Machine Learning Model to Predict Heart Disease and Heart Failure Mortality

Md. Naimur Rahman[1*], Fahmida Rahman Liza[2], Md. Shafak Shariar Sozol[3], S. M. Taohidul Islam[4],Md. Maniruzzaman[5]

[1]Dept. of Electrical and Electronics Engineering, Patuakhali Science and Technology University, Patuakhali.
E-mail: naimur.cse4th@pstu.ac.bd
[2]Faculty of Computer Science and Engineering, Patuakhali Science and Technology University, Patuakhali.
E-mail: liza14@cse.pstu.ac.bd
[3]Faculty of Computer Science and Engineering, Bangladesh University of Engineering and Technology.
E-mail: shahriarsajal05405@gmail.com, staohidul@gmail.com
[4]Electronics and Communication Engineering (ECE) Discipline, Khulna University.
E-mail: m-m-zaman@hotmail.com
[*]Correspondence: E-mail: naimur.cse4th@pstu.ac.bd,

Abstract: The heart performs an important role in the human body. A healthy heart is required for living a healthy life. But heart disease is a thorn in the path of living a healthy life and eventually leads to heart failure. Heart failure is a chronic condition that becomes worse over time. People with heart illness or who have a high risk of cardiovascular disease need to be detected early and managed, in that case, a machine learning model can be very helpful. This paper introduces two distinct systems to predict heart failure mortality. The first system predicts a patient who has cardiovascular disease or not and the second system can predict heart failure mortality. For the prediction, five supervised machine learning algorithms namely Extreme gradient Boost (XGB), Random Forest (RF), K Nearest Neighbor (KNN), Ada Boost (Ada) and an ensemble approach using Stacking CV Classifier has been used to compare their results using different classification metrics which can find the most effective model among them. According to the findings of this study, the ensembles algorithm is the most effective algorithm to predict heart disease and heart failure mortality, with an accuracy score of 91% and 93% respectively.

Keywords: Heart Disease, Heart Failure, Extreme Gradient Boost, K Nearest Neighbor, Random Forest, Ada Boost, Ensemble approach, Prediction.

Introduction

Heart disease or Cardio Vascular Disease the major cause of mortality worldwide, has long been a significant threat to public health, wreaking havoc on individuals, communities, and countries. As per the World Health Organization, 17.9 million people die each year as a consequence of cardiovascular disease, with 80% of such damage being between coronary artery disease and cerebral stroke [1]. Cardiovascular disease has been escalating at an alarming rate. That's why researchers have long been working in this field to create a fast, accurate, and reliable system that can predict heart disease and heart failure at an early phase using a variety of Machine Learning and data mining techniques. The overall number of heart failure patients continues to climb today as a result of a rising population. The frequency variety of heart failure, on the other hand, appears to be changing. These results, along with a slower-than-expected reduction in heart failure mortality, indicate that the epidemic is far from over. Despite the fact that heart disease has become the top cause of death globally in recent decades, it is also one of the diseases that can be successfully treated and maintained. The

correct moment of identification of a disease determines the entire effectiveness of illness care. Machine learning is frequently used in illness categorization. Smart electronic health records, drug development, biological signal processing, and illness detection and diagnosis are the core machine learning aspects in healthcare. Devansh Shah, et al. proposed a system that can predict the likelihood of having heart disease. They used five different machine learning algorithms namely Naïve Bayes, decision tree, K-nearest neighbor, and random forest. K-nearest neighbor provides the highest accuracy among them [2]. The high number of fatalities is prevalent in poor and middle-income countries [3]. An unhealthy lifestyle, such as smoking cigarettes, alcohol and eating too much fat can contribute to high blood pressure, which can lead to heart disease. The capacity to identify heart illness promptly, precisely, and properly is crucial in avoiding mortality. Heart failure was first identified as an oncoming epidemic around 25 years ago. It affects an estimated 64.3 million individuals globally. The prevalence of recognized heart failure in the general adult population in developed nations is believed to be between 1% and 2% [4]. Apurb Rajdhan, et al. used different techniques of data mining such as Decision Tree, Naive Bayes, Random Forest, and Logistic Regression for predicting the probability of having a heart disease and classified the risk label of the patients [5]. Archana Singh, et al. used various classification algorithm for predicting heart disease and then calculate their accuracy. They used decision tree, linear regression, k-neighbor and SVM for creating the model and on the basis of accuracy they conclude which model is the best among them [6]. Berina Alic, et al. performed a comparative analysis of 20 different research studies. They presented an overview of different techniques of machine learning for the classification of cardiovascular disease and diabetes where they have shown ANN and Naive Bayes are best to predict cardiovascular disease and diabetes [7]. Aditi Gavhane, et al. proposed a system to develop an application by which the vulnerability of a cardiovascular disease can be predicted [8]. Noor Basha, et al. used different classification algorithms such as Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Decision tree, Random Forest, and Naïve Bayes to predict and analyze the syndrome of heart [9]. Rubini PE, et al. presented a comparative analysis in cardiovascular diseases categorization. They used diverse machine learning approaches such as Logistic Regression, Random Forest, Support Vector Machine, and Naïve Bayes [10].

This study seeks to anticipate these cardiac diseases at an early stage in order to avoid catastrophic consequences such as heart failure and lower the heart failure death rate. The goal is to compare various machine learning approaches and choose the optimal choice for getting the maximum classification output accuracy.

Methodology

Data Source

For this research, two distinct datasets have been used in which the first one predicts whether a person has heart disease or not, whereas the other predicts mortality by heart failure. Dataset I is obtained from the open-source dataset platform UCI and is used to forecast heart disease [11]. Although there are 76 features in this dataset, most of the published studies only use a subset of 14 of them. In fact, so far, the only database that has been used by machine learning researchers is the Cleveland database. Previously, this dataset was utilized in numerous research, such as in the study of [12], where they used it to predict heart disease. The Dataset II is similarly obtained from the open-source dataset platform UCI which is used to forecast

mortality by heart failure [13]. The actual dataset was collected from the Faisalabad Institute of Cardiology and the Allied Hospital in Faisalabad (Punjab, Pakistan) which included 299 heart failure patients [14].

Data Description

The first Dataset I is used for predicting heart disease which has 303 instances and 14 attributes in total where 13 attributes are related to heart disease.

Table 1. Attributes detail of heart disease dataset

Attributes	Details	Type
Age	Age of the patients in year	Continuous
Sex	Sex (1 = male; 0 = female)	Binary
Cp	Chest Pain type 1: typical angina 2: atypical angina 3: non- anginal pain 4: asymptomatic	Categorical
Trest bps	resting blood pressure (in mm Hg on admission to the hospital)	Continuous
Chol	serum cholesterol in mg/dl	Continuous
Attributes	Details	Type
Fbs	(Fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)	Boolean
restecg	resting electrocardiographic results 0: normal 1: having ST-T wave abnormality 2: showing probable or definite left ventricular hypertrophy by Estes' criteria	Categorical
thalach	maximum heart rate achieved	Continuous
exang	exercise induced angina (1 = yes; 0 = no)	Binary
oldpeak	ST depression induced by exercise relative to rest	Continuous
Slope	the slope of the peak exercise ST segment	Categorical
Ca	number of major vessels (0-3) colored by fluoroscopy	Categorical
Thal	3 = normal; 6 = fixed defect; 7 = reversible defect	Categorical
target	Heart Disease or not (1 or 0) (the predicted attribute)	Boolean

A comprehensive explanation of the 14 characteristics used in the proposed study is shown in Table 1. The second Dataset II is used to forecast mortality by heart failure has 13 attributes and 299 instances in total. Out of them, 194 were men and 105 were women. Every one of them was over 40 years old and had systolic dysfunction in the left ventricle. Out of 13, 12 attributes have been related to heart failure like platelets, serum creatinine, serum sodium, etc. A detailed overview of 13 features utilized in this study is shown in Table 2.

Table 2. Attributes detail of heart failure dataset

Attributes	Details	Type
Age	Age of the patients in year	Continuous
anemia	Decrease of red blood cells or hemoglobin	Boolean
high blood pressure	Patient has hypertension or not	Boolean
creatinine phosphokinase (CPK)	Level of the CPK enzyme in the blood (mcg/L)	Continuous
diabetes	The patient has diabetes or not	Boolean
ejection fraction	percentage of blood leaving the heart at each contraction	Continuous
platelets	Number of platelets in the blood (kilo platelets/mL)	Continuous
Sex	woman or man	Binary
serum creatinine	level of serum creatinine in the blood (mg/dL)	Continuous
serum sodium	level of serum sodium in the blood	Continuous
smoking	if the patient smokes or not	Boolean
Time	follow-up period (days)	Continuous
DEATH_EVENT	if the patient deceased during the follow-up period	Boolean

Data pre-processing

The modifications applied to a dataset before feeding it to the algorithm are referred to as pre-processing. Data Preprocessing is a method for transforming raw data into a tidy collection of data. To put it another way, every time data is acquired from various sources, it is collected in raw format, which makes analysis impossible. In order to achieve better performance from the model, the data needs to be in a proper format. All the categorical features have been transformed into dummy variables of the Dataset I. Several scaling techniques including Min Max Scaler and Standard Scaler have been applied to dataset. All the features of the dataset are significant for medical diagnosis of heart disease and differs from patient to patient. That's why all the features have been used to train the model for predictive analysis of heart disease. A correlation-based feature selection approach has been used for the Dataset II. Out of the 12 features, 5 features have been chosen to train of the model:

- Time
- ejection_fraction
- serum_sodium
- serum_creatinine
- age

Figure 1 depicts the correlation coefficient of the Dataset II using a heatmap. All of the features have been chosen that have an absolute value of correlation with the target variable greater than 0.18.

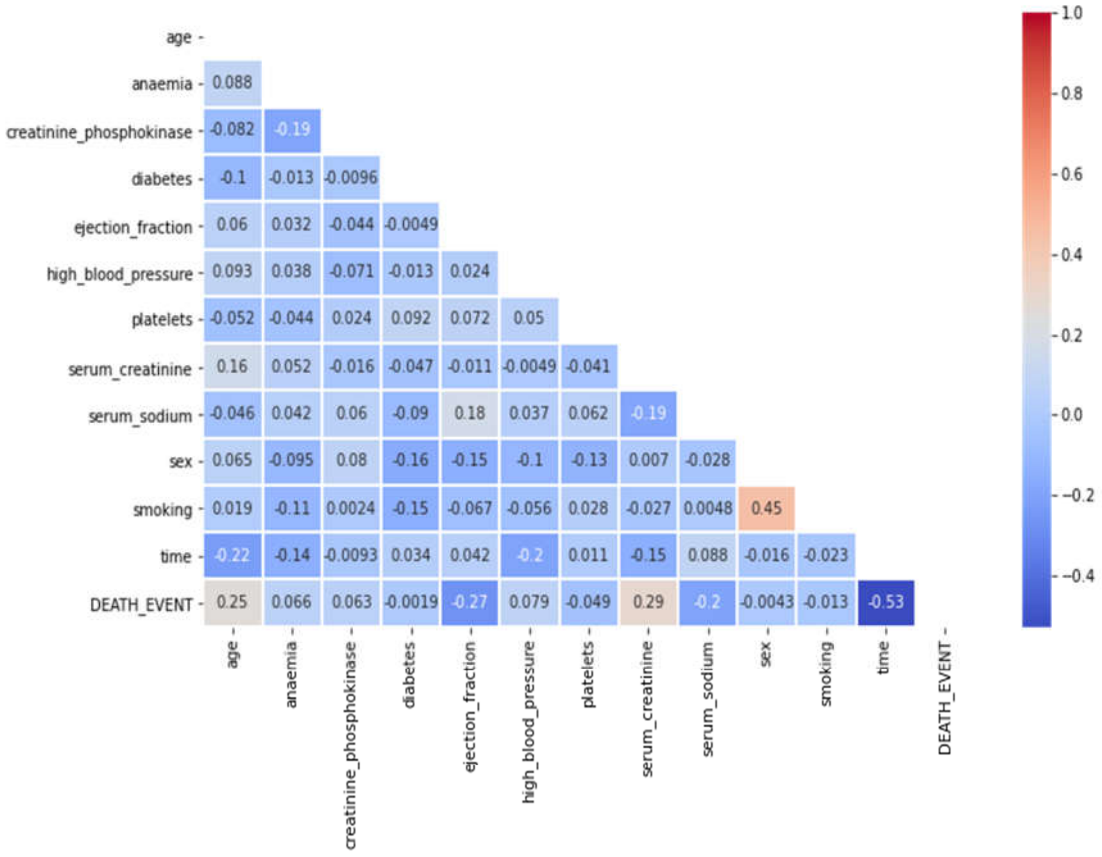

Figure 1. Heatmap of the feature correlation

Proposed Models

The proposed model predicts two different stages of heart disease such as early stage of heart disease and heart failure. To do this, four supervised machine learning algorithms have been used and then an ensemble approach has been performed using these algorithms. The flowchart in Figure 2 shows the datasets were entered as input and then processed in the second phase. Finally, methods and evaluation metrics were applied in two respective datasets. These algorithms are Random Forest, XGB, KNN, and AdaBoost. For the ensemble approach, Random Forest, XGB, and KNN are combined. Their outcomes are compared based on a number of distinct evaluation criteria to find out which of them performed best for the early prediction of heart disease and heart failure mortality.

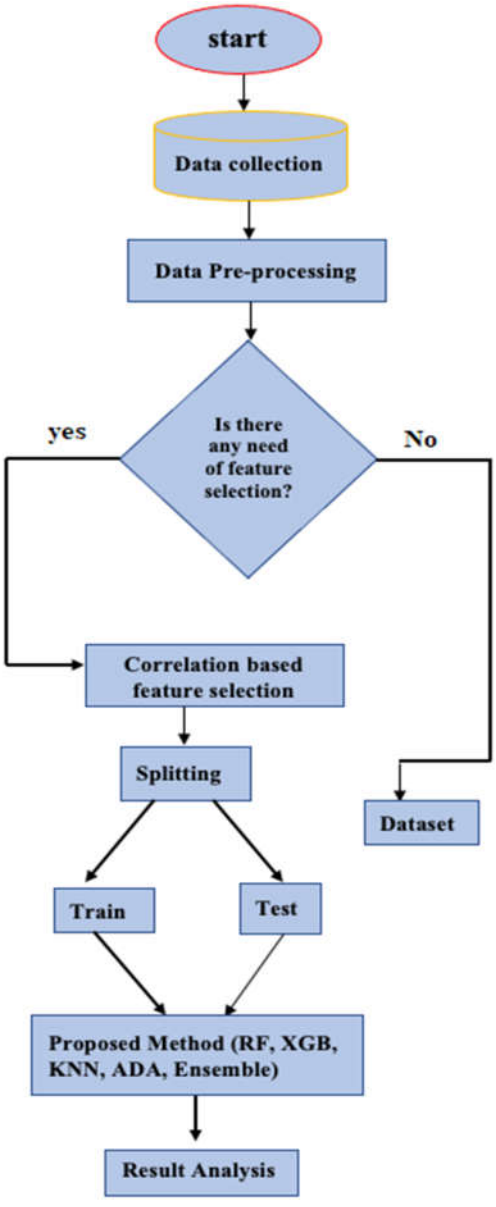

Figure 2. Workflow of the study

Random Forest

Random Forest is a machine learning approach which is used to tackle regression and classification issues using a random forest algorithm. In the case of categorization, the random forest produces the class picked by the majority of trees as the outcome. As a result of a reduction in generalization error and the likelihood of reaching that node, feature significance is computed [15].

$$ni_j = w_j C_j - w_{left(j)} C_{left(j)} - w_{right(j)} C_{right(j)} \quad (1)$$

Symbols are arranged as follows:
1. Node of significance (ni)
2. Number of samples reaching node j that are weighted (w),
3. When node j has an impurities score (G),
4. Child node,
5. Right split child

XGB-Boost

The gradient boosted decision trees have been optimized for power and stability using the XGBoost algorithm. For several advanced analytics issues, XGBoost represents a logical tree-boosting method commonly referred as GBDT or GBM. It is formed by three different models Including the learning rate in the Gradient Boosting Algorithm. It is a tree-based algorithm for classification and regression tasks and ranking with custom loss functions. It can control over fitting and give better performance. XGBoost employs the Taylor series to estimate the loss function value for a base learner [16].

$$F = \{f_1, f_2, f_3 \dots \dots \dots \dots \dots \dots \dots \dots . f_n\} \quad (2)$$

F is the set of base learners

Final Prediction, $y_i = \sum_{t=1}^{n} f_t(X_i)$

Taylor Series:

$$f(a+h) = f(a) + f'(a)h + \frac{1}{2}(f''(a)h^2 + \dots f^n(a)\frac{h^n}{n!} \quad (3)$$

XGBoost uses this loss function for calculating loss:

$$r_{im} = -\alpha[\frac{\partial(L(v_i, F(x_i))}{\partial F(x_i)}]_{F(x)=F_{m-1}(x)'} \quad (4)$$

New Prediction = Prediction of previous step + learning rate*output

KNN

The KNN approach is a Supervised Machine learning strategy for using in both classification and regression problems. It forecasts the importance for the new Data set using the K Nearest points. It saves all existing data and classifies new data points based on similarities. The algorithm uses different metrics for calculating distance. Euclidean distance and Manhattan distance are the widely used metrics for distance calculation [17].

Euclidian distance:

$$d = \sqrt{[(x_2 - x_1)^2 + (y_2 - y_1)^2]} \quad (5)$$

where (x1, y1) and (x2, y2) are data points

Manhattan distance:

$$d = |x_1 - x_2| + |y_1 - y_2| \quad (6)$$

where (x1, y1) and (x2, y2) are data points

G. ***Ada Boost***

This useful statistical format was developed by Yoav Freund and Robert Schapire in 2003. In order to increase performance, it may be used in concert with different learning methods. The outcome of any of the machine learning is combined into a weighting factor that reflects the final product of the boosted classifier. AdaBoost is flexible in the idea that potential weak learners are tinkered with in order to reward situations classified wrongly by past classifiers. Main principle of this algorithm is to assign the classifier weights and train data for each iteration so that predictions of unexpected observations can be made. Initial weights assigned

by using $\frac{1}{N}$, where N is the total number of records [18]. Final decision has been given using the following equation:

$$\alpha_t = \frac{1}{2} \ln \frac{(1-\text{Total Error})}{\text{Total Error}} \quad (7)$$

Where, α is the impact of a specific stump in the final decision. Total error can be defined by the number of incorrectly classified data. After each iteration, the weights are updated by:

$$w_i = w_{i-1} * e^{\pm\alpha} \quad (8)$$

Ensemble

Ensemble methods are strategies for developing multiple models and then combining them to generate better results. Ensemble techniques often yield more accurate results than a single model would. Bagging, boosting and stacking are the three basic types of ensembles learning methods. Ensemble techniques, on the other hand, take a variety of modeling techniques into consideration and mean them to generate a best solution. StackingCVClassifier has been used to stack Random Forest, XGBoost and K-Nearest Neighbor. Stacking usually produces better results than any individual learning algorithm. Stacking combines different models using another machine learning algorithm. The main concept is to train machine learning algorithms on training datasets and then use these models to generate new datasets. The new dataset is then sent into the combined machine learning algorithm.

Performance metrices

As part of this study, cutting-edge performance assessment measures are utilized to assist analyze the findings of a study. They are sensitivity, specificity, precision, accuracy, Brier Score statistics [19].

Sensitivity: Determine what rate of positive cases the classifier properly classifies as affirmative.

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

Specificity: Determine the percentage of all unfavorable occurrences that the classification properly classifies as negatively.

$$\text{Specificity: } \frac{TN}{TN+FP} \quad (10)$$

Precision: Determine which percentage of positive forecasts are true.

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

Accuracy: Divide the total amount of properly categorized observations by the entire input features.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FN+FP} \quad (12)$$

Brier Score: The mean squared error between predicted probability and simulated results is determined utilizing Brier score. The Brier score seems to be in the range of 0.0 to 1.0, with 0.0 indicating flawless competence and 1.0 indicating the poorest performance.

$$\text{Brier score} = (\text{Actual result} - \text{Forecast Probability}) \quad (13)$$

In comparison to other forecasts, Brier scores can reveal how successful a forecast was.

Result Analysis

The proposed method makes use of multiple classification algorithms as well as an ensemble approach, where all the approaches have been developed using Python. AdaBoost, KNN, Random Forest, and XGB have been used and then Random Forest, XGBoost and KNN have been ensembled by using Stacking CV Classifier. Figure 3 and Figure 4 depicts the Brier-score for Dataset I and II.

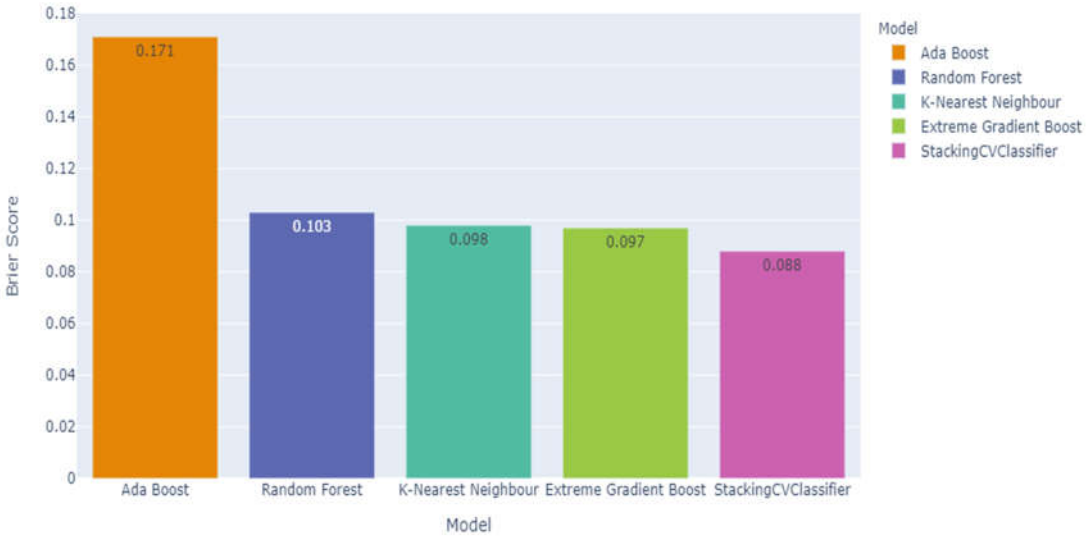

Figure 3. Brier Score of Dataset I

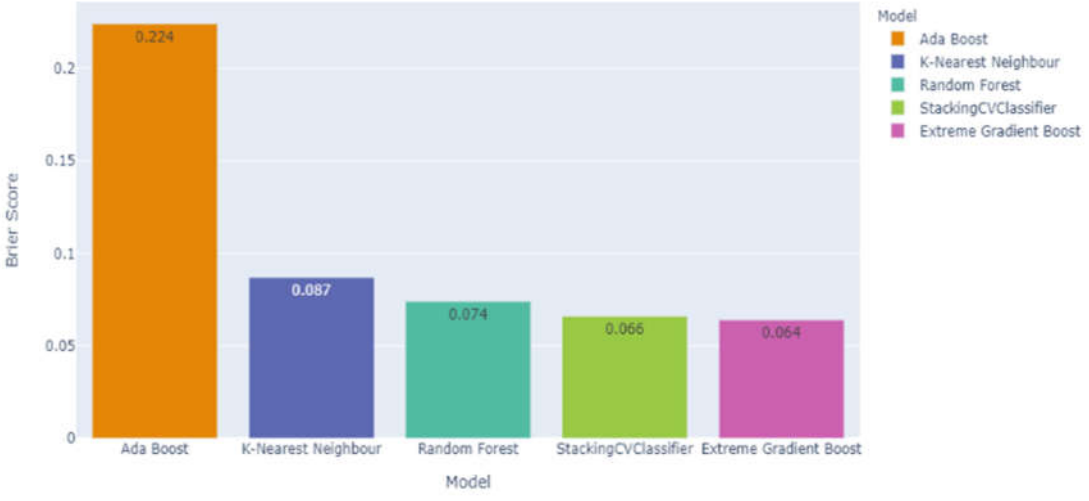

Figure 4. Brier Score of Dataset II

KNN and XGBoost is used as meta classifier for Dataset I and Dataset II respectively. Splitting the datasets by 80% and 20%, train and test sets were constructed.

Table 3. Result of different evaluation metrics for dataset I

Model	Precision	Sensitivity	Specificity	Brier Score
Random Forest	0.90	0.90	0.89	0.102
XG Boost	0.90	0.90	0.89	0.095
K-Nearest Neighbour	0.88	0.93	0.86	0.097
Ada Boost	0.87	0.84	0.86	0.170
Ensemble	0.93	0.90	0.93	0.086

Table 3 and Table 4 summarizes the evaluation metrics of all the classifiers that are essential to assess the strength of all the classifier.

Table 4. Result of different evaluation metrics for dataset II

Model	Precision	Sensitivity	Specificity	Brier Score
Random Forest	0.87	0.82	0.95	0.074
XG Boost	0.87	0.82	0.95	0.066
K-Nearest Neighbour	0.81	0.76	0.93	0.086
Ada Boost	0.86	0.76	0.95	0.224
Ensemble	1.00	0.82	1.00	0.067

Figure 5 and Figure 6 Depict the ROC-AUC curves of dataset I and dataset II

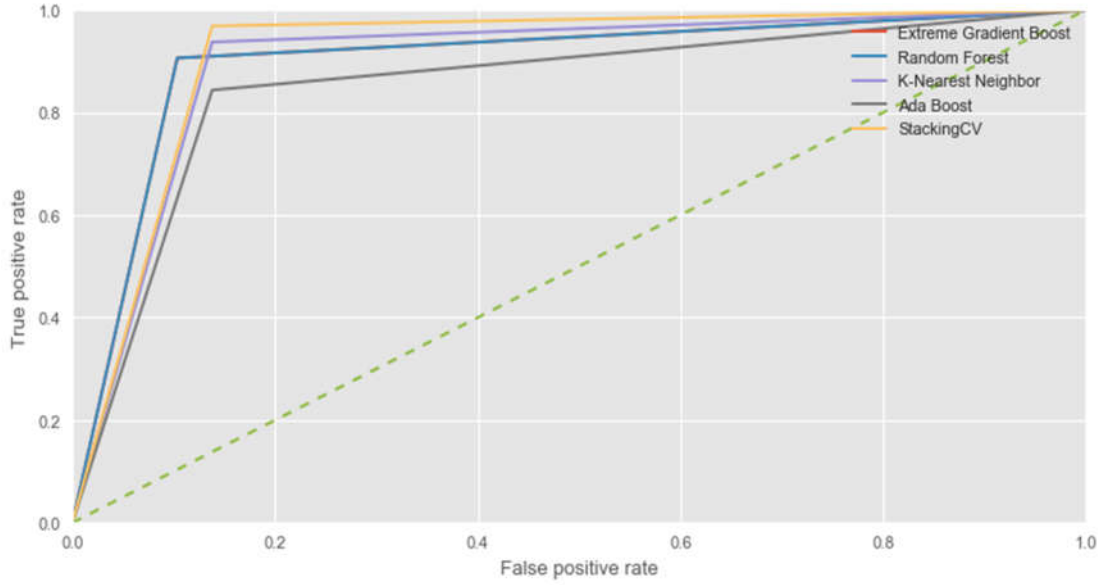

Figure 5. Proposed Model of ROC curve for Dataset I

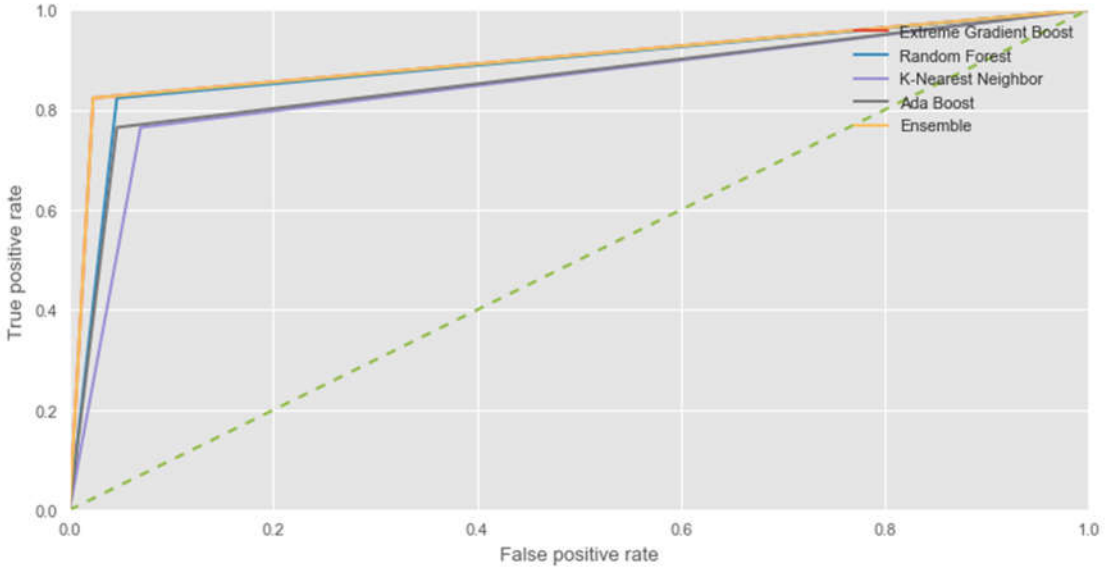

Figure 6. Proposed Model of ROC curve for Dataset II

Table 5 illustrates the accuracy of all the classifiers. For Dataset I AdaBoost achieves 85.2% accuracy, K-Nearest Neighbor achieves 90.2% accuracy, Random Forest (RF) achieves 90.2% accuracy, Extreme Gradient Boost (XGB) achieves 90.2% accuracy and the proposed Ensemble approach achieves 91.8% accuracy.

Table 5. Overall accuracy for both datasets

Model	**Dataset I**		**Dataset II**	
For Datasets	Train Accuracy	Test Accuracy	Train Accuracy	Test Accuracy
Random Forest	100.%	90.2%	100%	91.7%
XG Boost	92.5%	90.2%	92.8%	93.3%
K-Nearest Neighbour	90.0%	90.2%%	88.2%	88.3%
Ada Boost	85.1%	85.2%	90.7%	90.0%
Ensemble	91.3%%	91.8%	94.1%	93.3%

For Dataset II, AdaBoost achieves 90.0% accuracy, K-Nearest Neighbor achieves 88.3% accuracy, Random Forest (RF) achieves 91.7% accuracy, Extreme Gradient Boost (XGB) achieves 93.3% accuracy and the Ensemble approach achieves 93.3% accuracy. Obtained results show that the ensemble approach performs best for both of the datasets. For dataset I, the highest accuracy 91.8% has been achieved by the Ensemble approach, and for dataset II, the highest accuracy 93.3% has been achieved by the Ensemble approach. So, based on the experimental findings, it can be concluded that the Ensemble approach is effective in the early prediction of heart diseases and it surpasses the performance of previous studies. Figure 7 depicts the achieved accuracy of all classifiers.

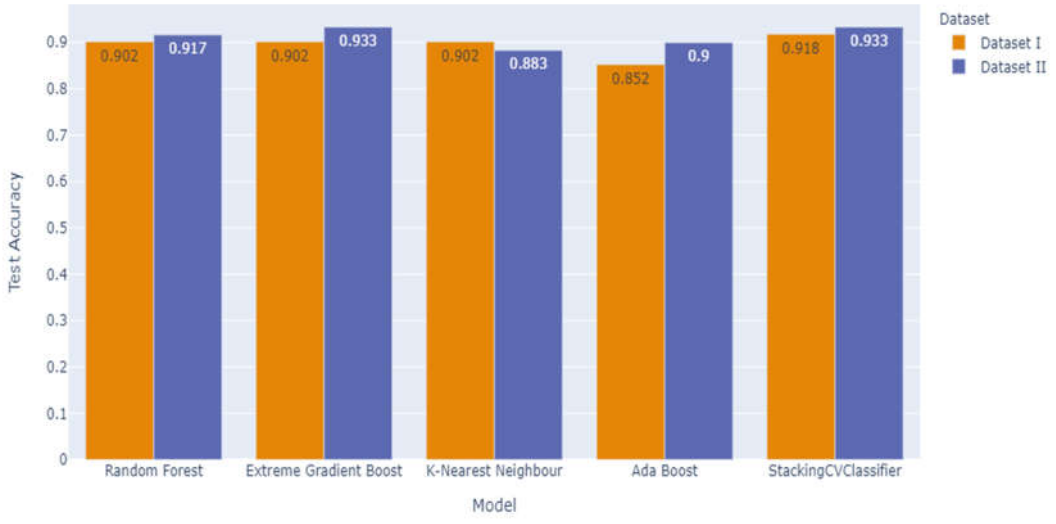

Figure 7. Accuracy of different approaches for Dataset I & Dataset II

Conclusion

Along with the number of deaths caused by heart disease on the upswing, designing an accurate and effective system become absolutely essential. This paper finds the most effective ML algorithm for recognizing cardiac issues early in the course of heart disease and heart failure. Dataset I and dataset II have been used to assess the accuracy of Random Forest, KNN, Ada-boost, XGB-boost and ensembles

(RF, XGB, KNN) algorithms for predicting early-stage heart disease and failure. According to the results found, the ensembles is the most effective approach to predict heart disease and heart failure with an accuracy score of 91% and 93%.

References

Jindal, H., Agrawal, S., Khera, R., Jain, R., & Nagrath, P. (2021). Heart disease prediction using machine learning algorithms. In *IOP Conference Series: Materials Science and Engineering* (Vol. 1022, No. 1, p. 012072). IOP Publishing.

Shah, D., Patel, S., & Bharti, S. K. (2020). Heart disease prediction using machine learning techniques. *SN Computer Science*, *1*(6), 1-6.

Gaziano, T. A., Bitton, A., Anand, S., Abrahams-Gessel, S., & Murphy, A. (2010). Growing epidemic of coronary heart disease in low-and middle-income countries. *Current problems in cardiology*, *35*(2), 72-115.

Groenewegen, A., Rutten, F. H., Mosterd, A., & Hoes, A. W. (2020). Epidemiology of heart failure. *European journal of heart failure*, *22*(8), 1342-1356.

Rajdhan, A., Agarwal, A., Sai, M., Ravi, D., & Ghuli, P. (2020). Heart disease prediction using machine learning. *International Journal of Research and Technology*, *9*(04), 659-662.

Singh, A., & Kumar, R. (2020, February). Heart disease prediction using machine learning algorithms. In *2020 international conference on electrical and electronics engineering (ICE3)* (pp. 452-457). IEEE.

Alić, B., Gurbeta, L., & Badnjević, A. (2017, June). Machine learning techniques for classification of diabetes and cardiovascular diseases. In *2017 6th mediterranean conference on embedded computing (MECO)* (pp. 1-4). IEEE.

Gavhane, A., Kokkula, G., Pandya, I., & Devadkar, K. (2018, March). Prediction of heart disease using machine learning. In *2018 second international conference on electronics, communication and aerospace technology (ICECA)* (pp. 1275-1278). IEEE.

Basha, N., PS, A. K., & Venkatesh, P. (2019, December). Early detection of heart syndrome using machine learning technique. In *2019 4th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT)* (pp. 387-391). IEEE.

Rubini, P. E., Subasini, C. A., Katharine, A. V., Kumaresan, V., Kumar, S. G., & Nithya, T. M. (2021). A cardiovascular disease prediction using machine learning algorithms. *Annals of the Romanian Society for Cell Biology*, 904-912.

"Dataset1," [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Heart+Disease.

Patel, J., TejalUpadhyay, D., & Patel, S. (2015). Heart disease prediction using machine learning and data mining technique. *Heart Disease*, *7*(1), 129-137.

"Dataset II," [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records.

Kaushik, S., & Birok, R. (2021, September). Heart Failure prediction using Xgboost algorithm and feature selection using feature permutation. In *2021 Fourth International Conference on Electrical, Computer and Communication Technologies (ICECCT)* (pp. 1-6). IEEE. https://towardsdatascience.com/the-mathematics -of- decision - trees - random - forest - and - feature - importance - in - scikit - learn - and-spark-f2861df67e3. Accessed on 11 March 2023. https://machinelearningmastery.com/gentle-introduction-xgboost-applied - machine-learning/. Accessed on 11 March 2023. https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm. Accessed on 11 March 2023. https://en.wikipedia.org/wiki/AdaBoost. Accessed on 11 March 2023.

Akobeng, A. K. (2007). Understanding diagnostic tests 1: sensitivity, specificity and predictive values. *Acta paediatrica*, *96*(3), 338-341.