:

1) Construct an Isolation Tree either from the entire feature set or a randomly chosen subset of the feature set.

2) Construct n such Isolation trees.

3) Calculate an Anomaly score for each data point using formula in (4).

$$s(x, n) = 2^{-E(\hbar(x)/c(n)} \qquad (4)$$

s = score (closer to 1: outlier, closer to 0: normal data point),

E($\hbar$(x)= Average path length taken by data point x,

c(n) = Average path length of every terminal nodes.

Isolation forest can be used for univariate as well as multivariate dataset. Let us consider our case of the multivariate dataset as shown in Fig. 2.

Isolation tree divides the data into “boxes”. It has the property that it segregates the region containing anomalies earlier than the boxes containing normal data points. If the feature has an anomaly, the anomalous point will be far away from the normal points in the data. It helps isolation forests to isolate out anomalies relatively early in the splitting process.

As shown in the figure, the anomaly can be detected at split 2 only. If we go on splitting the data, few normal points got isolated much later as shown in split 4. Isolation Forest can detect the outliers faster and require less memory as compared to other algorithms.

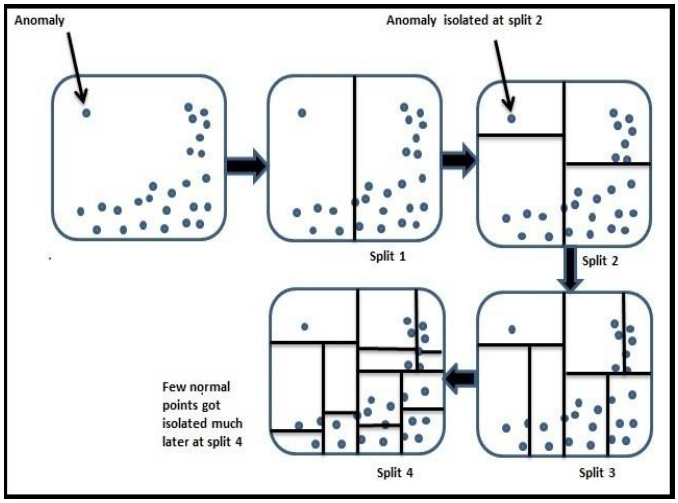


Fig. 2. Splitting process in isolation forest where anomalies are isolated early in split 2 as compared to normal data point isolated in split 4.

E. Prediction Module: Ensemble ML Techniques

Ensemble learning is a machine learning model in which numerous models (commonly referred to as weak learners or base models) are trained to handle the same issue and then integrated to provide improved results. We tested our model using bagging (RF, ETC), boosting (XGBoost, AdaBoost), voting (hard, soft voting), and staking (RF+SVM) ensemble techniques.

Here we discussed the ensemble techniques, implemented in our proposed work.

Bagging: In this kind of approach, several instances of the same base model are trained in parallel (independently from each other) on separate bootstrap samples and then aggregated in some form of "averaging" process.

We have implemented Random Forest and Extra Tree classifier as bagging techniques.

Random Forest: This classifier belongs to the ensemble classifier family. It employs decision tree models to improve prediction outcomes. It generates numerous trees from the training data set, and a bootstrap approach is used to each tree. The random forest technique is a bagging method in which deep trees fitted on bootstrap samples are blended to create an output with lower variance.

Extra Tree classifier: Extra Trees classifier is a type of ensemble technique that delivers classification result by accumulating the results of multiple uncorrelated decision trees grouped together in a "forest." It differs as compared to Random forest in a way the decision tree is built in the forest. It separates nodes by selecting cut-points completely at random, and it also grows the trees using the whole training sample.

Boosting: Boosting is an ensemble modeling strategy that seeks to construct a strong classifier from a collection of weak classifiers to reduce training errors. A random sample of data is chosen, fitted with a model, and then trained sequentially—that is, each model attempts to compensate for the shortcomings of its predecessor. In this work, proposed to use XGBoost and AdaBoost boosting techniques.

XGBoost: XGBoost is the state of the art gradient boosted tree algorithm that boosts the performance of weak learners. It uses greedy algorithm to calculate the best split. To begin, a weak classifier is fitted to the data. It adds another weak classifier to upgrade the present model's performance, without losing the prior classifier's performance. The same process is continued and it employs a gradient descent approach to reduce the loss when adding new models. Each new classifier must take into account where the prior classifiers failed to perform well. To generate a new model, the method constantly reduces the errors of prior models in the gradient direction.

AdaBoost: AdaBoost is a boosting machine learning algorithm that use weighted linear combination to cascade numerous weak learners into a particular classifier. AdaBoost uses a learning technique to re-weight samples of the original training data in a sequential manner. It is an iterative approach, with each iteration giving more weight to the misclassified occurrences from the preceding iteration. Each instance is initially allocated an identical weight, and iteratively, the weights of all wrongly classified instances are increased while the weights of successfully classified examples are decreased. The algorithm recursively applies the base classifier with fresh weights to the training data. The final classification model produced is a linear combination of all the models developed over the rounds. AdaBoost completely considers each classifier's weight; nonetheless, it is vulnerable to outliers and noisy data.

Voting Classifier: Voting classifier aggregates the output of each classifier provided to it and produces the final prediction of the class label of a new instance based on voting. The voting can be of two types, hard or soft. Simple majority voting is utilized in the situation of hard voting. In this situation, the class with the highest number of votes is projected. A forecast is created for soft voting by averaging the class-probabilities of each classifier. The projected class is the one with the highest average probability. In the proposed work, model is checked for both hard and soft voting. In the proposed model LR, NB, DT, SVM, KNN has been ensemble as base models in hard voting and LR, NB, DT, KNN are used in soft voting.

Stacking: The stacking approach is a two-layered ensemble technique. The top layer comprises of all the baseline models used to predict the outcomes on the test dataset. The second layer consists of a Meta-Classifier, which accepts all of the baseline model outcomes as input and generates new prediction. The second layer combines the output of the first layer. Here, RF is used as baseline model and SVM as Meta Classifier.

V. Results and Analysis

The experiment has been conducted on Python platform using different libraries on an Intel Core i5 processor 9300H CPU with 2.40GHz, 4GB NVIDIA GTX 1650 graphical processing unit Lenovo machine equipped with 8GB RAM. Exploratory data analysis and data pre-processing have been performed. The dataset is divided into 75% of training data and 25% of testing data. Stratified k fold has been introduced in the dataset's training phase to avoid sampling bias. The accuracy, precision, recall, sensitivity, specificity, PPV, NPV, F1 score, ROC_AUC score, and computing time of the model are used to validate the performance of the suggested technique.

The experiment focused on the evaluation of the model by implementing random oversampling and isolation forest for various ensemble classifiers. The number of instances after implementing Random oversampling to Cleveland dataset is as shown in Table II.

Before implementation of random oversampling the total number of instances in the dataset are 303 with 138 instances indicating absence of heart disease and 165 instances for presence of heart disease. After processing data for random oversampling, the total number of samples is 330 since the positive and negative class instances are evenly distributed and equal to 165.

TABLE II. UCI Cleveland Dataset over Sampling Results

Class	Absence of HD	Presence of HD
Before Random Oversampling	138	165
After Random Oversampling	165	165

The isolation forest is used to identify and clear outliers from the dataset. After removing outliers, the dataset has 313 samples. The performance assessment of the ensemble classifier model using the proposed methodology is shown in Table III.

It demonstrates that the implementation of random oversampling and an isolation forest has given excellent performance for all the ensemble classifiers. Many existing research papers put emphasis on feature engineering, but our proposed approach works on the whole featured dataset and demonstrates the significance of data balancing and outlier removal in the prediction of heart disease. The suggested approach outperformed numerous existing studies in the literature without requiring high computational time. The accuracy of the proposed model for ensemble classifiers is in the range of 97.47 % to 98.73 % for the Cleveland dataset. The highlighted column demonstrates that soft voting provides excellent result among all the implemented ensemble methods.

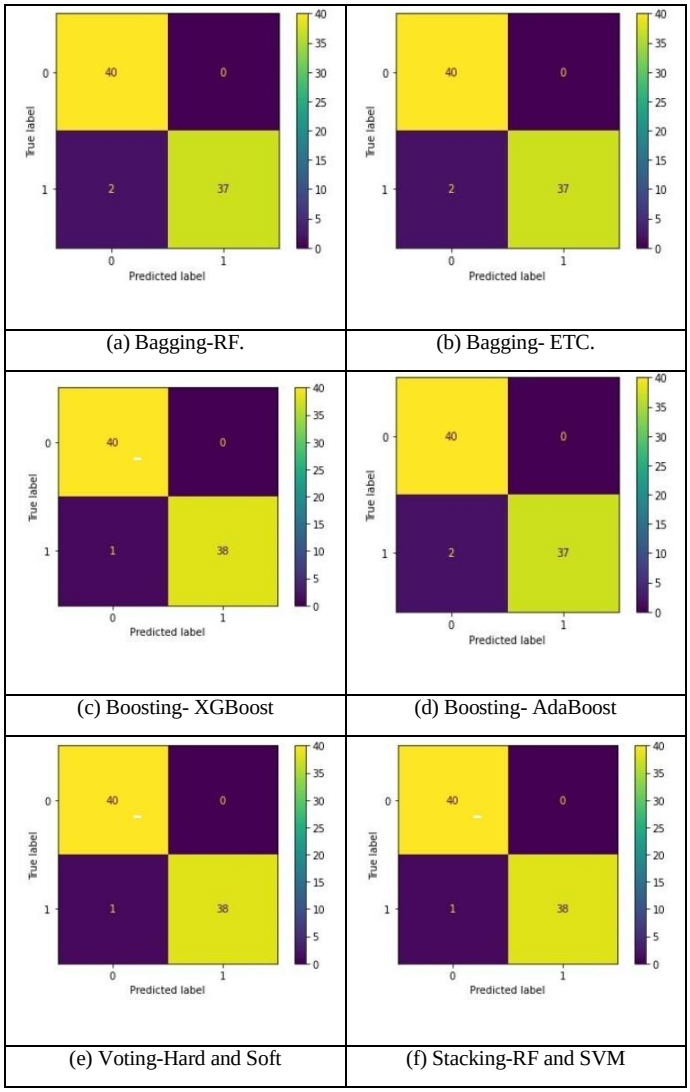
Confusion matrix for implemented ensemble classifiers is as shown in Fig. 3.

From the confusion matrix, it has been observed that, RF, ETC and AdaBoost Ensemble Techniques provide False Negative (FN) value of 2, that means two patients with actual heart disease are incorrectly predicted as non-heart disease persons. For XGBoost, Voting and hybrid ensemble techniques, FN value is 1 indicating only one heart disease patient incorrectly predicted as non-heart disease patient by the model.

Fig. 4 explores the graphical presentation of all the performance parameters of the proposed methodology with all the ensemble techniques implemented.

TABLE III. PERFORMANCE EVALUATION OF PROPOSED METHODOLOGY FOR CLEVELAND DATASET

Metric	RF	ETC	XGBoost	AdaBoost	Hard Voting	Soft Voting	Stacking
Accuracy	97.47	97.47	98.73	97.47	98.73	98.73	98.73
Sensitivity	95	95	98	95	98	98	98
Specificity	100	100	100	100	100	100	100
PPV	100	100	100	100	100	100	100
NPV	95	95	97	95	97	97	97
F1-Score	0.97	0.97	0.99	0.97	0.99	0.99	0.99
AUC	0.998	0.982	0.995	0.995	--	1	0.980
Computational Time in sec.	0.054	0.062	0.144	0.114	0.016	0.011	0.631



(a) Bagging-RF. (b) Bagging- ETC.

(c) Boosting- XGBoost (d) Boosting- AdaBoost

(e) Voting-Hard and Soft (f) Stacking-RF and SVM

Fig. 3. Confusion matrix of implemented ensemble classifiers.

It reveals that all the ensemble techniques provide excellent performance when applied to random over sampling to cater with imbalance data and isolation forest for anomaly detection to the Cleveland dataset with 303 instances without any feature selection.

Proposed work is additionally assessed for ROC_AUC analysis. The ROC curve is a depiction of the True positive rate vs the False positive rate. In other words, it is trade-off between sensitivity and specificity. The area the ROC curve (AUC) is said to be excellent for values between 0.9-1. Fig. 5 shows ROC-AUC curves for various ensemble techniques applied in the experimentation. Accuracy and AUC are two important metrics for the binary classification problem. The values of AUC for all the techniques are found to be excellent as per the observation in the ROC curve.

Fig. 6 demonstrates the two cases of the heart disease prediction system as heart disease positive and heart disease negative on a GUI application created using tkinter in Python.

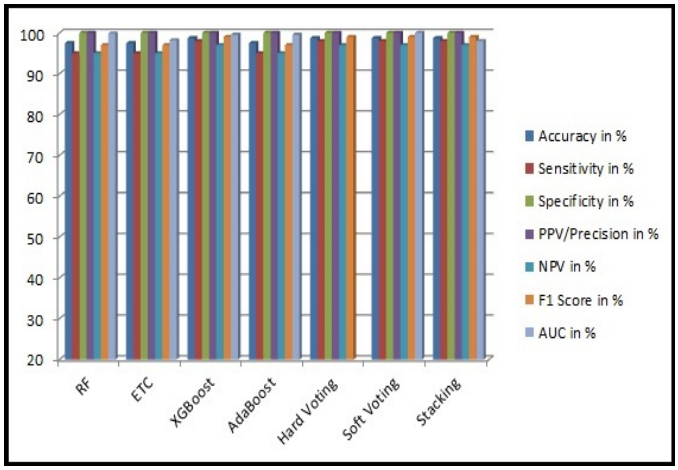


Fig. 4. Performance parameters of proposed ensemble technique for cleveland dataset.

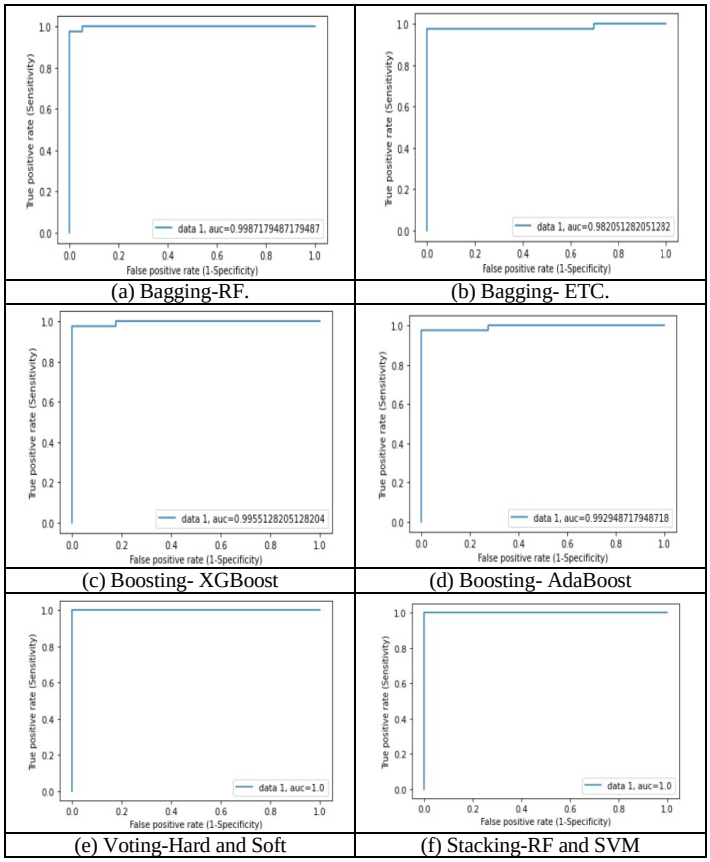
The proposed work's credibility is demonstrated by comparing its findings for the same dataset with three approaches listed below.

Approach 1: Without Random Over sampling and Isolation Forest for Ensemble Techniques.

Approach 2: Without Isolation Forest with Random Over sampling for Ensemble Techniques.

Approach 3: Without Random Over sampling with Isolation Forest for Ensemble Techniques.

In the first approach, Cleveland dataset is processed without random over sampling and no isolation forest. The results of this implemented methodology are tabulated as shown in Table IV.

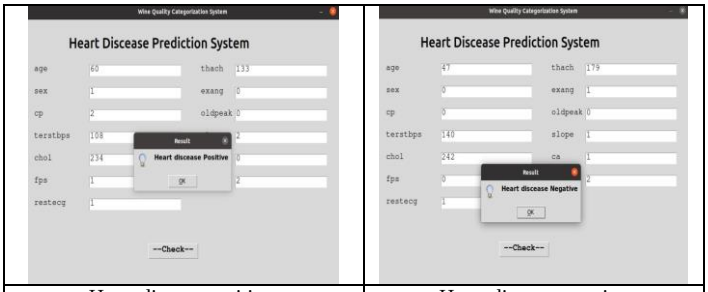


(a) Bagging-RF. (b) Bagging- ETC.

(c) Boosting- XGBoost (d) Boosting- AdaBoost

(e) Voting-Hard and Soft (f) Stacking-RF and SVM

Fig. 5. ROC curve of various ensemble Techniques used in the experiment, displaying corresponding AUC values.



Heart disease positive. Heart disease negative

Fig. 6. GUI for Heart disease positive and negative cases using tkinter in python.

The highlighted column in Table IV demonstrates that, soft voting is the best performing ensemble model with 96.05% accuracy for approach 1. But the performance of the proposed research is far better as compared to the implemented methodology in approach 1.

Similarly, in the second approach, Cleveland dataset is processed with random over sampling, no isolation forest and in the third approach Cleveland dataset is processed sampling with isolation Forest, no random over sampling.

The results of these implemented methodologies are tabulated as shown in Table V and Table VI respectively and the best results are highlighted.

The projected results in Tables IV, V and VI reveal that our proposed research work of implementation of Ensemble Techniques with random oversampling and isolation forest give excellent performance in terms of all performance matrices as compared to implementation strategy of all the three approaches.

TABLE IV. PERFORMANCE EVALUATION FOR CLEVELAND DATASET OF 303 INSTANCES WITHOUT RANDOM OVER SAMPLING AND ISOLATION FOREST (APPROACH 1)

Metric	RF	ETC	XGBoost	AdaBoost	Hard Voting	Soft Voting	Stacking
Accuracy	94.73	94.73	94.73	94.73	96.05	96.05	96.05
Sensitivity	92	92	92	92	92	92	92
Specificity	97	97	97	97	100	100	100
PPV	97	97	97	97	100	100	100
NPV	93	93	93	93	93	93	93
F1-Score	0.95	0.95	0.95	0.95	0.96	0.96	0.96
AUC	0.995	0.995	0.995	0.981	--	0.983	0.956
Computational Time in sec.	0.057	0.082	0.154	0.113	0.015	0.011	0.681

TABLE V. PERFORMANCE EVALUATION FOR CLEVELAND DATASET OF 303 INSTANCES WITHOUT ISOLATION FOREST WITH RANDOM OVER SAMPLING (APPROACH 2)

Metric	RF	ETC	XGBoost	AdaBoost	Hard Voting	Soft Voting	Stacking
Accuracy	95.18	95.18	95.18	95.18	95.18	95.18	95.18
Sensitivity	91	91	91	91	91	91	91
Specificity	100	100	100	100	100	100	100
PPV	100	100	100	100	100	100	100
NPV	90	90	90	90	90	90	90
F1-Score	0.95	0.95	0.95	0.95	0.95	0.95	0.95
AUC	0.966	0.947	0.980	0.950	--	0.971	0.971
Computational Time in sec.	0.067	0.071	0.173	0.123	0.024	0.015	0.719

TABLE VI. PERFORMANCE EVALUATION FOR CLEVELAND DATASET OF 303 INSTANCES WITHOUT RANDOM OVER SAMPLING WITH ISOLATION FOREST (APPROACH 3)

Metric	RF	ETC	XGBoost	AdaBoost	Hard Voting	Soft Voting	Stacking
Accuracy	94.44	94.44	94.44	94.44	94.44	94.44	94.44
Sensitivity	89	89	89	89	89	89	89
Specificity	100	100	100	100	100	100	100
PPV	100	100	100	100	100	100	100
NPV	89	89	89	89	89	89	89
F1-Score	0.94	0.94	0.94	0.94	0.94	0.94	0.94
AUC	0.966	0.941	0.966	0.941	--	0.969	0.951
Computational Time in sec.	0.062	0.077	0.140	0.093	0.015	0.0	0.625

VI. COMPARISON AND DISCUSSION

The reliability of the proposed work is demonstrated by comparing its findings to those of other state-of-the-art current systems conducting imbalanced data processing and outlier detection for the Cleveland dataset as shown in Table VII. Researcher [17] implemented imputation of missing values and outlier removing processes in order to provide quality data to ML model. Mahalanobis distance metric is used to drop the outliers in the data. Further, NB optimized with grid search found to provide accuracy of 84.8%. Instead of conventional model, an ensemble based majority voting scheme is proposed by researchers [18]. Outliers in the dataset are identified and removed using filter based techniques. From different combinations of ensemble, SVM +NB+ MLP ensemble provided highest accuracy of 84%. Researchers [19] proposed machine learning framework using data imbalance technique SMOTE and feature selection technique on ensemble (LR+KNN) classifier for heart disease prediction. Box plot technique is used to identify the outliers. The suggested architecture was assessed on the Framingham, heart disease, and Cleveland dataset and found to outperform them all. SMOTE based ANN [20] is proposed to Cleveland UCI dataset. The imbalance nature of the presented dataset is analyzed, and SMOTE oversampling strategy is proposed to improve the performance of the ANN classifier. Deep learning has been effectively used in heart disease prediction. Researchers experimented isolation forest for outlier detection for multivariate data using selected features from 13 features of Cleveland dataset [21]. Accuracy is found to be improved but sensitivity and specificity is very poor. Researchers used filter based feature selection and isolation forest for anomalies detection. The proposed approach found to be performing effectively for KNN with eight neighbours on UCI Cleveland dataset [22]. Researchers investigated anomaly detection using K-means clustering algorithm [23]. After removing anomalies, five classification techniques KNN, RF, SVM, NB and LR are used to build the prediction model. It is found that without anomalies RF and NB are performing better as compared to with anomalies in the dataset. DBSCAN is implemented to identify and remove the outliers, a hybrid SMOTE-Edited Nearest Neighbor (SMOTE-ENN) to balance the training data distribution and XGBoost as a classification model for heart disease prediction [24]. Most of the existing research work, proposed to use feature selection extensively. The state of the art research rarely used combination of data balancing and outlier detection in data pre-processing.

The suggested methodology outperforms prior research work utilizing data balancing techniques and outlier identification techniques. The performance of the data balancing solution using Random Oversampling and outlier detection using Isolation Forest on Ensemble Classifier is excellent in all performance metrics. The proposed model in our research outperformed previous models and research findings, with an accuracy of 98.73 % for Cleveland dataset.

TABLE VII. COMPARISON OF PROPOSED METHODOLOGY WITH STATE-OF-THE-ART RESEARCH FOR UCI CLEVELAND DATASET

Author Name, Year, Reference	Methodology	Accuracy	Sensitivity	Specificity	PPV	NPV	F Score	AUC Score
Sivaraman et al. (2021) [17]	Mahalanobsis distance+ Grid search optimization +NB	84.8	82.7	86.4	83.2	--	82.8	91
Bashir et al. (2021) [18]	Filter based outlier removing + Majority voting	84	84.80	83.22	--	--	84	--
Rahim et al. (2021) [19]	SMOTE+ Box plot +Feature importance (5 features)+ LR-KNN	98.0	--	--	--	--	--	--
Waqar et al. (2021) [20]	SMOTE+ ANN	96	95.7	--	96.1	--	95.7	100
Bharti et al. (2021) [21]	Isolation Forest+ Lasso FS+DL	94.2	82.3	83.1	--	--	--	--
Ramesh et al. (2022) [22]	Isolation Forest +Filter FS (7 features) +KNN	94.1	94.8	--	91.7	--	90.8	79.9
Ripen et al. (2021) [23]	K means clustering +RF	88	--	87	87	--	--	--
Fitriyani et al. (2020) [24]	DBSCAN+ SMOTE ENN+ XGBoost	98.4	98.3	98.3	98.5	--	98.3	--
Proposed Methodology	Random Oversampling+ Isolation Forest+ Ensemble Techniques	98.73	98	100	100	97	99	100

For binary classification problem of presence and absence of heart disease, accuracy and AUC are the most important metrics. The highlighted row in the Table VII demonstrates that the proposed methodology has highest accuracy and AUC as compared to state of the art research.

VII. Conclusion and Future Research

In this paper, we propose ensemble techniques that are supported by Random Oversampling and Isolation Forest for efficiently predicting heart disease. All the ensemble models are found to be performing excellently in all evaluation results. The accuracy range for the Cleveland dataset, for all models is 97.43% to 98.73%, sensitivity 95% to 98%, specificity 100%, precision 100%, NPV 95% to 97 %, F score 0.97 to 0.99, AUC score 0.98 to 1 with less computational overhead.

Most of the previous researchers implemented feature selection techniques to improve the accuracy of ML and DL models. Here we propose a model without any feature selection and achieve remarkably improved performance metrics. Experimental results prove that the ensemble approach with data pre-processing techniques resolves the issue of computational intricacy. Random oversampling and isolation forest significantly improve the performance of ensemble classifiers.

Data balancing with oversampling helps to make data more reliable by avoiding over fitting or under fitting the model. Noisy data removal with an isolation forest improves the quality of the data. The validity of the suggested framework on the UCI Cleveland dataset demonstrates that our framework is both trustworthy and efficient. It incorporates novel pre-processing techniques while also employing an inventive ensemble. Furthermore, the computational time is remarkably reduced with highly reliable results.

In the future, optimization techniques can be implemented for hyper parameter tuning to deploy the model. Because of the NP-hardness of feature selection approaches, a meta-heuristic feature selection method can be devised. There is a strong need for real-world clinical factors that are easily approachable and computed in real-time for the future of clinical cardiac disease detection via ML-centered systems.

References

[1] S. Vinayaka and P..K. Gupta, "Heart disease prediction systems using classification algorithms," Proceedigs of International conference on Advances in Computing and Data Sciences, vol. 1244, pp.395-404, July 2020.

[2] H. Jindal, S. Agrawal, R. Khera et al., "Heart disease prediction using machine learning algorithms," IOP Conference Series: Materials Science and Engineering, vol.1022(012072), pp.1-10, 2021.

[3] I.D.Mienye, Y. Sun, and Z. Wang Z, "An improved ensemble learning approach for the prediction of heart disease risk," Informatics in Medicine Unlocked, vol.20(100402), pp.1-5, 2020.

[4] R.Atallah and A. Al-Mous, "Heart disease detection using machine learning majority voting ensemble method," Proceedings of the 2nd International Conference on new Trends in Computing Sciences (ICTCS), pp.1-6, 2019.

[5] A.Javeed, S. Zhou, L. Yongjian et al. 4, "A. An Intelligent Learning System Based on Random Search Algorithm and Optimized Random Forest Model for Improved Heart Disease Detection," IEEE Access , vol. 27, pp.180235-180243, 2019.

[6] Y.Muhammad, M.Tahir,M. Hayat M et al., "Early and accurate detection and diagnosis of heart disease using intelligent computational model," Scientific Reports, nature research, vol 10(19747), pp.1-17, 2020.

[7] K.Dissanayake and M.G. Johar, "Comparative Study on Heart Disease Prediction Using Feature Selection Techniques on Classification Algorithms," Applied Computational Intelligence and Soft Computing, vol.5581806,pp.1-17, 2021.

[8] H.Koshimizu, H. Kojima and Y. Okuno Y, "Future possibilities for artificial intelligence in the practical management of hypertension," Hypertension Research, vol. 43(12), pp. 1327-1337, 2020.

[9] M.Rong M, D. Gong and X. Gao , "Feature Selection and Its Use in Big Data: Challenges, Methods, and Trends," IEEE Access , vol. 27, pp. 19709-19725, 2019.

[10] D.Yewale,S.P. Vijayaragavan and M. Munot M, "Decision support system for reliable prediction of heart disease prediction using machine learning techniques: an exhaustive survey and future directions," International Journal of Engineering Trends and Technology, vol. 7(4), pp. 316-331, 2022.

[11] R.C. Ripan,I.H. Sarker, M.H. Furhad, M.M. Anwarand M.M. Hoque, "An Effective Heart Disease Prediction Model Based on Machine Learning Techniques," Hybrid Intelligent Systems Advances in Intelligent Systems and Computing, preprints 2020, pp.280–288, 2020.

[12] H. Kang , "The prevention and handling of the missing data" Korean Journal of Anesthesiology, vol. 64(5), pp. 402-406, 2013.

[13] M.M.Ahsan, M.A.P. Mahmud, P.K. Saha et al., "Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance," Technologies , vol. 9(52), pp. 1-17, 2021.

[14] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," Progress in Artificial Intelligence, vol. 5, pp.221–232, 2016.

[15] I.D. Apostolopoulos, "Investigating the Synthetic Minority class Oversampling Technique (SMOTE) on an imbalanced cardiovascular disease (CVD) dataset," International Journal of Engineering Applied Sciences and Technology, vol. 4(2020), pp. 431-434, 2020.

[16] F.T.Liu, K.M. Ting and Z.H. Zhou,"Isolation-based Anomaly Detection," ACM Transactions on Knowledge Discovery from Data vol. 6(1), pp.1-39, 2012.

[17] K. Sivaraman and V. Khanna V, "Machine Learning Models for Prediction of Cardiovascular Diseases," Journal of Physics: Conference Series, 2040 012051, 2021.

[18] S. Bashir, A.A. Almazroi, S. Ashfaq et al., "A knowledge-based clinical decision support system utilizing an intelligent ensemble voting scheme for improved cardiovascular disease prediction," IEEE Access, vol. 9, pp. 130805-130822, 2021.

[19] A. Rahim, Y. Rasheed, F. Azam et al., "An Integrated Machine Learning Framework for Effective Prediction of Cardiovascular Diseases," IEEE Access, vol. 9, pp.106575-106588, 2021.

[20] M. Waqar, H. Dawood, H. Dawood et al., "An Efficient SMOTE-Based Deep Learning Model for Heart Attack Prediction," Scientific Programming, vol. 2021(6621622), pp.1-12, 2021.

[21] R. Bharti, A. Khamparia, M. Shabaz et al., "Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning," Computational Intelligence and Neuroscience, vol. 2021(8387680), pp.1-11, 2021.

[22] R. TR, U.K. Lilhore, M. Poongodi et al., "Predictive Analysis of Heart Diseases with Machine Learning Approaches," Malaysian Journal of Computer Science, vol. 1, pp.132–148, 2022.

[23] R.C. Ripan,I.H. Sarker,M.H.Furhad et al., "A Data-Driven Heart Disease Prediction Model through K-Means Clustering-Based Anomaly Detection," SN Computer Science, vol. 2(112), pp.1-12, 2021.

[24] N.L.Fitriyani,M. Syafrudin, G. Alfian et al., "HDPM: An Effective Heart Disease Prediction Model for a Clinical Decision Support System," IEEE Access, vol. 8(2020), pp.133034-133050,2020.