E-ISSN 2586-9027

Homepage : *https://tci-thaijo.org/index.php/SciTechAsia* Science & Technology Asia

Vol.30 No.1 January - March 2025 Page: [231-244]

Original research article

RFAB: The Hybrid Model for the Heart Disease Prediction

Priti Shinde*,[1], Mahesh Sanghavi[2]

[1]*Research scholar, Department of Computer Engineering, MET's IOE, Nashik 422-003, India*
[2]*Department of Computer Engineering, SNJB's KBJCOE, Chandwad 423101, India*

Received 4 July 2024; Received in revised form 15 January 2025
Accepted 11 February 2025; Available online 24 March 2025

ABSTRACT

The need for fast and accurate diagnostic systems is crucial as heart disease currently ranks as the leading cause of death worldwide. This study proposes a hybrid ensemble model called RFAB, which combines the predictions of Random Forest and Adaboost classifiers for reliable heart disease prediction. The proposed model introduces a significant improvement by utilizing an expanded dataset, increasing from 303 to 27,597 records, and applying advanced feature extraction and dimensionality reduction techniques. The testing accuracy was 95% which indicates the higher performance of the RFAB compared with SVM-81.71% and Extra Tree classifier-84%. These findings offer a non-invasive, low-cost method for early diagnosis, allowing for prompt clinical interventions and enhancing patient outcomes.

Keywords: Ensemble methods; Heart disease; Machine learning; Majority voting technique

1. Introduction

Cardiovascular disease (CVD) is a disease related to the narrowing of arteries. These arteries supply oxygenated blood to diverse parts of the body's organs. As per the report of WHO, it is the leading cause of death across the globe [1]. If arteries providing blood to the brain get blocked, it causes stroke. If the arteries, providing blood to the heart get blocked then it's called coronary artery disease. There are three main arteries in the human body that supply oxygenated blood to the heart. These are the left circumflex artery (LCX), right coronary artery (RCA), and left anterior descending artery (LAD). If any one of these arteries gets blocked due to plaque, the artery becomes stenotic. This results in a disease called coronary artery disease (CAD). This state causes a severe range of heart conditions such as angina, heart attack, and heart failure. All these terms can also be referred to individually as heart disease. It is becoming a dangerous disease

***Corresponding author:** pritivshinde25@gmail.com

doi: 10.14456/scitechasia.2025.15

and causes high mortality [2]. In developed and developing countries, for people more than 35 years old CAD is responsible for 33% of deaths [2]. Heart disease, especially coronary artery disease (CAD), represents a significant health issue in the world and makes up about one-third of all deaths in people over 35 [2]. The growing trend highlights the need for predictive models for early diagnosis to reduce the mortality rates. Conventional diagnostic approaches such as angiography are costly and invasive, which makes them inaccessible. Machine learning (ML) has emerged as a promising alternative for non-invasive detection of CAD based on patient data to make accurate predictions. Although challenges like the suboptimal performance of models on small clinical datasets have been overcome, individual classifiers such as SVM and Decision Trees still have difficulties in prediction. While techniques such as ensemble methods, which combine more types of algorithms, generally lead to better results, they usually depend on feature selection/extraction and, thus, may lose valuable information contained in the data. The medical field has evolved tremendously from a technical point of view in the last 20 years. There has been increasing growth in disease diagnosis of a patient by using newly developed devices. But these are costly, and not every person can afford it easily. Among all human chronic diseases, coronary artery disease is considered to be dangerous to mankind's life [3]. Hence the researchers have focused greatly on CAD disease diagnosis. The researchers are developing novel techniques for the diagnosis of CAD. Different surgeries and robotic technologies are involved in the diagnosis; hence, there becomes a need for non-invasive techniques for the diagnosis of disease.

In the medical field, image processing, data mining techniques, and machine learning methods have been widely applied. Image processing techniques have been applied for tumor detection and cancer cell identification, while data mining techniques are used for pattern finding and matching in input data. Machine learning techniques are used for disease detection as well as for prediction [4]. Nowadays machine learning techniques have been applied for disease prediction by many researchers. To detect CAD, many times clinicians use angiography which is an invasive technique. Hence, there is a need to develop a non-invasive technique using ML techniques. In this paper, we propose a methodology using SVM, ET, and hybrid techniques to predict heart disease. The Z Alizadeh Sani dataset, a clinical dataset containing 54 categorical features.

The dataset trains Random Forest and Adaboost classifiers using RFAB, a new hybrid ensemble (random forest combined with Adaboost classifiers) generating robust predictions using majority voting. It employs the extended Z Alizadeh Sani dataset with 27,597 records to avoid the exercise of feature selection. Representation learning method, surpassing SOTA with 95% accuracy, or get the best results. Gives stepwise reproducibility steps and details on relevance for future studies. Although previous studies have shown the effectiveness of single classifiers or feature selection techniques, they tend to compromise model accuracy and robustness. To fill the research gap, we have employed a hybrid ensemble method using the full-feature dataset to develop the solution for non-invasive and accurate prediction of heart disease. Also, we proposed a new hybrid ensemble method that integrates Random Forest and AdaBoost classifiers as weak classifiers and successfully achieved an accuracy of 95% in predicting heart disease. Compared to previous studies, the presented model uses the entire feature set of the extended Z Alizadeh Sani dataset to ensure that no important data is eliminated. Furthermore, this study tackles the problem of limited clinical datasets by increasing the number of records from 303 to 27,597, which strengthens model robustness. The method is validated via extensive benchmarking, suggesting its promise as a means of heart disease precursor detection

1.1 Proposed work

1. The dataset features are analyzed and indexed according to the type of variable, such as category, type, numerical, text, etc. The features are classified as categorical and continuous.

2. We propose a heart disease prediction system by using classifiers such as Linear SVM, Extra tree, and Hybrid methods. We have used the full feature dataset as an input to these classifiers.

3. The Hybrid methods contain the following algorithms: the Random Forest Classifier and the Adaboost Classifier. The ensemble methods give good prediction results as compared to individual classifiers. Hence, we used an ensemble.

4. Performance measures are calculated and compared with the state-of-the-art models.

The paper is divided into five sections. Section II identifies related research work done in this problem statement domain. The next section i.e. III explains the proposed methodology in detail. Experimental results and discussion of performance measures are given in section IV. Finally, the conclusion and future scope are given in section V.

2. Related Work

The most frequently used diagnostic methods for heart disease detection among clinicians are ECG tests, Echocardiograms, exercise stress tests, heart CT scans, Angio grams and Cardiac Catheterization [5]. In order to improve these diagnostic methods, machine learning (ML) techniques have been developed to predict heart disease with greater accuracy and efficiency. While significant progress has been made, this work aims to address shortcomings in feature selection, dataset handling, and model generalization.

By considering the severity of this deadly CAD disease, researchers are focusing on the precision and accuracy of the classifier. For the same, they are using feature selection methods. The hybrid PSO with extreme machine learning is used by Afzal Shahid, Maheshwari Singh et al. [6].

For instance, Shahid et al. [4] explored feature selection methods including Fisher, Relief, and minimum redundancy maximum relevance to enhance classification performance based on SVM. Although this method improved prediction accuracy (88.34%) and other measures, it was mainly based on certain feature extraction methods that might not allow generalization to datasets with different characteristics. Additionally, Nasarian et al. with SMOTE, Hua et al. [7] used a hybrid feature selection method with classifiers including XGBoost to achieve 81.23% accuracy. Nonetheless, the model's performance was largely based on a specific dataset (Z Alizadeh Sani), indicating limited focal applicability.

The ensembles are being used to predict the disease. Xiao-Yan et al. [8] proposed an ensemble by using 2 feature extraction methods and classifiers. They used LDA, and PCA for feature extraction, and KNN, SVM, DT, RF, and NB for classification. Along with these classifiers they used two ensemble methods, namely, bagging and boosting. The individual classifiers' accuracy was less as compared to ensembles.

Terrada et al. [9] employed supervised ML algorithms such as Artificial Neural Networks and Adaptive Boosting, obtaining superior prediction metrics (e.g., 94% accuracy on the Z Alizadeh Sani dataset). However, there was variance in performance on datasets such as Cleveland and Hungarian, suggesting difficulties in generalization. Similarly, Saboor et al. [9] did preprocess and hyper parameter tuning on nine classifiers, attaining 98% recall with SVM. Chif et al., (10) performed similar research, but their focus was not on class imbalance or ensemble so their results could have been better in terms of reliability.

Pathan et al. [10] used full-feature datasets with classifiers, such as Linear SVC, achieving moderate accuracy (72% and 66% on CVD and Framingham datasets, respectively) on information. While their work showed the practicality of full-feature models, their relatively lower performance highlights a need for robust models which better utilize these full feature spaces. Shahid and Singh et al. [6] focused on hybrid PSO with extreme machine learning which got accuracy 97.60% on Z Alizadeh Sani dataset. However, this effort focused on algorithmic optimization but did not tackle imbalances in datasets.

Supervised machine learning algorithms such as Artificial Neural Networks and Adaptive Boosting are used by Oumania Terrada [9]. This model gives prediction accuracy, Precision, Recall, and F1 scores as 94%, 92.58%, 97.73%, and 94.48%, respectively, on Z Alizadeh Sani dataset. They also compared their classifiers on Cleveland and Hungarian datasets. However, the Z Sani dataset achieved the highest accuracy.

Wahid et al. [11] showcase how imbalanced datasets require extra care. The reported 95.16% accuracy with Extra Trees classifier using SMOTE. They are effective; however, their dependence on certain resampling strategies may not generalize well to different datasets. Similarly, Xiao-Yan et al. [8] processed ensembles via bagging and boosting and attained 98.6% accuracy with decision trees as classifiers. Even if ensembles held strong, they relied ultimately on feature extraction methods (e.g. Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA)) which do not exploit the full potential of raw data.

Finally, Gabriel et al. [12] overcame the small number of clinical datasets by preprocessing and hyper tuning their approach. Although their model reached an accuracy of 97.7% on the CAD dataset, we believe there exists room for improvement as the authors do not further elaborate on advanced ensemble approaches nor suggest any advanced augmentation techniques in order to further boost model performance.

This study advances these progresses utilizing the complementary strengths of Random Forest and Adaboost, a full-feature dataset to prevent losing critical information, and using a majority voting method to make the predictions more robust. Moreover, in order to mitigate class imbalance and to enhance the generalization, dataset augmentation is also performed. This is a reliable and scalable heart disease prediction process that can address the shortcomings of contemporary methods.

Many times, authors have datasets other than the Z Alizadeh Sani dataset, such as Cleveland, Framingham, Statlog, and Long Beach VA. They are used as an individual or making a combination of all. However, the Z Alizadeh Sani dataset is preferable as per our proposed model. Also, many papers have not used the complete full features dataset. Instead, they have selected a feature dataset. The full feature dataset is still giving better prediction accuracy. By considering this gap we have proposed our system. There is no need that at every time one should extract features and optimize them to improve prediction accuracy. By using a full feature set one can also achieve it. Hence the contribution of this paper is to predict disease using a full feature set. While implementing the algorithms for the dataset it has been observed that, without using any feature extraction algorithm or method we are getting good performance measures for the proposed methodology.

3. Proposed Methodology

An input dataset is essential for developing and implementing any system. The dataset must be suitable for this disease prediction as the proposed methodology is implemented for coronary artery disease. There are many datasets available online, but the extended Z Alizadeh Sani dataset was found useful for disease prediction.

3.1 Dataset description

The extended Z Alizadeh Sani dataset contains a total of 58 features. The "Cath" feature indicates whether the patient has CAD. This dataset has features that give information on stenosis of coronary arteries. To predict CAD, the important feature is coronary arterial stenosis, which is given only in this dataset; hence, we have referred to this dataset. The features LAD, LCX, and RCA have been added to this dataset which are not present in any other heart disease dataset. The original dataset angiographic data served as the basis for the LAD, LCX, and RCA features. These characteristics indicate the existence of blockages in important coronary arteries, which are important markers of coronary artery disease. Clinical annotations in the dataset were used to calculate the binary values for these features (blockage present or not). The dataset is divided into four categories, namely, demographic features, electrocardiographic features, physical examination and symptoms, laboratory test and echocardiography [6]. The dataset is available on the UCI machine learning dataset repository which contains 303 records. The success of machine learning depends upon how many records are given as input to the system. If we increase the number of records for processing, the prediction results/output will be more effective. Hence in this proposed system, we have increased the number of records from 303 to 27597 by using Python functionalities such as NumPy and pandas.

Both categorical and continuous variables are present in the dataset. The features are visualized to analyze the distribution of categorical and continuous variables.

Data visualization helps to visualize and analyze/understand patterns present in data. For this we use graphs, bar charts, and scatter plots. The relationship between continuous variables can be displayed by using a scatter-plot. The dataset features are given in Table 1. As the dataset contains both categorical and continuous variables, preprocessing of the dataset is necessary. The preprocessing included the following:

1 Handling Missing Values: Missing data was imputed using median values.

2 Encoding Categorical Variables:

One-hot encoding was applied to ensure compatibility with ML algorithms.

3 Scaling Continuous Variables: Features were normalized using Min-Max scaling to ensure uniform ranges for all variables.

3.2 Methods and proposed system

The machine learning techniques were found to be useful for prediction of heart disease. The classification models are widely used for the same purpose. In this proposed method, we have used linear SVM (Support Vector Machine), an extra tree classifier, an ensemble of random forest classifier with different estimators, and an Adaboost classifier for disease prediction. By using these advantages, these methods are applied to a dataset for disease prediction. SVMs were selected due to their robustness in binary classification tasks, especially with high-dimensional data. ET was chosen due to its ability to efficiently deal with both categorical and continuous variables and its lower variance than traditional decision trees. Also, the hybrid RFAB model integrates the advantages of Random Forest (resistant of over fitting, handling imbalanced data) and Adaboost (revising misclassified samples in an iterative way), which matches with study characteristics of aiming at gaining high accuracy with the least preprocessing. CART and NB were excluded due to their comparatively weaker performance on similar data structures, and LR was less preferred due to its lack of ability to identify complex, non-linear relationships in the data. There is a need to implement an ensemble learning approach for the improvement of the accuracy of the classifier towards disease prediction [13]. The system can be implemented by using ensembles like Adaboost, bagging, boosting, etc. [14]. We used synthetic data generation techniques and algorithms implemented in Python libraries NumPy and pandas to augment the dataset. This involved oversampling and the generation of synthetic records to leverage the statistical distributions of existing features in order that variability remained realistic. This augmentation approach helped to mitigate the issue of small clinical datasets, allowing for improved model training and decreased susceptibility to over-fitting. We then extensively tested the effect of augmentation on model performance to validate it and make sure it is reliable.

3.3 Machine learning models used

3.3.1 Support Vector Machine (SVM)

SVMs can do classification and regression. They are based on the concept of hyperplanes, a boundary that z data is not going to be separated linearly, then the Kernel trick is used. In such cases, the linear kernel in the form of a mathematical expression is represented $K(\overrightarrow{x_i} \cdot \overrightarrow{x_j}) = \overrightarrow{x_i} \cdot \overrightarrow{x_j}$ where x is a feature and i and j represent two-dimensional space. Hence, an SVM is effective in high-dimensional spaces, i.e. it performs well in high-dimensional feature sets. Once the optimal hyperplane is determined in the training set, the SVM can be used for new, unseen data features [15].

The Extra tree classifier, a small name for an Extremely Randomized Tree classifier, is an ensemble learning technique used for classification purposes. It belongs to the family of Decision tree classifiers, as does Random Forest [16]. It is different from a random forest in sense that it introduces randomness while developing the tree. It selects features randomly at each

Table 1. Ex-Z Alizadeh Sani dataset [6].

Feature type	Feature name	Range
Demographic	Age	30-86
	Sex	M, F
	Weight	48-120
	BMI (kg/m^2)	18-41
	DM	Yes, No
	HTN	Yes, No
	Current smoker	Yes, No
	Ex Smoker	Yes, No
	Family History	Yes, No
	Obesity (BMI>25)	Yes, No
	Chronic Renal Failure (CRF)	Yes, No
	Cerebrovascular accident	Yes, No
	Thyroid disease	Yes, No
	Airway disease	Yes, No
	Congestive Heart Failure	Yes, No
	Dyslipidemia	Yes, No
ECG	ST elevation	Yes, No
	ST depression	Yes, No
	Q Wave	Yes, No
	T inversion	Yes, No
	LVH	Yes, No
	Poor R wave progression	Yes, No
Symptoms & examinations	BP (mm-Hg)	90-190
	Pulse rate (PR)	50-110
	Edema	Yes, No
	Weak peripheral pulse	Yes, No
	Lung rales	Yes, No
	Systolic murmur	Yes, No
	Diastolic murmur	Yes, No
	Typical chest pain	Yes, No
	Dyspnea	Yes, No
	Functional class	1,2,3,4
	Atypical	Yes, No
	Nonanginal chest pain	Yes, No
	Exertional chest pain	Yes, No
	Low threshold angina	Yes, No
Lab test & Echocardiography	Fasting blood sugar (mg/dL)	62-400
	Creatinine (mg/dL)	0.5-2.2
	Triglyceride (mg/dL)	37-1050
	LDL (mg/dL)	18-235
	HDL (mg/dL)	15-111
	Sodium (mEg/lit)	128-156
	WBC (cells/mL)	3700-18000
	Lymphosite (%)	7-60
	Neutrophil (%)	32-89
	Platelet (1000/mL)	25-742
	Blood urea nitrogen (BUN) (mg/dL)	6-52
	ESR (mm/h)	1-90
	HB (g/dL)	8.9-17.6
	Potassium (mEg/lit)	3.0-6.6

split. This randomness introduces more diversity at each split. This feature makes Extra Tree different from other ensemble families. This feature hence helps to reduce the variance of the model compared to traditional decision trees and random forests.

In Saboor et al. [9], the higher accuracies that they reported were found after large feature selection and hyper parameter tuning on a subset of those features. On the other hand, our approach employs the full-feature extended Z Alizadeh Sani dataset despite the absence of feature selection, thus being more generalizable and applicable for clinical use. Overall, our hybrid method achieves a classification accuracy of 95%, which confirms that ensemble learning provides robust predictions over other classifiers.

3.3.2 Extra tree classifier

The Extra tree classifier, a small name for an Extremely Randomized Tree classifier, is an ensemble learning technique used for classification purposes. It belongs to the family of Decision tree classifiers, as does Random Forest [16]. It is different from a random forest in sense that it introduces randomness while developing the tree. It selects features randomly at each split.

This randomness introduces more diversity at each split. This feature makes Extra Tree different from other ensemble families. This feature hence helps to reduce the variance of the model compared to traditional decision trees and random forests.

In Saboor et al. [9], the higher accuracies that they reported were found after large feature selection and hyper parameter tuning on a subset of those features. On the other hand, our approach employs the full-feature extended Z Alizadeh Sani dataset despite the absence of feature selection, thus being more generalizable and applicable for clinical use. Overall, our hybrid method achieves a classification accuracy of 95%, which confirms that ensemble learning provides robust predictions over other classifiers.

3.3.3 Hybrid classifier

The evolution of ML techniques for CAD detection has progressed significantly. Early works, such as those by Afzal Shahid, utilized SVM with feature selection methods, achieving accuracies of ~ 88.34%. Subsequent studies integrated ensemble methods like XGBoost and SMOTE to handle class imbalances, achieving higher performance (e.g., 94.7% by Zhang et al. [17]).While these models showed promise, many relied on feature extraction or dataset-specific optimizations, limiting generalizability. Recent works have explored hybrid ensembles, yet few have used the full-feature dataset approach. Our study addresses this gap by combining RF and Adaboost, leveraging their complementary strengths to enhance accuracy without feature extraction.

We have proposed the flow of the system as shown in Fig. 1. The proposed system aims to predict heart disease early and will be helpful for doctors as clinical assistants. As the chosen dataset for the system contains data about blocking arteries, it is useful for early disease prediction. These features are namely LAD (Left anterior descending artery), LCX (Left circumflex artery), and RCA (Right coronary artery) indicating whether they have a blockage or not. If any of these arteries have blockages, then the patient develops angina and then a heart attack. Hence in the early stage of angina developed patient, one can predict heart disease.

This extended Z Alizadeh Sani dataset contains these extra features (LCX,

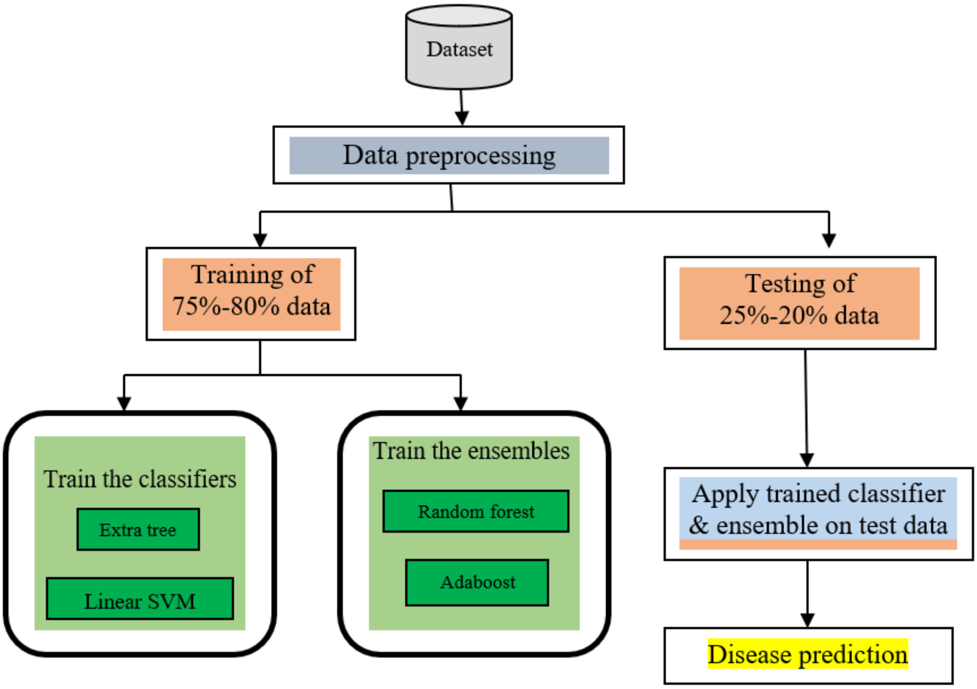

Fig. 1. Proposed system for heart disease prediction.

LCA, RCA); hence, we have chosen this dataset for disease prediction. We have used all features of the dataset for disease prediction. The proposed system is divided into three phases, the first phase is data preprocessing which includes finding missing values, data cleaning, and data readiness for the classifier; the second phase is applying the classifier to the dataset; and the final phase contains a prediction of disease along with calculating the performance of system. A representation of the hybrid model's work-flow can be added for clarity.

Algorithm: Hybrid Ensemble Classifier for Heart Disease Prediction
Input: Z Alizadeh Sani dataset with 58 features. Data Preprocessing: Handle missing values and clean data. Expand dataset records using Python libraries (e.g., NumPy, pandas).
Classifier Training: Train Random Forest with optimized hyper parameters. Train AdaBoost with the same dataset.
Combine Random Forest and AdaBoost using majority voting for final prediction.
Performance Evaluation: Calculate metrics: Accuracy, Precision, Recall, F1-score, ROC-AUC.
Output: Predicted class (heart disease: Yes/No).

To increase prediction accuracy, the method combines the complementary advantages of Adaboost and Random Forest. It prevents the possible loss of important information by using a full-feature dataset without feature extraction, guaranteeing that all pertinent data contribute to the model's performance. As a reliable technique for final predictions, majority voting improves reliability over conventional ensemble methods. To improve generalization and performance in a variety of scenarios, dataset augmentation is also used to increase the training data and correct imbalances.

4. Experimental Results & Discussion

4.1 Set-up

The experiments have been done in Python using Google Collab on HP, Intel(R) Core (TM) i3-6006U CPU @ 2.00GHz with 8 GB RAM, Windows 10 64-bit OS.

4.2 Results of classifiers

The Z Alizadeh Sani dataset has 303 samples and 58 features consisting of demographic, clinical, and laboratory data. With the help of Python libraries, such as NumPy and pandas, we expanded the dataset to 27,597 records by synthetic data generation via the oversampling approach. The idea is to train and test a model in a robust manner.

We use ensemble techniques to propose hybrid Random Forest and AdaBoost algorithms. We presented a hybrid classifier having majority voting, trained with various classifiers like Standalone classifiers (SVM, Extra Tree Classifier). Unlike many approaches that reduce features or select features, the new method uses all 58 features of the extended Z Alizadeh Sani dataset. This method preserves important frequencies that are relevant for heart disease prediction. Using features including stenos of coronary arteries (LAD, LCX, RCA) to predict conditions like angina that are precursors to heart attacks, the study focuses on the early detection of heart disease.

In this study, the SVM, the Extra tree, and Hybrid classifiers are used. The comparative analysis of these algorithms was carried out in the experiments. We compared these algorithms on the extended Z Alizadeh Sani dataset. 75%-80% and 25%-20% of data have been taken for training and testing respectively. We have experimented on a full feature dataset i.e. on 58 features. By using full features, we are getting good performance measures. For training we are getting accuracies of 81.6%, 86.0%, 96%, and for testing 81.71%, 84%, and 95% for SVM, Extra tree, and Hybrid classifiers respectively.

Table 2. Confusion Matrix for SVM.

		Predicted Values	
		No	Yes
Actual	No	1796	0
Values	Yes	1262	3842

Table 3. Confusion Matrix for extra tree classifier.

		Predicted Values	
		No	Yes
Actual	No	1209	228
Values	Yes	655	3428

Table 4. Confusion Matrix for hybrid classifier.

		Predicted Values	
		No	Yes
Actual	No	1681	74
Values	Yes	271	4874

4.3 Experimental results with support vector machine, extra tree classifier & hybrid classifier

We have calculated different performance measures such as Confusion matrix, Accuracy, F1 score, Precision, and Recall to validate our system. Table 2 shows the Confusion Matrix for SVM; Table 3 shows the Confusion Matrix for Extra Tree Classifier; and Table 4 shows the Confusion Matrix for Hybrid Classifier.

The Confusion Matrix is calculated for the machine learning techniques. The

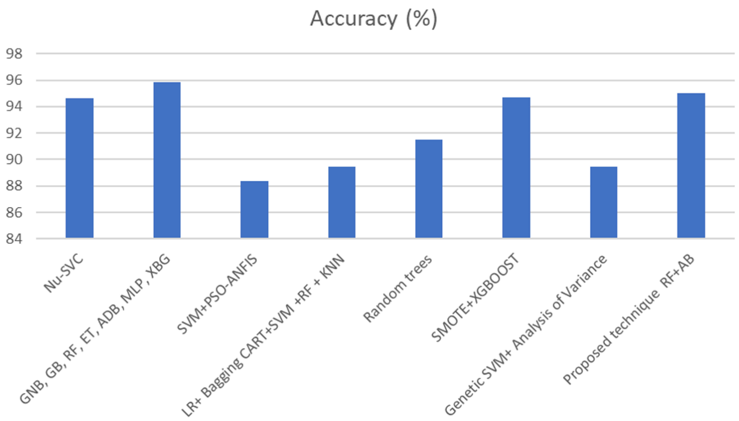

Fig. 2. Comparison between proposed technique & other related work.

Table 5. Evaluation metrics for implemented ML techniques.

Ref.	Authors	Techniques/Methods	Accuracy %
[18]	Abdar, Moloud, et al	Nu-SVC	94.66
[19]	Wang, Jikuo, et al.	GNB, GB, RF, ET, ADB, MLP, XBG	95.84
[20]	Shahid, Afzal Hussain, and M. P. Singh	SVM+PSO-ANFIS	88.34
[21]	Dahal, Keshab R et.al.	LR+ Bagging CART+SVM +RF + KNN	89.47
[15]	Joloudari, Javad Hassannataj, et al	Random trees	91.47
[17]	Zhang, Shasha, et al	SMOTE+XGBOOST	94.7
[5]	Hassannataj Joloudari, Javad, et al.	Genetic SVM+ Analysis of Variance	89.45
	Proposed system	**Hybrid technique using Random Forest + Adaboost**	**95**

parameters of the Confusion Matrix are TP, TN, FP, and FN. TP occurs when the model accurately and correctly predicts the occurrence of disease. TN occurs when the model accurately and correctly predicts the absence of disease. FP occurs when the model incorrectly classifies the occurrence of disease. And finally, FN is defined as when the model incorrectly classifies the absence of disease. The analysis of the model is done on a testing dataset by calculating the model's accuracy, precision, recall, F-measure, and ROC-AUC values. Precision gives the reliability of a model in predicting the correct class, i.e. patient will have a disease. The recall gives the proportion of correct prediction of positive cases to the total number of positives cases. The performance measures are given as below for our model.

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN}, \tag{4.1}$$

$$\text{Precision} = \frac{TP}{TP+FP}, \tag{4.2}$$

$$\text{Recall} = \frac{TP+TN}{TP+FN}, \tag{4.3}$$

$$\text{F-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{4.4}$$

The evaluation metrics for all the supervised classifiers are given in Table 5.

4.4 Comparative analysis with previous studies and discussion

The comparative analysis given in Table 5 for disease prediction indicates that there is a significant difference in the proposed system and the existing system. These differences are visualized in Fig. 2

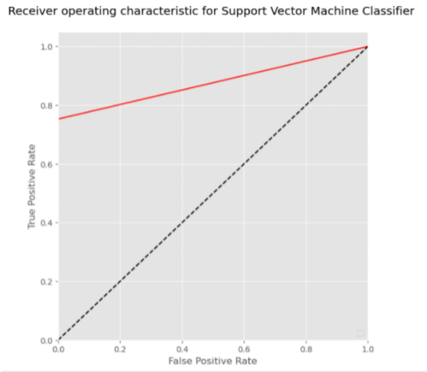

Fig. 3. ROC curve for SVM classifier.

which identifies that the proposed system has achieved maximum accuracy for disease prediction. In this proposed technique, we have used a full-featured dataset for the prediction of heart disease. The misclassified samples were examined for trends. We found borderline cases only, with their clinical profile very similar to CAD patients. False negatives were ascribed to patterns not well represented in the original dataset. Compared to existing methods, the RFAB hybrid model yields improved results (see Table 6), which suggests that the mixture of general and specific features in full-feature datasets is beneficial when using ensemble learning.

Although Wang et al. [15] shows a slight increment in accuracy, stack-based ensembles can be resource-heavy. However, our hybrid model simplifies the network structure and achieves similar performance with high efficiency, which allows for its deployment in real applications.

4.5 Performance analysis with AUC for SVM, Extra Tree classifier, Hybrid classifier

Figs. 3-5 shows the ROC-AUC curves for the experiments conducted on exz sani dataset.

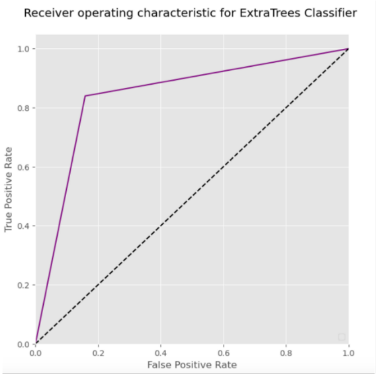

Fig. 4. ROC curve extra tree classifier.

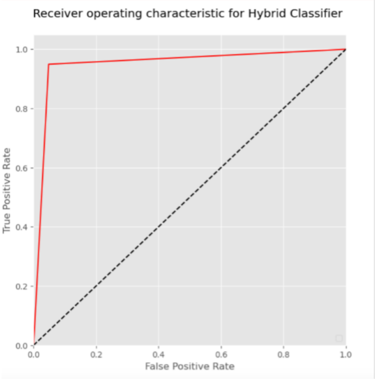

Fig. 5. ROC curve for hybrid classifier.

5. Conclusion

This paper presents a novel hybrid ensemble method that combines different estimators, such as Random Forest and AdaBoost, to predict heart disease. The experiments were conducted on a comprehensive dataset sourced from the UCI machine learning repository, and included machine learning techniques such as SVM, Extra Tree, Decision Tree, and the proposed Hybrid algorithm. The hybrid method achieved an accuracy of 95%, outperforming the SVM (81.71%) and Extra Tree (84%) models. A key insight from the system implementation was the importance of feature extraction and dimensionality reduction, which allowed for the expansion of

Table 6. Performance comparison of proposed method with existing related studies on Z Sani dataset for heart disease prediction.

Evaluation metrics **ML techniques**	**Training Accuracy %**	**Testing Accuracy %**	**F measure %**	**Precision %**	**Recall %**
SVM	81.6	81.71	0.81	0.86	0.82
Extra tree	86	84	0.83	0.84	0.84
Hybrid classifier	96	95	0.95	0.95	0.95

the dataset from 303 to 27,597 records using a Python function. This significant increase in data volume contributed to the enhanced performance of the hybrid model.

For future research, real-time data could be integrated to predict heart disease dynamically, further enhancing the practicality of the model. Furthermore, the suggested hybrid approach may be expanded to more difficult tasks, like artery stenosis detection, opening up new clinical application opportunities.

Acknowledgement

The authors would like to thank MET's Institute of Engineering, Bhujbal Knowledge City, and SNJB's KBJCOE for their support in this research, and thank the UCI Machine Learning Repository for hosting the dataset.

References

[1] World Health Organization (WHO). Cardiovascular Diseases. https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1

[2] Kandaswamy E, Li Z. Recent advances in treatment of coronary artery disease: role of science and technology. International journal of molecular sciences. 2018; 19(2): 424.

[3] Shorewala V. Early detection of coronary heart disease using ensemble techniques. Informatics in Medicine Unlocked. 2021; 26: 100655.

[4] Shahid AH, Singh MP. A novel approach for coronary artery disease diagnosis using hybrid particle swarm optimization based emotional neural network. Biocybernetics and Biomedical Engineering. 2020; 40(4): 1568-85.

[5] [28] Mayo Clinic Staff. Coronary Artery Diseases-Diagnosis and Treatment. https://www.mayoclinic.org/diseases-conditions/coronary-artery-isease/diagnosis-treatment/drc-20350619

[6] Alizadehsani R, et al., A data mining approach for diagnosis of coronary artery disease. Computer Methods and Programs in Biomedicine. 2013; 111(1):52-61.

[7] Nasarian E, et al. Association between work-related features and coronary artery disease: A heterogeneous hybrid feature selection integrated with balancing approach. Pattern Recognition Letters. 2020; 133: 33-40.

[8] Gao XY, et al. Improving the accuracy for analyzing heart diseases prediction based on the ensemble method. Complexity. 2021; 2021: 1-10.

[9] Saboor A, et al. A method for improving prediction of human heart disease using machine learning algorithms. Mobile Information Systems 2022 (2022).

[10] Pathan MS, et al. Analyzing the impact of feature selection on the accuracy of heart disease prediction. Healthcare Analytics. 2022; 2: 100060.

[11] Wahid N, et al. Detection of Coronary Artery Disease Using Extra Tree Categorization.

[12] Gabriel JJ, Anbarasi L J. Optimizing Coronary Artery Disease Diagnosis: A Heuristic Approach using Robust Data Preprocessing and Automated Hyperparameter Tuning of eXtreme Gradient Boosting. IEEE Access (2023).

[13] Swain D, et al. An efficient heart disease prediction system using machine learning. Machine Learning and Information Processing: Proceedings of ICMLIP 2019. Springer Singapore, 2020.

[14] Bashir S, et al. A knowledge-based clinical decision support system utilizing an intelligent ensemble voting scheme for improved cardiovascular disease prediction. IEEE Access. 2021; 9: 130805-22.

[15] Wang H, et al. Research survey on support vector machine. 10th EAI International Conference on Mobile Multimedia Communications. 2017.

[16] Geurts P, Damien E, Louis W. Extremely randomized trees. Machine learning. 2006; 63: 3-42.

[17] Zhang S, et al. Improvement of the performance of models for predicting coronary artery disease based on XGBoost algorithm and feature processing technology. Electronics. 2022; 11(3): 315.

[18] Hassannataj JJ, et al. GSVMA: a genetic support vector machine ANOVA method for CAD diagnosis. Frontiers in cardiovascular medicine. 2022; 8: 760178.

[19] Chandramouli S, Saikat D, Amita D. Machine learning: Pearson Education India; 2018.

[20] Terrada O, et al. Prediction of patients with heart disease using artificial neural network and adaptive boosting techniques. 2020 3rd International Conference on Advanced Communication Technologies and Networking (CommNet): IEEE; 2020.

[21] Shahid AH, et al. Coronary artery disease diagnosis using feature selection based hybrid extreme learning machine. 2020 3rd International Conference on Information and Computer Technologies (ICICT): IEEE; 2020.