A Soft Voting Machine Learning Model with Explainable AI for Cardiovascular Disease Management

Thesis · November 2022

DOI: 10.13140/RG.2.2.32578.76486

1 author:

Fahim Muntasir
BRAC University

A Soft Voting Machine Learning Model with Explainable AI for Cardiovascular Disease Management

Fahim Muntasir

Department of Electrical and Electronic Engineering

Rajshahi University of Engineering and Technology, Rajshahi

Email: muntasirfahim.niloy@gmail.com

Abstract

Cardiovascular disease has become the reason to more than 30% of the deaths globally. It is becoming prevalent because of poor lifestyle choices, a dearth of health consciousness, and unhealthy eating habits. Medical facilities and hospitals confront significant difficulties in accurately predicting and diagnosing different types of CVDs. The risk factor of CVD in an individual can be predicted based on their different health parameters. To effectively classify and anticipate complicated, highly nonlinear medical data, machine learning methods have been attempted by various researchers in previous studies. However, these research works failed to gain trustworthiness from the healthcare perspective as the focus was given only on increasing the model performance and not on the explainability. Thus, discouraging medical practitioners to implement them in real life scenarios. In this study, easy explanation and understanding of the results were given priority along with achieving better results for predictions. The proposed methodology encapsulates a highly accurate and efficient CVD predictor with an explainable AI. A comprehensive CVD dataset, containing data from 5 hospitals was used. There were 1190 instances, where 629 were CVD patients, and 561 cases were non-CVD with several CVD attributes. After implementing various data cleaning, feature scaling and analysis, eight machine learning algorithms, artificial neural network (ANN), decision tree (DT), random forest (RF), extreme gradient boosting (XGBoost), logistic regression (LR), naïve bayes (NB), k nearest neighbor (KNN) and support vector classifier (SVC) were selected as base classifiers. The best parameters for each base algorithm were found using grid search and fivefold cross validation. Then soft voting ensemble was implemented on three levels of accuracies. Weighted Ensemble of RF, XGBoost and SVC performed the best, reaching 94% accuracy and F1-score of 94%. Lastly to explain the model Shapley additive explanations (SHAP) was used on the best ensemble to produce local and global explanations, and various graphical representations were produced for easy understanding of the results. The proposed approach can be used for determining early CVD risk possibility in a patient, help medical professionals in understanding of CVD diagnosis. This can also be a great step towards preventive medicines.

Keywords

Cardiovascular disease prediction; SHAP Explainable AI; Ensemble Learning, Preventive Medicine

1. Introduction

1.1 Definition and Facts

The term cardiovascular refers to both the heart (cardio) and the blood arteries (vascular). Arteries, Arterioles, Capillaries, the Heart, and Venules comprise the cardiovascular system in a human body [1]. Cardiovascular diseases (CVDs), is among the top causes of death globally, accounting for approximately 18 million deaths annually and more than 30% of all fatalities globally [2]. There are mainly four types of cardiovascular diseases [3]. Firstly, coronary heart disease (CHD), which means the blood supply to the heart tissue is blocked by fatty substances in the coronary arteries. This blockage of blood flow will cause angina and lead to heart attack. This is the most common of the four. Secondly, there is stroke, which is similar blockage of blood flow, but towards the brain. With each pulse, arteries transport around one fourth of the blood to the brain, where billions of cells consume approximately 20% of the oxygen and fuel carried by the blood. While doing brain engaging activities, the brain may burn up to half of its fuel and oxygen [4]. If this flow gets disrupted, this causes stroke. The third one is peripheral arterial disease. This happens when the blood can't reach the limbs, specifically legs. This can start from being a tiny pain to an extended paralysis condition when the limbs stop functioning due to lack of blood flow. Lastly, the aorta, the largest vessel that propagates blood to the rest of the body, can get damaged. This is called aortic disease. CVD illnesses are generally associated with blood vessel blockages or narrowing, which can lead to heart attacks, strokes, and angina. Heart disorders come in a variety of forms, including those that damage the heart's rhythm, valve, or muscle. The most mistaken sign of cardiovascular disease is heart disease. It is sometimes also referred to as a cardiac condition [5]. Basically, all heart diseases are CVDs, but not all CVDs are heart diseases.

1.2 Problem Statement

CVDs are much too common in today's time. Lifestyle risk factors, such as food choices, physical inactivity, smoking, and obesity have a significant impact on traditional cardiovascular risk factors [6]. Minor changes in these lifestyle risk variables might have a significant impact on cardiovascular risk. Early prediction of CVD is very essential, as the treatment in most cases is not only expensive, but also complex and life threatening. In some cases, if not treated on time, sudden death is also possible. Currently available medical lab tests are time consuming and require expert personnel. This is an issue for healthcare management, while handling the vast number of CVD patients. Moreover, understanding how the risk factors and different health conditions are impacting cardiovascular outcomes of a human body can make a great aid to eliminating CVD risks, helping to provide accurate suggestions.

1.3 AI Research challenges in CVD Diagnosis

Over the years, there have been attempts to develop reliable machine learning (ML) models, to predict CVDs [7]. Some early works got lower accuracies due to adapting wrong methods and not having proper datasets to work on. In recent research, the models, being good starting points, have failed to attain exceptional performance from baseline models. Moreover, the researchers focusing towards complex and sophisticated models, have deviated from the actual purpose of effectively providing knowledge about the patient to the medical professional. This minimal influence is mostly the result of the ML models' enigmatic character [8]. Explaining the medical situation by depicting the causes and effects, while performing the prediction would make AI approaches easier to understand and trustworthy for real life use cases.

1.4 Aim, Contribution and Paper Organization

The focus of the study is to develop a robust ML model along with explaining the impact of the input factors on the prediction. There are two parts to the research. Firstly, to develop the best prediction model and secondly make the predictions interpretable. The soft voting ML classifier along with SHAP explainable AI, has enabled us in crafting this easy-to-use framework. The suggested method can assist in initiating appropriate action at the very earliest stages of the disease and enhance clinical knowledge of CVD diagnosis. The unique contributions are listed below:

- Different machine learning models, based on different algorithms were applied and soft voting ensemble was later developed by combining three of them.
- SHAP Analysis was done to give a clear view of the important features. The visualizations along with feature importance scores would help with decision-making.
- The model's internal explainability was included to make the model more dependable and to establish a good equilibrium between accuracy and interpretability, making it easier for doctors or clinicians to grasp and implement the model.

The following section contains the literature review on the current state of ML application in CVD study. In section 3, the dataset, proposed algorithms, and approaches are described. Section 4 has the findings explained and the study was concluded in section 5.

2. Literature Review

Many researchers have spent their time and hard work in developing prediction models for cardiovascular disease. Here, some of the most impactful works have been reviewed.

Alqahtani et al. in 2022, implemented ensemble learning for predicting cardiovascular disease on UCI Heart Disease dataset. Their work was based on both neural networks and ensemble of generic machine learning models. The ensemble mechanism used here was hard voting or, majority voting. Their approach found the Deep Neural model to gain 87.59% accuracy. Their ensemble of Extreme Gradient Boost (XGBoost), K Nearest Neighbor (KNN) and Decision Tree (DT) model reached an accuracy of 88.70% [9]. No data preprocessing was opted in the work. Same dataset was used in [10], Logistic Regression (LR), Naive Bayes (NB), DT, Random Forest (RF), Support Vector Machine (SVM), XGBoost, Light Gradient Boosting Machines (LGBM), and KNN were used in ensemble. Ensemble 1 with SVC and LR and Ensemble 2, with NBC, RF, SVC, and LGBM, both had an accuracy of 88.33%. Ensemble of LR, NB and KNN reached 84% accuracy by [11], A similar approach was seen in [12], where same three algorithm was cross validated and later ensembled to improve the accuracy to 90%. The study also showed that ensemble can gain significant performance boost than solo algorithms. SMOTE was used to handle the imbalance in the dataset of [13]. The dataset was consisted of 299 patients from Allied Hospital Faisalabad (Pakistan) and Faisalabad Institute of Cardiology (Punjab) from April to December 2015. Using

AdaBoost algorithm, their accuracy reached an excellent 96%. A dataset of 70000 patients were used by [14]. They found LR to perform better than NB, RF and SVM. This reached only 72% accuracy. They saw an improvement in accuracy by opting ensemble method [15]. Achieving 87% accuracy by aggregating KNN, LR, SVM, RF and AdaBoost. This time they also treated the dataset for class imbalance with SMOTE as the ratio of 0 and 1 were 30:70. On the same dataset, using cross validated Multi-Layer Perceptron (MLP) model gained 87% accuracy [16], splitting the dataset in 80:20 train to test ratio. LR was found to be better than DT, KNN and ANN, with 84% accuracy in [17].

Another approach to classify heart disease is based on ECG signal. In [18] XGBoost performed well with long-term data but had a lengthy inference time due to time-consuming calculations during the pre-processing step. In both the CinC 2017 and CinC 2020 datasets, the 1D ResNet, produced the greatest results, with F1 scores of 85% and 71%, respectively. [19] used ML methods such as XGBoost, RF to predict the effects of different metals in CVDs. They found RF to achieve 96% accuracy. SHAP values were also calculated to understand which metal had most impacts.

3. Methodology

3.1 Proposed Approach

The whole workflow is given in figure 1. The dataset was downloaded from IEEE data port. Several preprocessing was done on the dataset to ensure there were no garbage values. After the preprocessing steps, the dataset was divided into training and testing. 75% of the data was used for training the models. The performance was tested on the remaining 25%. Several ML techniques were attempted and tested. Decision Tree (DT), Random Forest (RF), Extreme Gradient Boosting (XGBoost) were selected as tree-based algorithms. Logistic Regression (LR), Nave Bayes (NB), K Nearest Neighbor (KNN), Support Vector Classifier (SVC) were selected as classification algorithms. Lastly, an Artificial Neural Network (ANN) was also implemented to check if, ML algorithms tend to outperform neural networks in statistical predictions [20]. Data pipeline was created for each algorithm with min-max normalization as scaling method. This converted all the data into a range of 0 and 1. Each algorithm was hyper-parameter tuned using Grid search cross validation (GridsearchCV), with 5-fold cross validation. This helped to choose optimal parameter values.

For aggregating the models, weighted soft voting was chosen. This takes into consideration each classifier's confidence in its class probabilities [21]. As the classifiers can deliver well-calibrated probability estimates, weighted soft voting can be more informative and perhaps more accurate. It applies varying weights to the output of each classifier dependent on its dependability. When dealing with unclear or confusing cases, weighted soft voting tends to be more effective than hard voting [22]. This ensemble is also suitable in handling imbalance of the dataset.

Three ensembles were made, firstly three lower performing models, then one lower and two higher performing models and lastly three high accuracy models. Higher weights were given to better models. ANN was not taken into ensemble as ensemble of traditional ML algorithms with neural networks tend to increase complexity by a huge margin, gaining little to no performance gain.

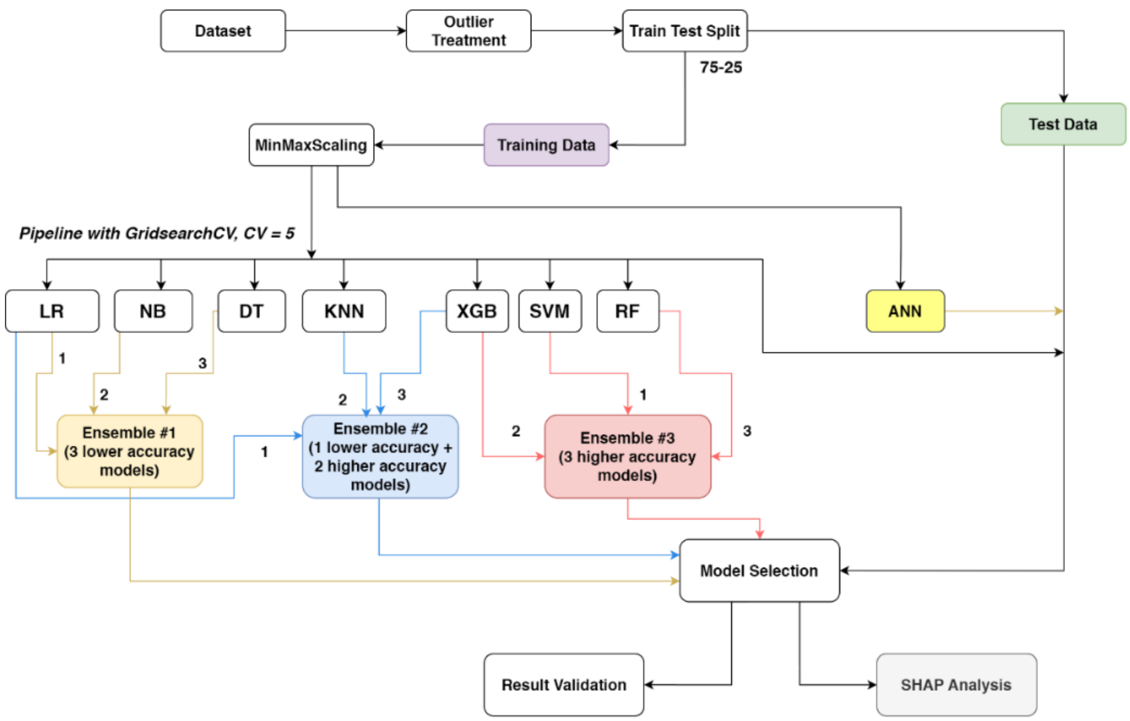


Figure 1. Methodology

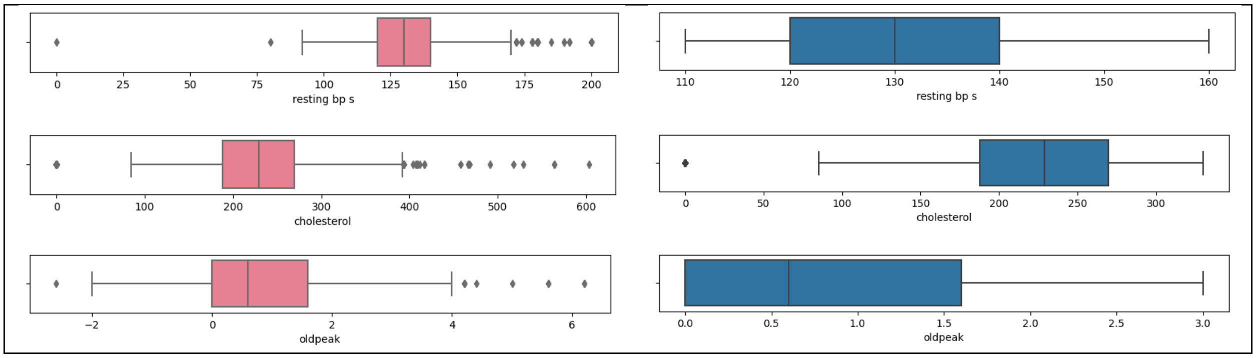
3.2 Dataset Description

The work is based on the most comprehensive heart disease dataset, containing real life data from 5 different databases (Cleveland, Hungarian, Switzerland, Long Beach VA and Statlog) and 1190 patients among which 629 had some kind of heart disease and 561 were healthy [23]. The dataset had eleven features related to CVDs. The main reason for choosing this dataset was its availability and authenticity. Moreover, having data from different sources made the dataset more complex, thus helping the ML models to understand more variations.

3.3 Preprocessing

Loading into notebook, the dataset was explored for better understanding. It was checked if the dataset had any null or missing values. There were no missing values. Dataset balance was checked next. From the target column, it was observed that the dataset had 53% positive and 47% negative values. So, the dataset was nearly balanced. As all the datapoints were real life, no oversampling or under sampling was done on the dataset. The next step to preprocessing was dealing with outliers. Outliers are the noise in the features, or irregular datapoint that has abnormal distance from the rest of the population. The resting blood pressure, cholesterol and old peak had outliers in them. Those were dealt with by replacing the lower values by 10th percentile values and upper values by 95th percentile values. This helped in removing the extreme values without losing the dimensionality of the dataset.

Table 1. Distribution of features with outlier and after replacing them.



The correlation between the features and targets were then checked by Pearson correlation. The correlation coefficients varied from -1 to +1. A number closer to 0 indicates a lower association. The value 0 indicates that there is no association. A number closer to +1 or -1 implies a higher positive or negative association. There weren't any inter-correlating features based on the 0.7 threshold, so all the features were kept.

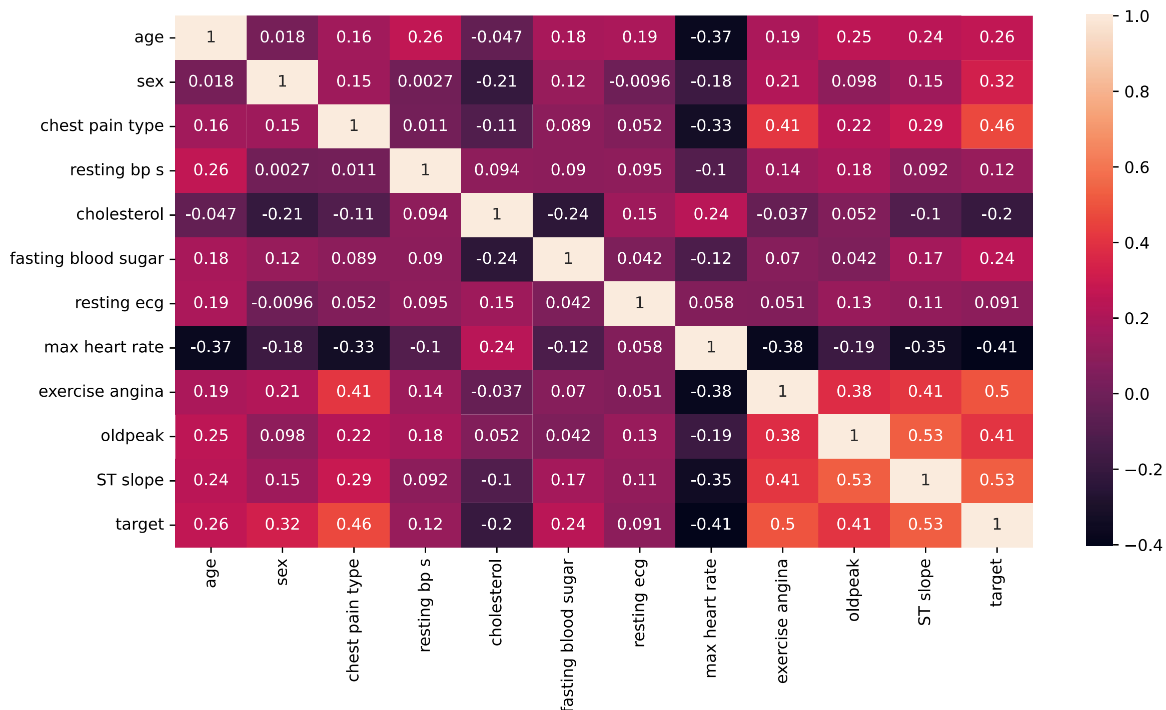


Fig 2. Pearson correlation between different columns.

3.4 Models and Algorithms

3.4.1 XGBoost

This technique generates decision trees sequentially. Prior to entering each independent variable into the decision tree that predicts results, each independent variable is assigned a weight. Before entering the second decision tree, variables that the first tree mis predicted are given a harsher weighting. A strong and precise model is created by combining these various classifiers. Poor base learners have a large bias and slightly better prediction ability than chance. The boosting approach combines these weak learners to create a strong learner, reducing both bias and variation. This technique is made possible by parallel and distributed computing [24].

3.4.2 Random Forest

Random forest is a classification algorithm that uses decision trees to improve prediction accuracy and reduce overfitting. It employs averaging and bagging of decision trees to increase prediction accuracy [25]. Random vectors are selected randomly across all trees, combining multiple tree predictors. As the number of trees increases, the generalization error corresponds to a limit. The strength of each tree and their correlation determine the generalization error. Random node division provides error rates comparable to other boosting techniques and is more noise resistant.

3.4.3 Logistic Regression

Regression methods are essential in data studies to determine the connection between a response variable and explanatory factors. Logistic regression, a binary classifier, outputs binary data and is used for classification problems [26]. Logarithmic conversion on the result variable can represent non-linear connections in a linear fashion, with a 50% chance of success when the log of the odd ratio is positive.

3.4.4 Support Vector Machine

SVM is a classification model that applies labels to objects and learns patterns. It uses a training method that can delete unusual instances and automatically adjust the model's capability by maximizing the gap between training instances and class boundaries [27]. The SVM algorithm divides the dataset into labelled data and splits them into positive and negative hyperplanes. Support vectors are chosen to create the hyperplane. SVM is efficient for binary classification and fast for small to moderately large datasets.

3.4.5 K Nearest Neighbor

The k-nearest neighbor method, or KNN for short, is a supervised learning classifier that employs proximity to produce classification or predictions about how a particular data point will be grouped. While it can be applied to classification or regression issues, it is commonly employed as a classification method because it relies on the idea that comparable points can be discovered close to one another.

3.4.6 Naïve Bayes

Naive Bayes is a simple learning technique that assumes characteristics are conditionally independent, performing Bayes' rule. Despite being often broken, it consistently produces competitive classification accuracy, making it widely used. It also has desirable properties like being computationally efficient. The Bayes Theorem determines the probability of an incident based on earlier events [28].

3.4.7 Decision Tree

Decision tree methodology is a widely used data mining technique for creating classification systems and prediction models for target attributes. It divides a population into branches, forming an inverted tree with root nodes, internal nodes, and leaf nodes. This non-parametric approach can handle complex datasets without imposing a challenging parametric framework [29]. Data can be split into training and validation datasets, and a decision tree model is constructed using the training dataset and the validation dataset. Root nodes divide records into mutually exclusive sets, internal nodes indicate available options, and leaf nodes reflect the outcome of a string of choices or occurrences.

3.4.8 Artificial Neural Network (ANN)

Neural network approaches have been proposed for this sort of statistical analysis in many studies. Here, a simple neural network was implemented to test against the available tree-based algorithms. Neural networks, though quite old concepts, were not used due to their computational complexity. They work as human brains, by setting weights and updating them in different neurons. The ANN architecture used in this study is stated below.

Table 2. ANN Architecture

Epoch	50
Input layer	11
Hidden layer 1	16, activation=relu
Hidden layer 2	20, activation=relu
Dropout layer	2%
Output	1, activation = sigmoid

As all the algorithms were parameter tuned using grid search cross validation method, the tuned parameters and hyperparameters are stated in the table below:

Table 3. Tuned parameters of different algorithms

Algorithm	**Tuned Parameter**
XGBoost	N estimators = 100, learning rate = 0.1, max depth = 10
Logistic Regression	Max iter = 100
Decision Tree	criterion = gini, max depth = 5, max features=7, splitter = best
Random Forest	Max depth = 12, n estimator = 1000
KNN	N neighbor = 5
SVM	C = 100, gamma = 0.1, kernel = rbf

3.4.9 Soft Voting Ensemble

This work combined various algorithms to create a robust model using a voting classifier method. This technique aggregates multiple classifiers to act as one, learning from the total dataset. The model's effectiveness is determined by the weight assigned to each algorithm, allowing more efficient models to engage more in the ensemble. There are two types of voting, hard and soft. Hard voting involves a majority vote, while soft voting involves averaging the results of multiple algorithms. Here we opted for soft voting ensemble in three ways. firstly with 3 models having the lowest accuracy, then 3 models where two high accuracy models were assembled with one low accuracy model and lastly, 3 higher accuracy models were assembled. Higher weight was put on the best performing one in each ensemble.

3.4.10 SHAP Analysis

SHapley Additive exPlanations (SHAP) analysis works based on game theory, producing feature importance score for each feature in each instance [30]. We performed the feature impact study based on SHAP values for each algorithm. There were both local and global explanations, portrayed in various graphs and visualizations for better interpreting the result by medical personnel. Summary plot from SHAP library showed the overall impact of the features behind the classification. Force plot and bar plots were used to show the feature importance of single instance or single patient.

4. Performance Analysis

As the problem was a classification one, accuracy, precision, recall and f1 score were used as the performance metrics. Apart from these, confusion matrix was also viewed as it shows the number of type-1 and type-2 errors made in test predictions. Accuracy represents the total number of correctly predicted outcomes in the test set. Precision is a metric that measures the accuracy of the positive predictions made by a classification model. Recall, also known as sensitivity or true positive rate, measures the ability of a classification model to identify all relevant instances in the dataset. F1 Score provides a balanced measure of a model's accuracy that considers both false positives and false negatives. It is especially useful when a balance between precision and recall is required, as it combines both metrics into a single value, with higher F1 scores indicating better model performance.

4.1 Result

The base models with the best parameters from grid search cross validation, performed exceptionally well. The Random Forest and XGBoost performance were among the highest. The Neural Network approach did good spite of the tabular nature of the dataset. As the research proposal, three ensembles were made from different base models. Soft voting helped with the low performing models getting a higher accuracy. The ensemble of high performing models mostly followed the results of the single best models.

Table 3. Performance Comparison of Models

	Accuracy	**Precision**	**Recall**	**F1**
LR	85	84	84	84
DTC	88	87	88	87
NB	85	85	85	85
XGBoost	92	92	92	92
RFC	93	93	93	93

ANN	90	90	90	90
KNN	88	88	88	88
SVM	89	89	89	89
DTC + LR + NB	89	89	89	89
LR + XGB + KNN	92	92	92	92
RFC + XGB + SVC	**94**	**94**	**94**	**94**

4.2 Comparison with Previous Works

As the table suggests, comparing with recent works on the same dataset, our work has outperformed all the baseline and previous models with excellent accuracy and f1 score. This suggests that our approach was better than the existing baseline models, which combined with the extracted data insight makes the work a unique one.

Table 4. Result Comparison with Previous Works

Year	*Author, Dataset*	*Accuracy*	*Precision*	*Recall*	*F1*
2022	Alqahtani et al., UCI	88	88	87	88
2022	Gupta et al., UCI	88.4	85	88	88
2022	Itoo et al., UCI	90	89	88	90
2022	Kameswara et al., Author Provided	96	95	96	96
2022	Majumder et al., UCI	83	83	83	83
2021	Nguyen et al., UCI	87	81	-	83
2023	***Proposed Method, UCI Comprehensive***	***94***	***94***	***94***	***94***

5. Feature Importance Analysis

5.1 Chi-Square Test

The feature importance analysis was done both statistically and with explainable AI method. Chi-squared tests are commonly used in statistics to determine if there is a significant association between categorical variables. Chi-Square helps us understand which features are statistically significant for the change of target column. The chi square test was done, and the following result was obtained:

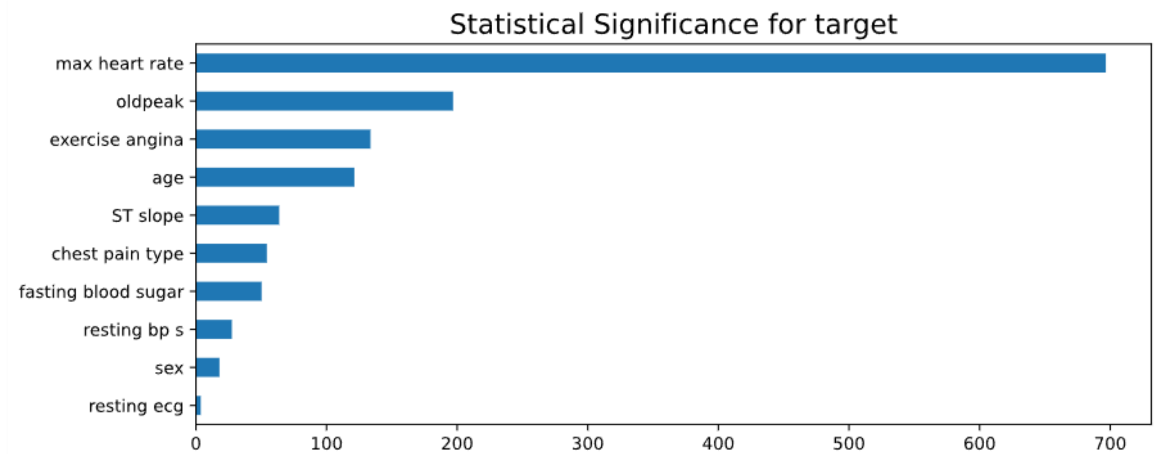


Figure 3. Statistically significant features to target change.

According to statistics, max heart rate, old peak and angina from exercise were the most important features for CVD detection. Keeping this in mind, the SHAP analysis was performed.

5.2 SHAP Results

In a SHAP plot, the visualization shows how the prediction is affected by different features. Generally, any machine learning model will create a base prediction, which is an arbitrary value based on the possibility of a sample being in majority class. The base prediction is then tuned by taking different features in different nodes and performing necessary calculations. Here all the predictions are divided into two classes, 0 for non-CVD and 1 for CVD patients. For any randomly taken sample the feature importance can be shown with force plot and bar plot. The overall features important to a model can be seen from summary plot. Here using kernel explainer API from shap, the best voting model was used for feature importance analysis. The summary plot was obtained as below:

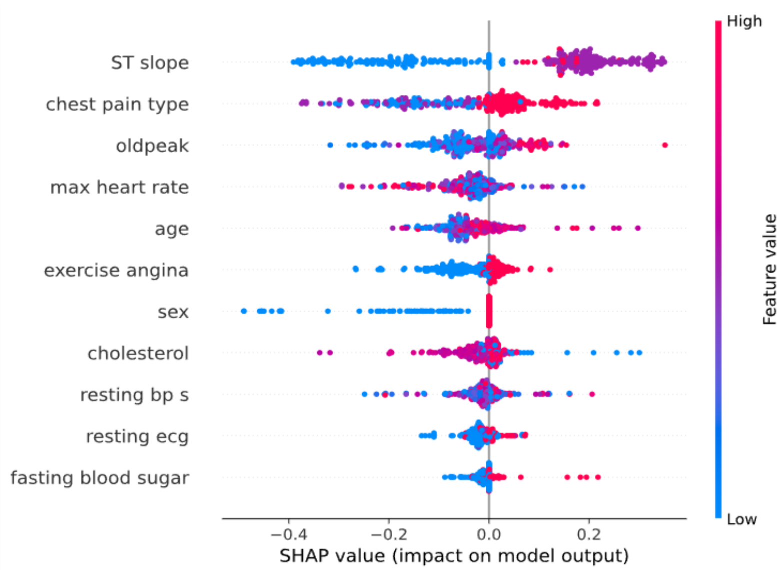
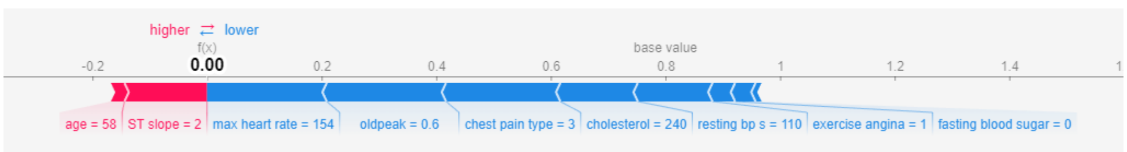


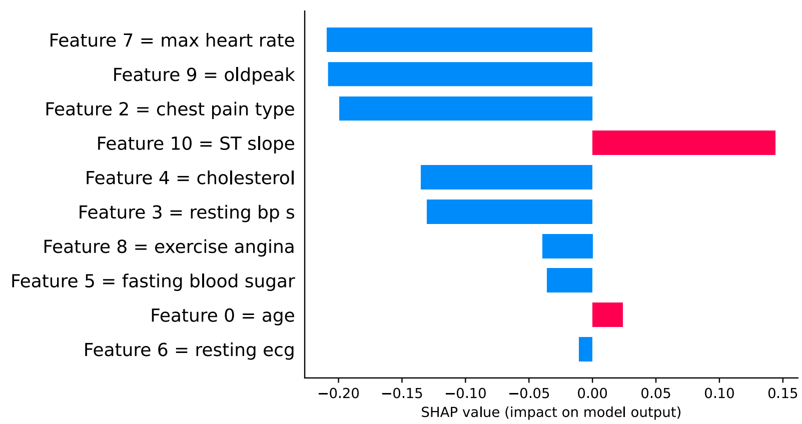
Fig 4. SHAP Summary Plot for Best Ensemble Model

Here each feature distribution shows the impact of feature value on SHAP values, thus describing the effect on the prediction class. The ST Slope had four values, 0,1,2 and 3. The higher the value, the higher the SHAP value was. This means as the ST slope value increased it was more prone to being class 1. The same trend is seen in chest pain type as well. It had four levels, from no chest pain (1) to severe chest pain (4) and, higher levels had higher SHAP values. Meaning high chest pain levels are likely to be of CVD patients. All the features showed this relationship as we see from the summary plot. This plot can also be used to understand socio-demographic factors if those are present in the feature set. Here, male were class 1 and female were class 0 and it is seen that women are less likely to have CVD than men. Also, based on age, older people are more likely to have CVD than younger people.

We took a random sample from the dataset and made predictions in our classifier and found out the feature roles for that individual prediction. Patient 63 from test set was correctly predicted class 0 (non-CVD) with a possibility of 86%. The SHAP values from this prediction were plotted in force plot.



(a) Force plot

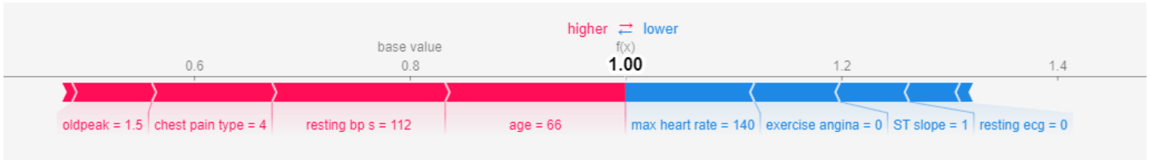


(b) Bar plot

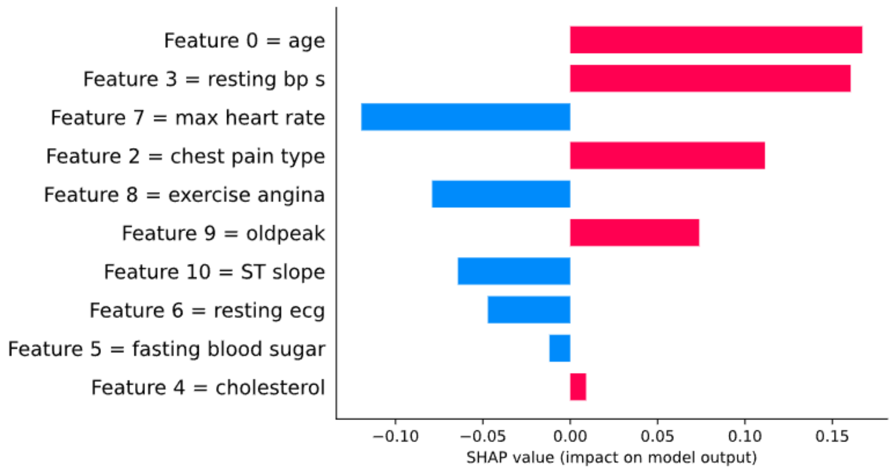
Fig 5. SHAP Values in class 0 prediction

Here, the model's base prediction was 0.8, which was then affected by all the features. The ST slope and age pushed the prediction higher, but the max heart rate, old peak, chest pain type, cholesterol, resting bp, angina, fasting blood sugar all these pushed the prediction lower. In this way the feature impact can be seen for each prediction. This can also be represented in a bar plot as shown.

Patient 203 from test set was correctly predicted as class 1 (CVD) with a possibility of 54%. Here the SHAP force plot was more challenging as we investigate.



(a) Force plot



(b) Bar Plot

Fig 6. SHAP Values in class 1 prediction

Age, resting blood pressure, chest pain and old peak pushed the prediction to 1, but max heart rate, angina, ST slope and resting ecg suggested the prediction to be class 0. As the class possibility here was merely more than 50%, the force plot was almost equal in each side of the spectrum.

6. Conclusion

In this study, a robust and interpretable ensemble model was developed using a combination of several machine learning (ML) methods to accurately predict the risk of cardiovascular disease. The ensemble model employed a cross-validation approach and harnessed the power of three distinct ML algorithms: Random Forest (RF), XGBoost, and Support Vector Classifier (SVC). This ensemble approach utilized a weighted soft voting classifier, demonstrating superior predictive performance compared to the individual algorithms. Remarkably, the system achieved an outstanding accuracy rate of 94% and a similarly impressive F1 score, placing it in a highly competitive position among the models presented in the existing literature. Furthermore, the suggested ensemble model not only produced accurate findings, but it also provided a high degree of interpretability by utilizing SHAP (SHapley Additive exPlanations) plots. These interpretability tools gave logical, clinically credible insights and practical judgements which can help a lot in risk predictions and early diagnosis. The implication of such a verbose model can greatly benefit the preventive medicine culture, as it can be used in a simple mobile application, which from inputs will be able to assess the risk of CVD, enabling healthy lifestyle suggestions. This can also extend into a hospital's CVDs management facility, substituting costly monitoring equipment such as

ECG machines, heart rate monitors etc. for inexpensive wearable devices. Enabling easy remote monitoring, personalized treatments and pointing to exact areas of improvement of a patient.

References

[1] Vorvick, L. J. (2023, 2/2/2023). *Cardiovascular*. National Library of Medicine (US). https://medlineplus.gov/ency/article/002310.htm#:~:text=The%20term%20cardiovascular%20refers%20to,Capillaries

[2] (WHO), W. H. O. *Cardiovascular diseases*. WHO. https://www.who.int/health-topics/cardiovascular-diseases/#tab=tab_1

[3] Scotland, N. I. (2022, 18/11/2022). *Cardiovascular Disease*. NHS Inform. https://www.nhsinform.scot/illnesses-and-conditions/heart-and-blood-vessels/conditions/cardiovascular-disease

[4] Hohler, S. E. (2011). *Caregiver's Guide: Care for Yourself While You Care for Your Loved Ones*. McFarland.

[5] Chang, V., Bhavani, V. R., Xu, A. Q., & Hossain, M. (2022). An artificial intelligence model for heart disease detection using machine learning algorithms. *Healthcare Analytics*, *2*, 100016.

[6] Mozaffarian, D., Wilson, P. W., & Kannel, W. B. (2008). Beyond established and novel risk factors: lifestyle risk factors for cardiovascular disease. *Circulation*, *117*(23), 3031-3038.

[7] Huang, J.-D., Wang, J., Ramsey, E., Leavey, G., Chico, T. J., & Condell, J. (2022). Applying artificial intelligence to wearable sensor data to diagnose and predict cardiovascular disease: a review. *Sensors*, *22*(20), 8002.

[8] Choudhary, S., Chatterjee, N., & Saha, S. K. (2022). Interpretation of black box nlp models: A survey. *arXiv preprint arXiv:2203.17081*.

[9] Alqahtani, A., Alsubai, S., Sha, M., Vilcekova, L., & Javed, T. (2022). Cardiovascular disease detection using ensemble learning. *Computational Intelligence and Neuroscience*, *2022*.

[10] Gupta, P., Mala, S., Shankar, A., & Asirvadam, V. S. (2022). Heart Disease Detection Scheme Using a New Ensemble Classifier. In *Advances in Data and Information Sciences: Proceedings of ICDIS 2021* (pp. 99-110). Springer.

[11] Majumder, A. B., Gupta, S., & Singh, D. (2022). An Ensemble Heart Disease Prediction Model Bagged with Logistic Regression, Naïve Bayes and K Nearest Neighbour. Journal of Physics: Conference Series,

[12] Itoo, N. N., & Garg, V. K. (2022). Heart Disease Prediction using a Stacked Ensemble of Supervised Machine Learning Classifiers. 2022 International Mobile and Embedded Technology Conference (MECON),

[13] Kameswara Rao, B., Prasan, U., Jagannadha Rao, M., Pedada, R., & Kumar, P. S. (2022). Identification of heart failure in early stages using smote-integrated adaboost framework. In *Computational Intelligence in Data Mining: Proceedings of ICCIDM 2021* (pp. 537-552). Springer.

[14] Dritsas, E., Alexiou, S., & Moustakas, K. (2022). Cardiovascular Disease Risk Prediction with Supervised Machine Learning Techniques. ICT4AWE,

[15] Dritsas, E., & Trigka, M. (2023). Efficient data-driven machine learning models for cardiovascular diseases risk prediction. *Sensors*, *23*(3), 1161.

[16] Bhatt, C. M., Patel, P., Ghetia, T., & Mazzeo, P. L. (2023). Effective heart disease prediction using machine learning techniques. *Algorithms*, *16*(2), 88.
[17] Sonia, S. E., Nedunchezhian, R., Ramakrishnan, S., & Kannammal, K. (2023). An empirical evaluation of benchmark machine learning classifiers for risk prediction of cardiovascular disease in diabetic males. *International Journal of Healthcare Management*, 1-16.
[18] Pham, H., Egorov, K., Kazakov, A., & Budennyy, S. (2023). Machine learning-based detection of cardiovascular disease using ECG signals: performance vs. complexity. *arXiv preprint arXiv:2303.11429*.
[19] Li, X., Zhao, Y., Zhang, D., Kuang, L., Huang, H., Chen, W., Fu, X., Wu, Y., Li, T., & Zhang, J. (2023). Development of an interpretable machine learning model associated with heavy metals' exposure to identify coronary heart disease among US adults via SHAP: Findings of the US NHANES from 2003 to 2018. *Chemosphere*, *311*, 137039.
[20] Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature machine intelligence*, *2*(1), 56-67.
[21] Kumari, S., Kumar, D., & Mittal, M. (2021). An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier. *International Journal of Cognitive Computing in Engineering*, *2*, 40-46.
[22] Géron, A. (2022). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. " O'Reilly Media, Inc.".
[23] Siddhartha, M. (2020). Heart disease dataset (comprehensive). *IEEE Dataport*.
[24] Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining,
[25] Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5-32.
[26] Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons.
[27] Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. Proceedings of the fifth annual workshop on Computational learning theory,
[28] Webb, G. I., Keogh, E., & Miikkulainen, R. (2010). Naïve Bayes. *Encyclopedia of machine learning*, *15*(1), 713-714.
[29] Song, Y.-Y., & Ying, L. (2015). Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, *27*(2), 130.
[30] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.