SAUCIS

Sakarya University Journal of Computer and Information Sciences

http://saucis.sakarya.edu.tr/

SAKARYA UNIVERSITY

DOI: 10.35377/saucis...1367326

RESEARCH ARTICLE

Prediction of Cardiovascular Disease Based on Voting Ensemble Model and SHAP Analysis

Erkan Akkur [1]

[1] Turkish Medicines and Medical Devices Agency, Çankaya, Ankara, Türkiye

Corresponding author:
Erkan Akkur, Turkish Medicines and Medical Devices Agency, Çankaya, Ankara, Türkiye
E-mail address:
eakkur@gmail.com

Submitted: 27 September 2023
Revision Requested: 15 October 2023
Last Revision Received: 11 November 2023
Accepted: 15 November 2023
Published Online: 15 November 2023

Citation: Akkur E.(2023). Prediction of Cardiovascular Disease Based on Voting Ensemble Model and SHAP Analysis *Sakarya University Journal of Computer and Information Sciences*.6 (3)
https://doi.org/10.35377/saucis...1367326

ABSTRACT

Globally, cardiovascular diseases (CVD) account for a large number of deaths. Early detection plays a critical role in reducing the mortality rate. Early detection can be achieved by utilizing machine learning algorithms on existing data of patients. Ensemble learning methods are one of the techniques applied to improve the classification performance of ML algorithms. This study suggests a prediction model based on voting ensemble learning for the prediction of CVD. The hyperparameters of classification algorithms are optimized by using grid search. The results of each model are validated by using a 10-fold cross-validation schema. The IEEE Data port dataset is used for all experiments Furthermore, the SHAP technique is employed to interpret the proposed prediction model, including the risk factors that play a role in detecting this disease The proposed model for CVD prediction achieved an accuracy of 0.937 and an AUC-ROC score of 0.936. The model presented in this study has a high classification rate compared to previous similar studies.

Keywords: Cardiovascular disease, Machine learning, Voting ensemble model, SHAP value

1. Introduction

Cardiovascular diseases (CVD) are considered conditions that negatively affect the heart and blood vessel system. Risk factors such as smoking, obesity, diabetes, lack of exercise, and unhealthy diet can cause these diseases. According to global statistics, CVD causes a large number of deaths worldwide. The high mortality rates also emphasize the importance of diagnostic methods [1-2]. In available clinical data, finding the difference between healthy and heart-diseased individuals has been a powerful approach in classification studies. The ability to classify CVD is a critical basis for the diagnosis of patients. In recent years, machine learning (ML) algorithms have played a significant role in research like the detection of CVD. The ML algorithms are one of the approaches applied to predict the status of CVD based on clinical data. These algorithms can make predictions by training with existing data sets. Models built with such methods with high classification rates can be used for diagnosis in new patient records. ML algorithms can help doctors with accurate prognostic predictions based on the patient's clinical data [3-5].

It is possible to increase the performance of ML algorithms by applying different methods. Ensemble learning (EL) techniques are one of the methods used to improve model performance by eliminating the disadvantages of classifiers. These algorithms aim to build models using multiple classifiers instead of a single classifier. When multiple classifiers are applied to train the input data, the actual predictions may outperform the result obtained by a single classifier [6]. The “black box” nature of ML algorithms also poses some challenges regarding the interpretability of the prediction models developed. An interpretable forecasting model is significant from a medical perspective as it enables people to understand the rationale behind the predictions and decisions made by the model. The SHapley Additive exPlanations (SHAP) method can illustrate the effect of attributes in the dataset on the final prediction. It can also effectively refine and explicate model predictions [7].

This study aims to improve the performance of classifiers for the prediction of CVD using voting techniques. The voting EL prediction model has been built using two different approaches, hard and soft voting. The SHAP analysis method is utilized to interpret the prediction model presented in this study.

2. Related Works

ML models have been proven to be effective in predicting CVD in numerous studies. The UCI Heart Disease dataset in the UCI Machine Learning Repository is publicly accessible and is a widely used dataset in this research area [8].

Mohapatra et al [5] applied a stacked EL model on Cleveland Heart Disease for CVD prediction. Ten different classifiers were used as base learners. The classification performance of the suggested EL model was compared with the base learner classifiers. The suggested model in the study achieved an accuracy of 92%. The study indicated that ensemble learning algorithms improve classification performance. Sangya et al [9] presented a comparative analysis of different ML algorithms including Logistic Regression (LR), K-Nearest Neighbor (K-NN), Support Vector Machine (SVM), Decision Tree (DT), and Random Forest (RF) algorithms for predicting CVD on the Cleveland dataset. Data preprocessing steps were applied to fill in missing data and remove noise from the dataset. As a result of the experiments, the SVM algorithm achieved the best result with an accuracy of 89.34%. Shah et al. [10] compared different ML algorithms to predict CVD. NB, DT, K-NN, and RF algorithms were used classification process. The experiments were carried out on the Cleveland dataset. Data preprocessing stages were performed on the dataset before the classification process. As a result of the comparisons, the K-NN algorithm achieved the best classification rate with 90.78 % accuracy. Rajdhan et al. [11] employed various classification algorithms containing NB, RF, LR, and RF. The experiments were carried out on the Cleveland dataset. The dataset was divided into 80% training data and the rest test data. The RF algorithm outperformed with an accuracy of 90.16%. Poorani and Hemalatha [12] carried out the comparison of SVM, DT, MLP, RF, and J48 algorithms to predict CVD. The NB algorithm outperformed with 90.33% accuracy. Ozhan and Kucukakcalı [13] utilized the XGBoost (XGB) model to estimate the risk prediction of CVD. A 10-fold CV was applied to measure the performance of the classifier. The suggested model was achieved with an accuracy of 89.4%. Das and Sinha [14] suggested a voting-based EL model to predict CVD. The experiments were implemented on the Statlog Heart Disease dataset. The proposed model yielded 90.74% accuracy comparing K-NN, SVM, NB, DT, LR, and ANN algorithms. The study showed that EL models provide higher success rates than classical classifiers.

Akyol and Atilla [15] conducted a study comparing Gradient Boosting Machines, RF and NB algorithms for CVD detection. Recursive Feature Elimination with a cross-validation (CV) technique was applied to select the most discriminative features. The experiments were performed on the Statlog Heart Disease dataset and the SPECT dataset. In both datasets, the NB algorithm achieved the highest classification rate with accuracy of 86.42% and 77.78%. Jan et al. [16] proposed a voting EL model for CVD prediction using SVM, ANN, NB, RF, and Regression Analysis algorithms. The proposed model is tested on the dataset created by combining the Cleveland and Hungary datasets. The Weka Data Mining Tool is used for the analysis. The proposed method achieved an accuracy of 93%. Tiwari et al. [17] suggested an ensemble approach containing the stacked model for the prediction process. EXC, SVM, RF, and XGB algorithms were utilized as a base classifier. The results obtained by the suggested model compared with basic classifiers. The suggested model showed 92.34% accuracy on the Heart Disease Dataset (IEEE DataPort).

Yilmaz and Yagin [18] suggested a predictive model containing SVM, LR, and RF algorithms for CVD prediction. The performance of models was evaluated on the Heart Disease Dataset (IEEE DataPort). The hyperparameters of the ML algorithms were tuned using a 10-fold repeated CV. RF algorithm yielded an accuracy rate of 92.9%. Doppala et al. [19] utilized the EL approach for the prediction of CVD. NB, RF, SVM, and XGB algorithms were adopted as base classifier algorithms. The Majority Voting technique was utilized as an EL approach using the Cleveland, IEEE Dataport, and Mendeley Data Center datasets, respectively. The presented EL method demonstrated accuracy rates of 88.24%, 93.39%, and 96.75%. The suggested model achieved higher classification rates than classical classifiers. García-Ordás et al. [20] utilized a Conventional Neural Network (CNN) algorithm for CVD prediction. A 10-fold CV approach was utilized to avoid randomness. The proposed model achieved a better result with an accuracy rate of 90.09% compared to ML algorithms.

Table 1 summarizes some recent studies on CVD prediction in the literature. Although machine learning is used for CVD prediction, the majority of studies have been conducted with individual classifiers. However, the number of studies using EL approaches is quite small. When determining the hyperparameters of classifiers, hyperparameter optimization is not commonly utilized. To address these limitations in the literature, a CVD prediction model based on EL is proposed in this study. The Grid Search technique is used for hyperparameter tuning. The SHAP analysis is also used, which allows machine learning algorithms to make interpretations on models by removing black box features.

Table 1: Some recent studies on CVD prediction in the literature

Author	Dataset	Techniques used	Claimed outcome	Accuracy	Limitations/Gaps
Sangya et al. [9]	Cleveland	LR, SVM, K-NN, RF, NB, DT	SVM	89.34 %	Small dataset is used. Hyperparameter tuning is not used.
Shah et al. [10]	Cleveland	NB, DT, K-NN, RF	K-NN	90.78 %	Small dataset is used Hyperparameter tuning is not used.
Rajdhan et al. [11]	Cleveland	DT, RF, LR, NB	RF	90.16 %	Small dataset is used Data pre-processing is not specified.
Poorani and Hemalatha [12]	Cleveland	DT, RF, NB, MLP, SVM	NB	90.33 %	Small dataset is used. Data pre-processing is not specified. Hyperparameter tuning is not used.
Ozhan and Kucukakcalı [13]	Cleveland	-	XGB	89.4 %	Small dataset is used. Data pre-processing is not specified. A single ML algorithm is used.
Das and Sinha [14]	Statlog	K-NN, SVM, NB, DT, LR, ANN, Voting EL	Voting EL	90.74 %	Small dataset is used. Hyperparameter tuning is not utilized.
Akyol and Atilla [15]	Statlog	GBM, NB, RF	NB	86.42 %	Small dataset is used Hyperparameter tuning is not utilized.
Jan et al. [16]	Cleveland+ Hungarian	SVM, ANN, NB, RF	Voting EL	93.00 %	Hyperparameter tuning is not utilized.
Tiwari et al. [17]	IEEE Dataport	EXC, SVM, RF, XGB Stacked EL	Stacked EL	92.34 %	Hyperparameter tuning is not utilized.
Yilmaz and Yagin [18]	IEEE Dataport	SVM, LR, and RF	RF	92.9 %	Data pre-processing is not specified.
Doppala et al. [19]	Cleveland	NB, RF, SVM, XGB	Majority Voting	88.24 %	Hyperparameter tuning is not utilized.
	IEEE Dataport			93.39 %	

2. Material and Method

2.1. Dataset Description

The IEEE Data Port is used to obtain the CVD dataset. The dataset was built by merging five previously individually available CVD datasets identified as Hungarian, Cleveland, Long Beach VA, Switzerland & Statlog. The dataset contains 1190 instances of 12 features, including 11 attribute values and one target variable [21]. The target variable of the dataset is composed of 561 samples without CVD (0) and 629 samples with CVD (1). Table 2 shows the attributes and their descriptions in the dataset.

2.2 Data Preprocessing

Data preprocessing is one of the essential steps before building predictive models. This process includes the steps that ensure that datasets are suitable for prediction models. Well-processed and organized data can significantly determine the effectiveness of the models designed [22]. There is no missing data in the data set. Data visualization provides an easier understanding of datasets. Figure 1 and Figure 2 indicate a visual representation of the categorical and numeric variables in the dataset.

Table 2: The attributes of the CVD dataset

No	Attributes	Description	Unit	Types of Attributes
1	age	in years	28-77	numerical
2	sex	gender (0=female, 1=male)	0-1	categorical
3	chest pain type	typical angina (1), atypical angina (2), non-anginal pain (3), asymptomatic pain (4)	1-4	categorical
4	resting bp s	resting blood pressure in mmHg	0-200	numerical
5	cholesterol	serum cholesterol in mg/dl	0-603	numerical
6	fasting blood sugar	(fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)	0-1	categorical
7	resting ECG	normal (0), ST-T wave abnormality (1), LV Hypertrophy (2)	0-2	categorical
8	max. heart rate	max. heart rate achieved	60-202	numerical
9	exercise-induced angina	yes (1), no (0)	0-1	categorical
10	oldpeak	ST depression	-2.6-6.2	numerical
11	ST slope	the slope of the peak exercise ST segment upsloping (1), flat (2), downsloping (3))	1-3	categorical
12	target	heart disease (1), no heart disease (0)	0-1	categorical

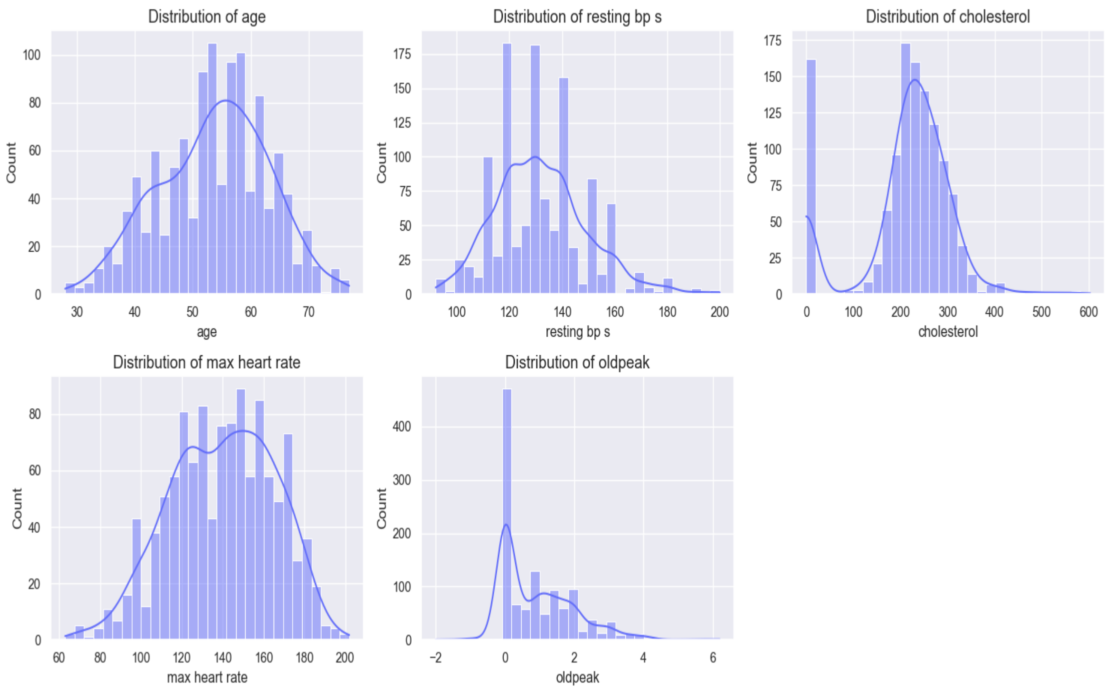


Figure 1: Numerical features

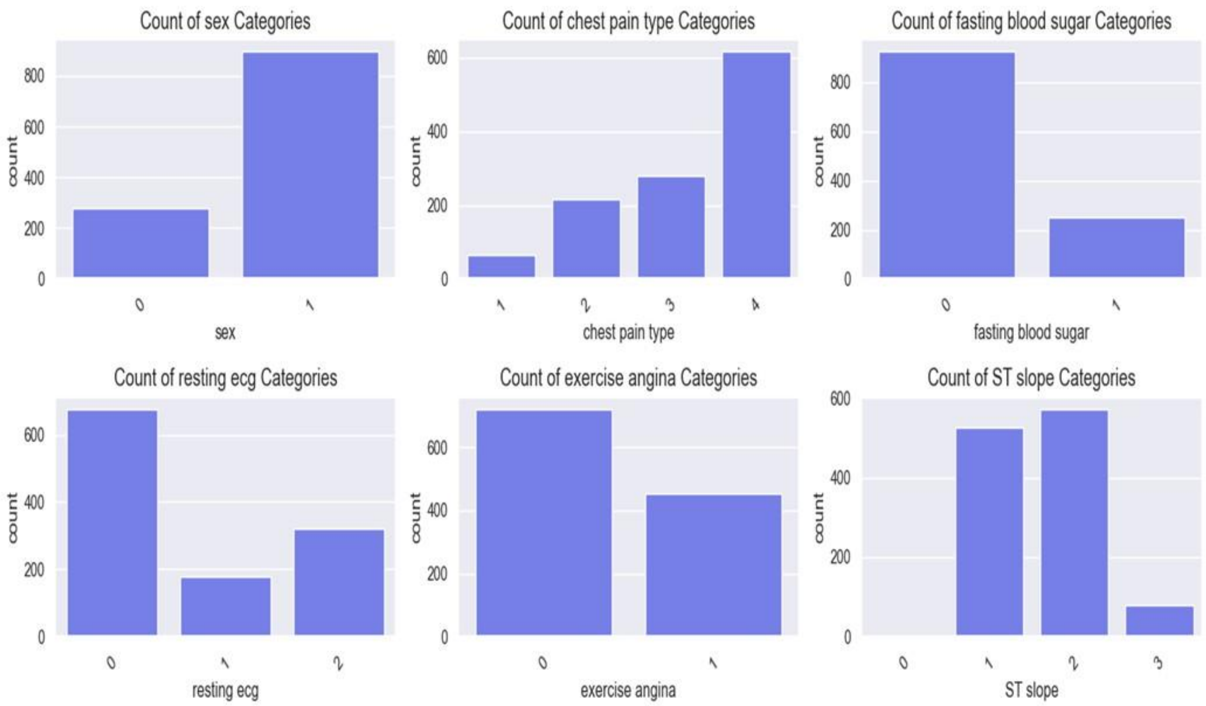


Figure 2: Categorical features

When the figures are analyzed, the following conclusions can be drawn about the dataset.

- There are five numerical and six categorical variables.
- Numerical variables in the dataset have a normal distribution.
- People with heart disease in the dataset are most frequently male.
- The most common type of chest pain is asymptomatic pain.
- Fasting blood levels are mostly below 120 mg/dl.
- ECG values are in the normal range
- Most people do not have angina.
- Most people have a flat ST slope.

In the next stage, a data transformation process is applied. The aim of this is to preserve the quality of the data and improve the data structure. The attributes are rescaled utilizing the Min-max technique between [0-1]. This technique leverages Equation 1 for rescaling. In equation 1, x_{scaled} depicts the rescaled value, x denotes the attribute value, and x_{max} and x_{min} depict the maximum and minimum attribute values [23].

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

The correlation heatmap graphically expresses the strength of relationships between numerical variables in data sets. It assists in analyzing which variables are correlated and the power of that relationship [24]. Figure 3 illustrates the correlation heatmap. Figure 4 shows the correlations observed between the target variable and the independent variables in the dataset. Accordingly, ST slope, exercise angina, chest pain type, and gender variables have a positive correlation. However, max. heart rate and cholesterol are negatively correlated with the target variable.

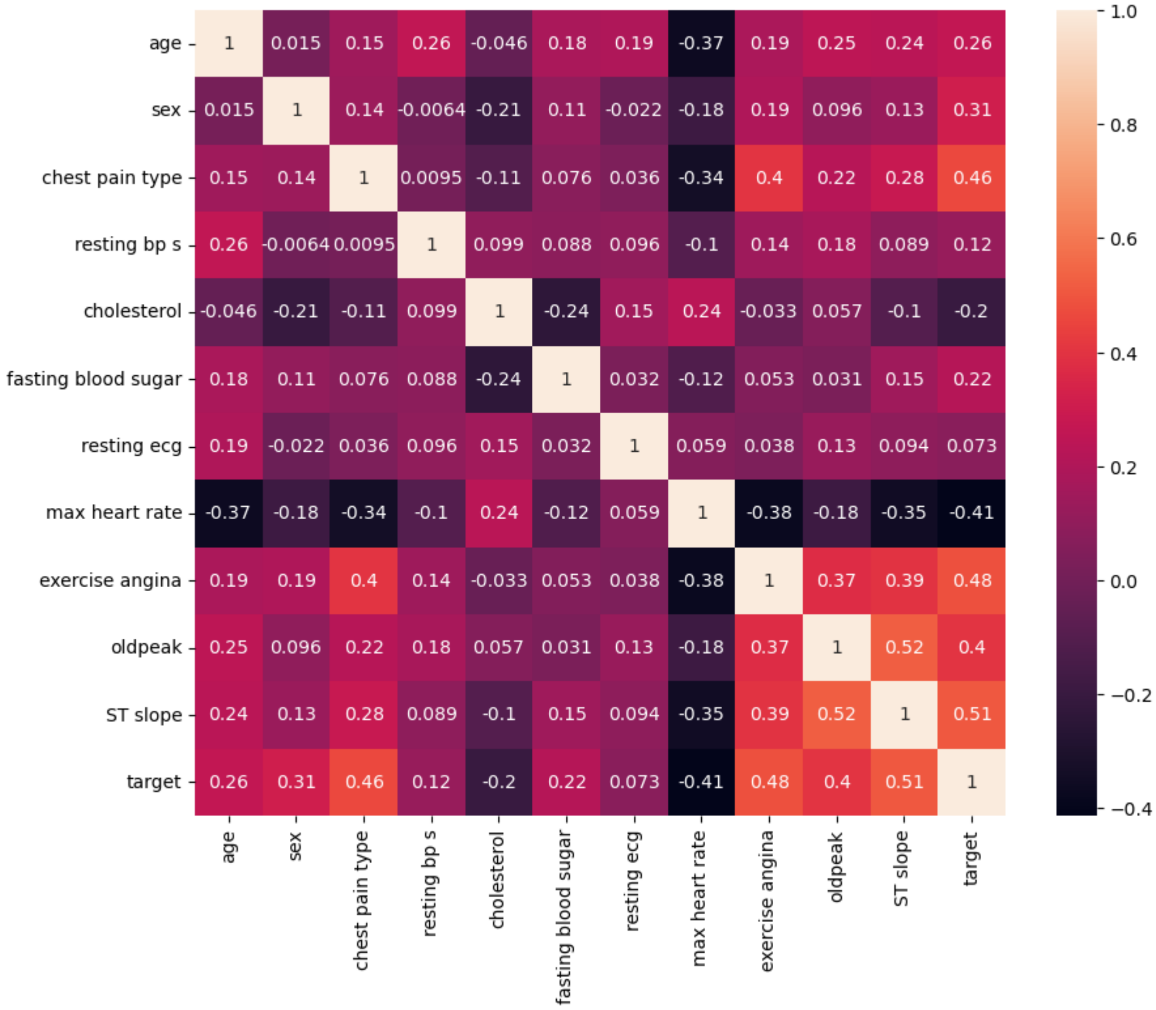
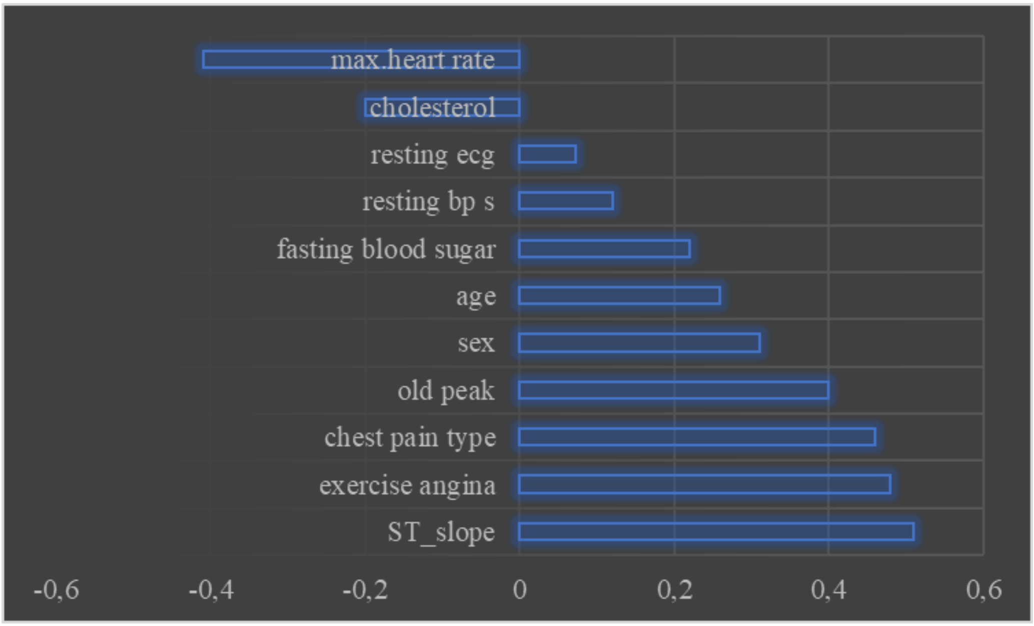


Figure 3: The correlation heatmap of all attributes in the dataset.



.

Figure 4: Correlation with target attribute

2.3 Ensemble Learning Approach

An EL technique usually acquires training data and trains models. After the training process, this technique provides testing data to the models and then each model predicts a class label for each sample in the test data. Then, each sample is put through a voting process for prediction. The voting ensemble model aims to predict an outcome based on the maximum probability of the output label that is trained on an ensemble of multiple classifiers and selected as output. The basic approach in this technique is to combine the predictions of each classifier passed to the Voting Classifier and predict the resulting label depending on the highest voting majority. Instead of building separate custom models and trying to find the accuracy for each of them, this approach is to build a single model that is trained with these models and estimates the output based on the combined majority vote for each output label. This approach provides flexibility in the combination strategy and helps to achieve the maximum possible classification accuracy. In general, two forms of voting techniques exist: Hard Voting (HV) and Soft Voting (SV). In the HV technique, each model is voted for each test case and the one that receives more than half of the votes is the final output prediction. It can be concluded that the ensemble approach does not provide a stable prediction for this problem if none of the predictions gets more than half of the votes. Instead of building discrete models and calculating accuracy for each one, a single model is built that trains on a group of multiple models and predicts the output for each output label based on the total majority of votes. In the SV technique, each classifier assigns a probability value that a given data point falls into a particular class. The predictions are then weighted by the importance of the classifier and summed. The target with the highest weighted probability sums the label wins the vote [25, 26]. In this study, soft and hard voting ensemble models are implemented on a set of algorithms including Random Forest (RF), Decision Tree (DT), K-Nearest Neighbor (K-NN), Support Vector Machines (SVM), AdaBoost (ADAB) and XGBoost (XGB) for CVD prediction. Figure 5 demonstrates the concept of a Voting Classifier, one of the EL methods that unifies multiple ML algorithms.

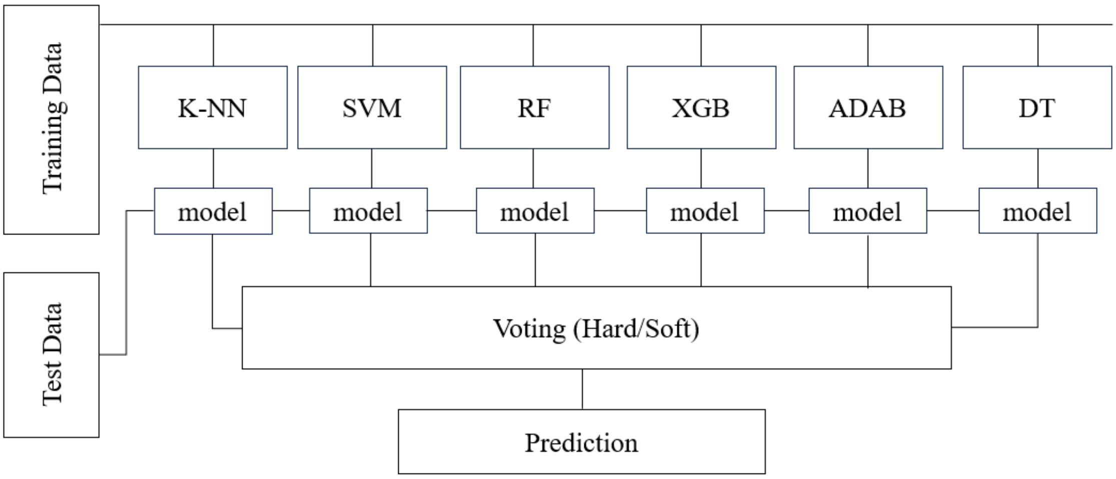


Figure 5: The workflow of the Voting Classifier

2.4 Data partition

The input dataset is partitioned into training sets and test sets utilizing a ten-fold cross-validation (CV) technique. The CV technique minimizes the chances of over-fitting and under-fitting models. The ten-fold CV technique randomly divides the dataset into ten equal subsets. Nine subsets are used for the training set and the rest for the test set. This process is iterated until each subset represents the test set. The final accuracy of each ML algorithm is decided by the mean of the accuracies derived through iterations in the model [27].

2.5 The hyperparameters of ML algorithms

The hyperparameters are parameters for ML algorithms whose values are set before training the model. Hyperparameter tuning denotes the tuning of the parameters of the model, which is a long process. This process improves the performance of ML algorithms. In this study, the Grid Search (GS) technique was used for hyperparameter tuning of the classifiers. It subdivides the space of hyperparameters into a separate grid. The performance of the model is then appraised for each combination of hyperparameters through k-fold CV. A 10-fold CV technique is used for the evaluation process. The grid point that maximizes the mean value in CV is the optimal combination of values for the hyperparameters [28]. Table 2 summarizes the hyperparameters of each ML algorithm.

Table 2: Hyperparameter tunning summary

Algorithm	Hyperparameters	Search Range	The best hyperparameter
RF	N_estimators	[10-500]	200
	Min_samples_split	[2-10]	3
	Max_depth	[2-10]	2
	Min_samples_leaf	[1-10]	2
XGB	N_estimators	[10-300]	150
	Min_samples_split	[2-10]	3
	Max_depth	[2-10]	2
	Min_samples_leaf	[1-10]	2
ADAB	N_estimators	[10-200]	[100]
	Learning_rate	[0.001-0.5]	[0.1]
K-NN	N_neighbors	[1-31]	5
DT	Max_feature	[auto, sqrt, log2]	log2
	Ccp_alpha	[0.001- 0.1]	0.01
	Max_depth	[2-10]	2
	Criterion	entropy, Gini	Gini
SVM	C	[1-1000]	100
	Gamma	[0.001-1]	0.1
	Kernel	[rbf, linear]	rbf

2.6 Model interpretation and feature importance

It is a challenging issue to comment on these models as ML algorithms are often considered a black box. SHapley Additive Explanations (SHAP) takes these algorithms out of the black box and allows comments to be made on the model. With this technique, it is possible to interpret attribute importance scores obtained after a training process and make interpretable predictions for a test instance. It provides an annotated representation of the feature value of each variable that is influential in determining the output of a ML model. It also offers the possibility of determining whether the contribution of each input characteristic is positive or negative. In this study, the Tree SHAP algorithm was used to calculate the SHAP values. For a model with prediction function f(x) and m attributes, SHAP values are obtained by Equation 2.

$$\varnothing_i = \sum_{p \subseteq N \setminus \{i\}} \frac{|p|!\,(m-|p|-1)!}{m!} [f_x(p \cup \{i\}) - f_x(p)] \qquad (2)$$

The formula expressed in Equation 1 is the sum of all possible subsets (p) of all attribute values except the i_{th} attribute value. |p|! is the number of permutations of attributes that precede the i_{th} attribute value. (m-|p|-1)! is the number of permutations of attributes that follow the i_{th} attribute value. The difference operation in the equation expresses the marginal contribution of adding the i_{th} attribute value to p [29].

3. Results and Discussion

The results and analysis of the suggested framework for the prediction of CVD have been presented in this section. The suggested framework is performed on a 64-bit machine with an 8th Generation Intel i7 CPU (16 GB DDR3 - 1 TB Hard drive 256 GB SSD). The experiments have been performed utilizing Jupyter Notebook 3.8.16 in Python containing packages such as Numpy, Pandas, Matplotlib, Seaborn, and Scikit-Learn. The Pandas 1.5.3 package was used for reading and analyzing the dataset, and The Numpy 1.23.5 package was chosen to realize mathematical functions with multidimensional arrays and matrices. The Matlplotlib 3.6.3 package was used to gain insight into how attributes are distributed in the dataset. The Seaborn 0.12.2 package was used to provide appealing and instructive data visualization graphics, especially the distribution and heatmap functions. The Scikit-Learn 1.2.2 package was utilized for splitting the dataset, model selection, and calculating statistical measures to evaluate the performance of the models. A confusion matrix is employed to measure the performance of the classifiers created within the scope of the study. Using this matrix, the performance metrics such as accuracy, precision, sensitivity, and F1-score can be calculated. The necessary mathematical expressions for these metrics are presented in Equations 3-6.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FP}} \tag{3}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5}$$

$$\text{F1} - \text{Score} = \frac{2 \text{ x Precision x Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$

Table 3 and Figure 6 summarize the results of the performance comparison of ML algorithms concerning the four performance metrics. When the classification rates of the individual classifiers are compared, the XGB and RF algorithms achieved higher classification rates with accuracies of 0.9185 and 0.9084. To improve classification performance, soft and hard voting ensemble learning models are generated by combining individual classifiers. The hard voting ensemble classifier (HVE) and soft voting ensemble classifier (SVE) outperform the baseline classifiers with accuracy of 0.9278 and 0.937, precision of 0.9387 and 0.9459, recall of 0.9253 and 0.9355 and F1-score of 0.9319 and 0.9407. Analyzing the performance of the two proposed voting ensemble models, it is seen that SVE performed better with an accuracy of 0.9370.

Table 3: Comparison of suggested voting-based models with the baseline classifier

Classifiers	**Accuracy**	**Precision**	**Recall**	**F1-Score**
K-NN	0.8431	0.8553	0.8472	0.8513
SVM	0.8580	0.8696	0.8628	0.8662
ADB	0.8706	0.8808	0.8752	0.8780
DT	0.8815	0.8919	0.8849	0.8884
RF	0.9084	0.9173	0.9101	0.9137
XGB	0.9185	0.9269	0.9196	0.9232
HVE	**0.9278**	**0.9387**	**0.9253**	**0.9319**
SVE	**0.9370**	**0.9459**	**0.9355**	**0.9407**

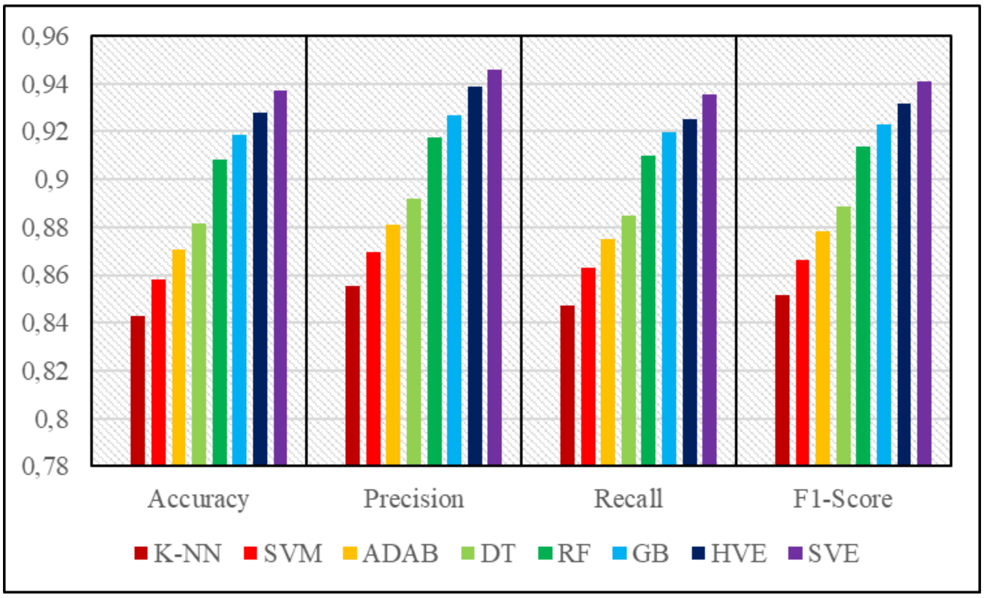


Figure 6: The comparison of classification rates in terms of performance metrics

The confusion matrix summarizes the prediction results of the proposed SVE and HVE models and is given in Figure 7 and Figure 8 respectively. Besides these metrics, the AUC-ROC curve can also be leveraged to evaluate the performance of classifiers. Figure 9 presents the AUC-ROC scores of the suggested voting ensemble techniques comparatively. Comparing the results of voting EL techniques, SVE has a higher performance rate than HVE with an AUC value of 0.936.

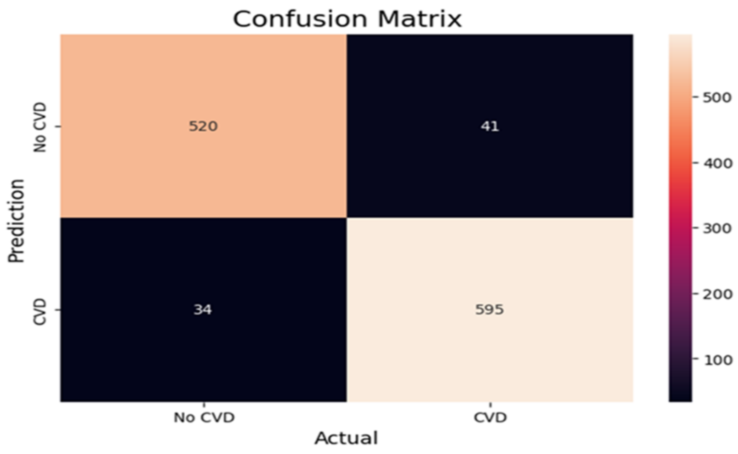


Figure 7: The confusion matrix of the suggested soft-voting ensemble model

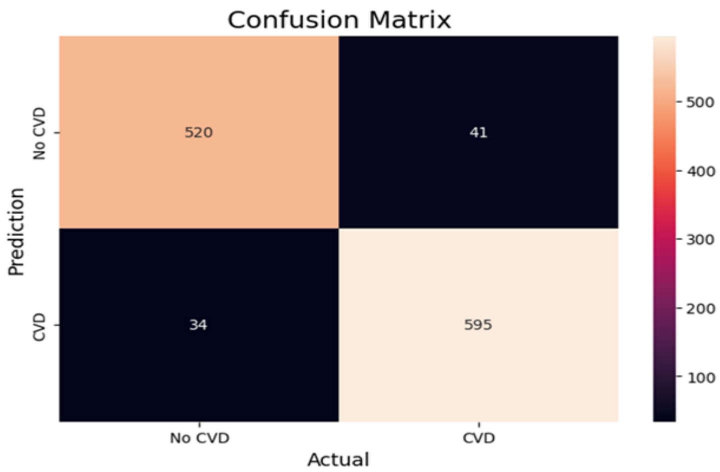


Figure 8: The confusion matrix of the suggested hard-voting ensemble model

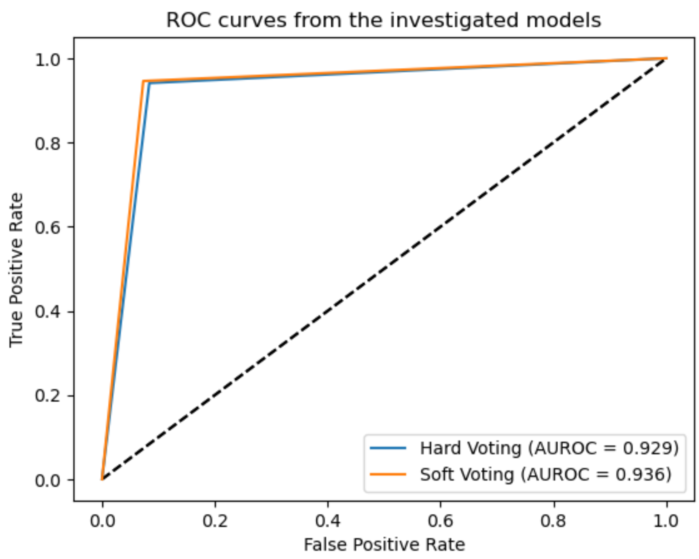


Figure 9: AUC-ROC curve for suggested voting ensemble models

With the help of Shap analysis, it is possible to learn the importance of attributes in CVD prediction and the contribution of each attribute to the accuracy of the model. Since the Tree SHAP algorithm is used, SHAP values are calculated for the XGB algorithm, which achieves the highest accuracy among tree-based algorithms. XGB is an innovative algorithm based on a decision tree and uses the method of gradient boosting in its computations. Figure 10 shows the variables evaluated by their mean absolute SHAP values. Figure 11 shows the variables in order of importance. SHAP values that hurt the predictions have a negative sign, whilst those that have a positive impact have a positive sign. The ST slope is the most significant risk factor for this disease when both graphs are examined. This means that the presence of the ST slope variable in a patient increases the patient's risk of suffering from heart disease. Moreover, the attributes with the least contribution to the CVD prediction model are "resting ecg" and "resting bp s".

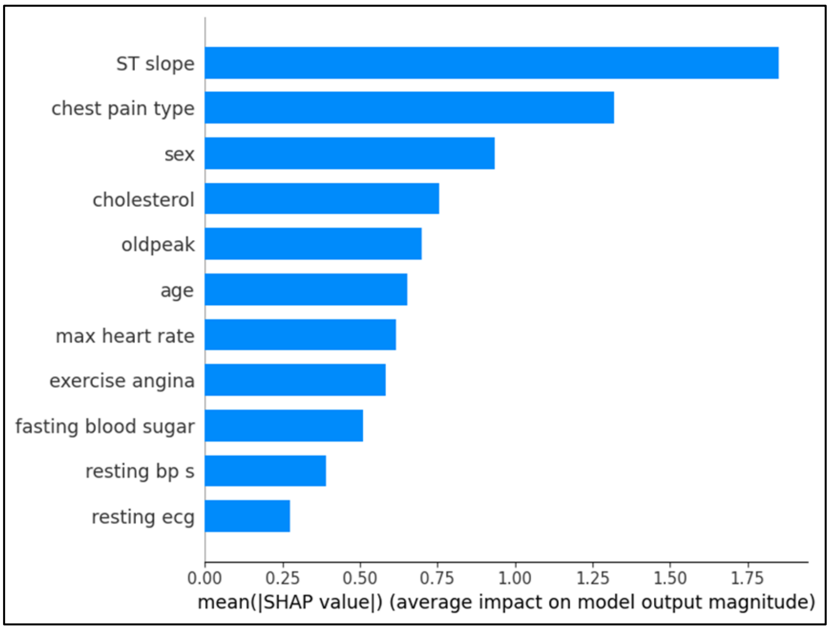


Figure 10: The attribute importance ranking according to the mean |SHAP| value

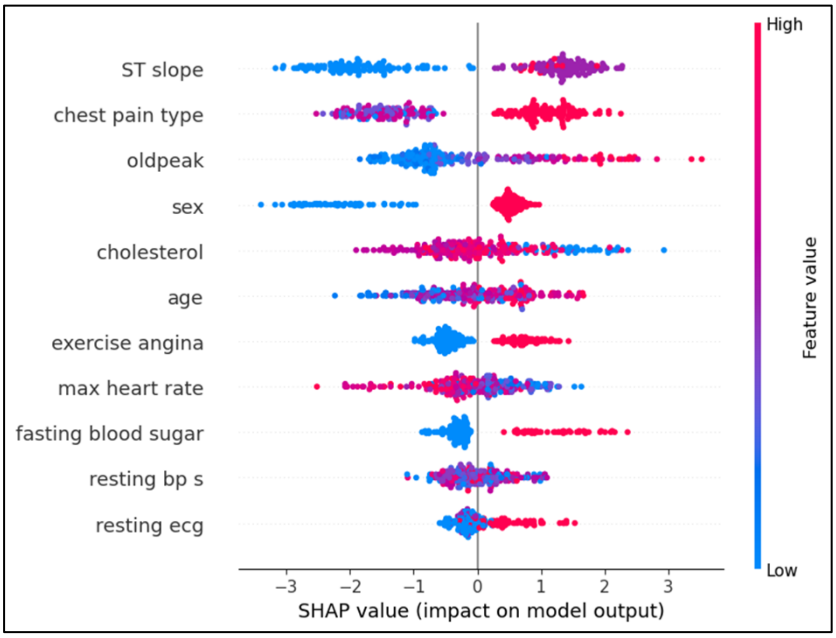


Figure 11: The attribute importance with the stability and interpretation

Table 4 provides a comparative analysis with previous similar studies for CVD prediction. As a result of the comparison with similar studies, it can be said that the SVE has a good classification rate.

Table 4: Comparison of the suggested model with some existing studies using the CVD dataset

References	Year	Dataset	Methods	Accuracy	Precision	Recall	F1-Score
Shah et al. [10]	2020	Cleveland	K-NN	0.9078	-	-	-
Rajdhan et al. [11]	2020	Cleveland	RF	0.9016	0.937	0.882	
Poorani and Hemalatha [12]	2021	Cleveland	NB	0.9033	-	-	-
Ozhan and Kucukakcalı [13]	2022	Cleveland	XGB	0.894	-	0.894	0.884
Das and Sinha [14]	2023	Statlog	Voting EL	0.9074	-	-	0.9230
Akyol and Atilla [15]	2019	Statlog	NB	0.8642	-	0.7188	-
Tiwari et al. [17]	2022	IEEE Data port	Stacked EL	0.9234	0.92	0.9349	0.9274
Yilmaz & Yagin [18]	2022	IEEE Data port	RF	0.929	-	0.928	0.928
Doppala et al. [19]	2022	Cleveland	Majority Voting	0.8824	0.85	0.90	0.88
		IEEE Data port		0.9339	0.99	0.88	0.90
Our suggested model		IEEE Data port	**Soft Voting**	**0.9370**	**0.9459**	**0.9355**	**0.9407**

4. Conclusion

In conclusion, this study presents a prediction model based on the voting ensemble technique for the detection of cardiovascular disease. Two different models, hard and soft voting, have been developed as voting ensemble models. Six different classifiers are selected as baseline classifiers. In the proposed method, GSCV is utilized to obtain the best hyperparameters of the classifiers. Moreover, the presented EL approaches have better classification rates when compared to the baseline classifiers. Among the voting techniques, the proposed SVE model achieved the highest classification rate with 0.9370 in accuracy, 0.9459 in precision, 0.9355 in sensitivity, 0.9407 in F1-score, and 0.936 AUC-ROC values. In addition, the SHAP technique is used to extract the black box structure of classifiers and to investigate the effects of variables in the dataset on the model. According to the results of the analysis, the ST slope variable is found to be the most important risk factor for this disease**.** Although voting EL is not the optimal solution for all problems, it provides a higher classification rate than individual classifiers. In the future, we intend to evaluate and test the proposed model on different datasets. Machine learning algorithms face major challenges owing to the limited amount of data. If hospitals and other data-generating organizations collaborate to obtain a larger amount of high-quality medical data, more study and research can be done.

References

[1] F. Coronado, S. C. Melvin, R. A Bell and G. Zhao, "Global Responses to Prevent, Manage, and Control Cardiovascular Diseases." *Prev Chronic Dis*, 2022, 8:19:E84.

[2] R. Hajar, "Risk Factors for Coronary Artery Disease: Historical Perspectives." Heart Views, 2017; 18(3), 109-114.

[3] J Azmi, M. Arif, M.T. Nafis, M. A. Alam, S. Tanweer, G. Wang, "A systematic review on machine learning approaches for cardiovascular disease prediction using medical big data." *Medical Engineering & Physics*, 2022, 105, 103825.

[4] K. P. Kresoja, M. Unterhuber, R. Wachter, H. Thiele, P. Lurz, "A cardiologist's guide to machine learning in cardiovascular disease prognosis prediction." *Basic research in cardiology*, 2023, 118(1), 10.

[5] S. Mohapatra, S. Maneesha, S. Mohanty, P. K. Patra, S.K. Bhoi, K. S. Sahoo and A.H. Gandomi. "A stacking classifiers model for detecting heart irregularities and predicting cardiovascular disease." *Healthcare Analytics*, 2023, 3, 100133.

[6] I.D. Mienye and Y. Sun, "A survey of ensemble learning: Concepts, algorithms, applications, and prospects." *IEEE Access*, 2022, 10, 99129-99149

[7] K.Wang et al. "Interpretable prediction of 3-year all-cause mortality in patients with heart failure caused by coronary

heart disease based on machine learning and SHAP." *Computers in Biology and Medicine*, 2021, 137, 104813.

[8] M. Ahsan and Z. Siddique, "Machine learning-based heart disease diagnosis: A systematic literature review.", *Artificial Intelligence in Medicine*, 2022, 128, 102289.

[9] Sangya W., Shanu KR, C. Bharat., Heart Attack Prediction by using Machine Learning Techniques. *International Journal of Recent Technology and Engineering* 2020;8(5):1577–80.

[10] D. Shah, S. Patel, S.K. Bharti. "Heart disease prediction using machine learning techniques." *SN COMPUT. SCI*. 2020, 1:345.

[11] Rajdhan A, Agarwal A, Sai M, Ravi D, Ghuli P. Heart disease prediction using machine learning. *International Journal of Research and Technology* 2020;9(04): 659–62.

[12] Poorani S, Hemalatha D. Machine Learning Techniques for Heart Disease Prediction. *Journal of Cardiovascular Disease Research* 2021;12(1):93–6.

[13] O. Ozhan and Z. Kuçukakcali, "Estimation of risk factors related to heart attack with XGBoost that machine learning model." *Middle Black Sea Journal of Health Science*, 2022, 8(4), 582-591.

[14] T. Das and B. B. Sinha, "A comprehensive study on machine learning methods for predicting heart disease: a comparative analysis," *8th International Conference on Computing in Engineering and Technology (ICCET 2023)*, Hybrid Conference, Patna, India, 2023, pp. 205-210.

[15] K. Akyol and U. Atilla, "A study on performance improvement of heart disease prediction by attribute selection methods.", *Academic Platform Journal of Engineering and Science*, 2019; 7-2, 174-179.

[16] M. Jan, AA Awan, MS Khalid & Salman Nisar, Ensemble approach for developing a smart heart disease prediction system using classification algorithms, *Research Reports in Clinical Cardiology*, 2018; 9: 33-45.

[17] A. Tiwari, A. Chugh, A. Sharma, "Ensemble framework for cardiovascular disease prediction." *Computers in Biology and Medicine*, 2022, 146, 105624.

[18] R. Yilmaz and F.H. Yagin, "Early detection of coronary heart disease based on machine learning methods." *Medical Records*, 2022, 4(1), 1-6.

[19] BP. Doppala, D. Bhattacharyya D, M. Janarthanan, N. Baik, "A reliable machine intelligence model for accurate identification of cardiovascular diseases using ensemble techniques." *J Healthc Eng*. 2022, 8:2022:2585235

[20] MT. García-Ordás, M. Bayón-Gutiérrez, C. Benavides et al. "Heart disease risk prediction using deep learning techniques with feature augmentation.", *Multimed Tools Appl* 2023, 82, 31759–31773.

[21] M. Siddhartha, November 5, 2020, "Heart Disease Dataset (Comprehensive)", IEEE Dataport, doi: https://dx.doi.org/10.21227/dz4t-cm36. (Accessed -10.09.2023).

[22] S. Garcia, S. Ramírez-Gallego, J. Luengo, J.M. Benítez & F. Herrera, "Big data preprocessing: methods and prospects.", *Big Data Analytics*, 2016, 1(1), 1-22.

[23] SGK Patro and KK Sahu, Normalization: A preprocessing stage. *arXiv preprint arXiv*: 2015, 1503.06462.

[24] BC. Haarman, RF. Riemersma-Van der Lek, WA Nolen, R. Mendes, HA. Drexhage, H. Burger. "Feature-expression heat maps--a new visual method to explore complex associations between two variable sets." *J Biomed Inform*. 2015, 53:156-61.

[25] S. Tewari, U.D. Dwivedi. "A comparative study of heterogeneous ensemble methods for the identification of geological lithofacies." *J Petrol Explor Prod Technol*. 2020, 10, 1849–1868.

[26] N. Chandrasekhar, S. Peddakrishna, "Enhancing heart disease prediction accuracy through machine learning techniques and optimization." *Processes* 2023, 11, 1210.

[27] Y. Xie, C. Zhu, W. Zhou, Z. Li, X. Liu, T. Tu. "Evaluation of machine learning methods for formation lithology identification: a comparison of tuning process and model performance." *J Pet Sci Eng* 2018, 60:182–193.

[28] D.M. Belete, M. D. Huchaiah. "Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results.", *International Journal of Computers and Applications*, 2022, 44:9, 875-886.

[29] S.M. Lundberg and S.I. Lee, "A unified approach to interpreting model predictions." *Advances in neural information processing systems*, 2017, 30.

Conflict of Interest Notice

The author declare that there is no conflict of interest regarding the publication of this paper.

Ethical Approval and Informed Consent

It is declared that during the preparation process of this study, scientific and ethical principles were followed, and all the studies benefited from are stated in the bibliography.

Availability of data and material

Not applicable

Plagiarism Statement

This article has been scanned by iThenticate ™.