Enhancing Cardiovascular Disease Prediction Using Hard Voting Technique in Machine Learning

R.Gowthamani[1], K.Sasi Kala Rani[2], V.C.Ambarish[3], J.Binesh[4], C.R.Jayanth[5], B.Jeba Regan Raj[6]

[1]Head of the Department, Computer Science and Engineering, SKCET
[2]Associate Professor, Computer Science and Engineering, SKCET
[3,4,5,6]Student, Computer Science and Engineering, SKCET

Abstract

Cardio Vascular Disease (CVD) or Heart disease is one of the leading causes of death around the globe. Early identification of the disease can significantly save precious lives. But the identification of heart-related diseases is a challenging task as it relies on a wide range of factors. Machine Learning algorithms have strong potential in prediction-related domains. In this paper, we have used an Ensembled model called the Hard Vot-ing Ensemble Model to detect heart disease. A dataset containing 13 features is taken from the UCI repo using Kaggle. Seven different algorithms are used, tested, and trained, accuracy is measured and out of those, models with the best accuracy are picked and ensembled together. The ensemble model resulted in higher accuracy than all other individual models.

Keywords: Machine Learning, Supervised Learning, Naive Bayes, Logistic Regression, Random Forest, Extreme Gradient, K-Nearest Neighbors, Decision Tree, SVM, Hard Voting Ensemble Technique.

1. Introduction

The largest cause of death worldwide, heart-related illnesses (CVD) claim 17.9 million lives annually. CVD refers to a range of diseases that affect the heart namely coronary artery disease, heart failure, arrhythmias, and so on. The most prevalent kind of heart illness is coronary artery disease, also known as atherosclerosis, which happens when fatty substances constrict or block the arteries that carry blood to the heart. When the heart is unable to pump blood effectively, heart failure occurs. Arrhythmias refers to abnormal heartbeats. More than four CVD fatalities out of every five are attributable to heart attacks and strokes. Heart disease risk factors include a poor diet, a lack of exercise, cigarette use, and alcohol abuse. Heart disease identification is an extremely difficult task. Many factors have to be taken into consideration to predict the result. It will be a challenging task for humans to make such a pre-diction which involves such a complex process. Thankfully, in recent times Machine Learning has emerged in many fields especially in making predic-tions. It would be an optimal solution for this kind of scenario. In our solution, a type of Machine Learning technique called the Hard Voting Ensemble method is to predict CVD with high accuracy.

2. Literature Survey

Machine Learning has already made a great impact in this field. Various algorithms were trained, and tested and have yielded some great accuracy. TABLE 1 shows the algorithms and their accuracy achieved by various works.

TABLE 1. Comparison of various algorithms

Year	Author	Algorithm	Result
2020 [1]	Archana Singh, Rakesh Kumar	1. SVM 2. Decision Tree 3. Linear Regression 4. K-Nearest Neighbor	87% - K-Nearest Neighbor
2019 [2]	S.J. Krishnan, S. Geetha	1. Naïve Bayes 2. Decision Tree	91% - Decision Tree
2019 [3]	AvinashGolande	1. K-Mean Clustering 2. K-Nearest Neighbor 3. Decision Tree	86.6% - Decision Tree
2019 [4]	M.Marimuthu, S. Deivarani, R. Gayathri	1. SVM 2. Decision Tree 3. Naive Bayes 4. K-Nearest Neighbor	83.6% - K-Nearest Neighbor
2018 [5]	Abhay Kishore	1. CNN 2. Decision Tree 3. SVM 4. KSOM 5. DMNRNN	92% - RNN

3. Existing System

The performance of any algorithm depends on the dataset fed into that. A noisy dataset will affect the accuracy of the algorithm. Also having features whose contributions are significantly less to the target class will also reduce the accuracy. So it is crucial to select the features via a thorough examination before feeding them into the model. [1] have used the UCI dataset consisting of 13 features and 1 target class which refers if the instance had heart disease or not. All those 13 features were taken into consideration for testing and training the model. Supervised learning is carried out on four algorithms namely Support Vector, Decision Tree, Linear Regression, and K-Nearest Neighbor. The highest accuracy was 87%, produced by the KNN algorithm. By considering the above facts, the accuracy of the algorithms can be improved by selecting optimum features and eliminating the others. Also, ensembling might have yielded even higher accuracy.

4. Proposed System

The proposed system uses the Hard Voting Ensemble technique to predict CVD effectively. The dataset is collected from the UCI repo which is verified by several researchers and authorities of UCI. It consists of 13 features. Preprocessing is done to remove noisy, empty, null values in the dataset. Feature Selection is carried out using the Filter method with the help of the correlation matrix. Finally, a dataset consisting of 12 features (removing the FBS feature) is used in building the model instead of using all the features. The dataset is divided into 8:2 ratios for training and testing the model. Supervised Learning is carried out on various algorithms namely Naive Bayes, Logistic Regression, Random Forest, Extreme Gradient, K-Neighbors, Decision Tree, and SVM and they are trained, tested, and their accuracy is noted down. Among those models, Random Forest, Decision Tree, and SVM models gave promising accuracy of 93.65%, 91.95%, and 92.68%. Hard Voting Ensembling is implemented on those three models and its accuracy is tested. The Ensembled model produced an accuracy of 95.12% which is higher than the accuracy of all individual models.

5. Proposed System Modules

5.1. Data Preprocessing:

Using the Kaggle website, the dataset of 13 features is downloaded in CSV format from the UCI repository. TABLE 2 shows the features of the dataset and their description. The CSV file containing the dataset is imported and cleaned. Empty spaces and duplicate entries are eliminated, and null values are substituted with the column's mean. Data Preprocessing is a crucial step to remove noisy data which prevents the model from recognizing wrong patterns.

TABLE 2.Features of the Dataset

S.No	Feature	Description
1	age	Human’s age
2	sex	Person’s gender
3	cp	Chest pain type
4	trestbps	Resting blood pressure
5	chol	Serum cholesterol in mg/dl
6	fbs	Fasting blood sugar
7	resting	Resting ECG results
8	thalach	Maximum heart rate achieved
9	exang	Exercise induced angina
10	oldpeak	Exercise-induced depression relative

		to rest
11	slope	The slope of the peak exercise
12	ca	Number of major vessels
13	thal	0: normal, 1: fixed, 2: defect is reversible
14	target	The person has CVD or not

5.2. Feature Selection:

It is a process that helps in identifying optimum sets of features essential for prediction [6]. The filter method is used in our solution to select the best features. Using a correlation matrix and some trial and error methods, among 13 features one feature namely FBS had significantly less contribution towards the target class. That feature is removed from the dataset. All other features are kept as it was as they had a good correlation with the target class. The final dataset consisting of 12 features is fed into the model. Figure.1 correlation matrix shows how the features and related to the target class.

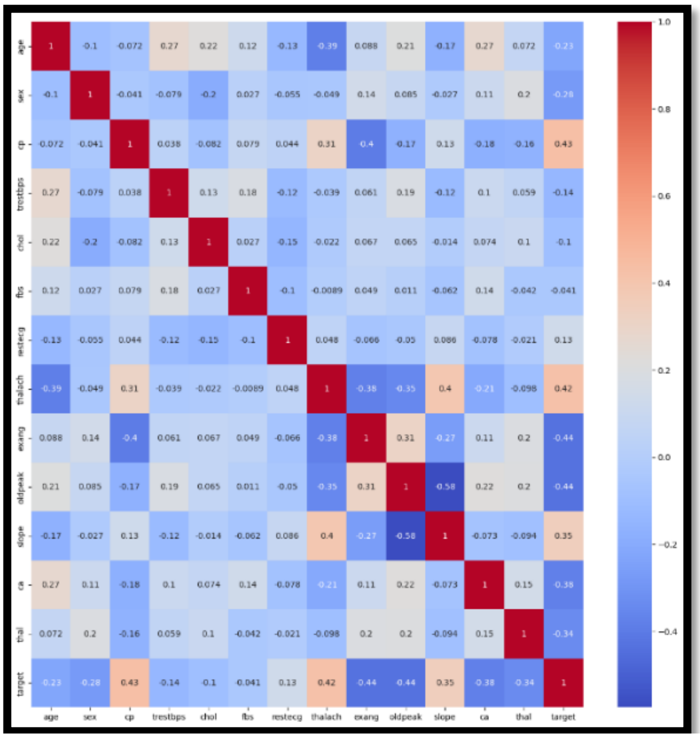


Figure 1. Correlation matrix of Cleveland dataset before feature selection

5.3. Splitting the dataset and Balancing the target class :

The entire dataset is divided into 8:2 ratios to train and test the models. The datasets should consist of an equal amount of target classes (CVD, No CVD) which makes the model predict both the target classes effectively with consistent accuracy. Figure 2 depicts the dataset's approximately 800 target classes distributed evenly.

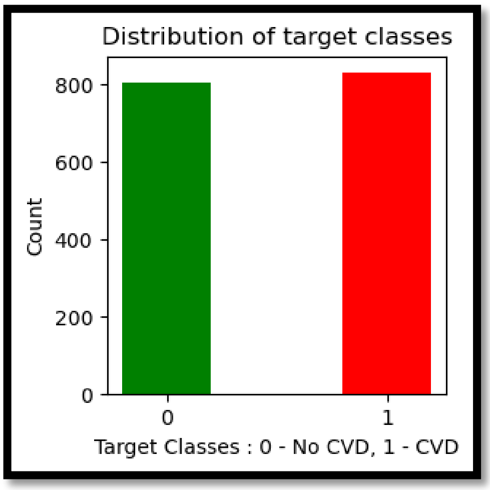


Figure 2. Distribution of target classes in the dataset

5.4. Selecting the Models:

Seven supervised learning algorithms namely Naive Bayes, Logistic Regression, Random Forest, Extreme Gradient Boost, K-Neighbors, Decision Tree, and SVM are trained using the training dataset. Then they are tested with the Test Dataset to calculate the accuracy.

5.5. Accuracy Calculation :

The accuracy of the algorithms was calculated using four values namely True Positive(TP), False Positive(FP), True Negative(TN), and False Negative(FN) which are present in the confusion matrix.

Accuracy = (FN + TP) / (TP + FP + TN + FN)

The confusion matrix is used to determine the values for the expression above. Testing the model using a test dataset yields a confusion matrix. Their accuracy score is compared to find the models with higher accuracy. Random Forest, Decision Tree, and Support Vector Machines have higher accuracy than others with an accuracy of 93.65%, 91.95%, and 92.68%.

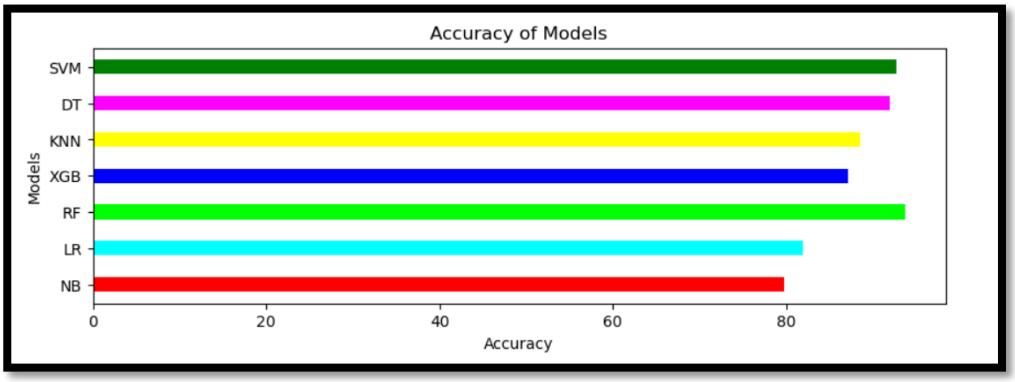


Figure 3. Accuracy of the ML models

5.6 Ensembling the Models using the Hard Voting method:

In the Hard Voting Ensemble method[7], the classification process is based on the models' consensus vote. The same input is fed into all the models and each model votes for a target class i.e make a prediction. The target class with the majority vote is the final output. Hard Voting Ensembling is carried out between various models. Of those, ensembling the Random Forest, Decision Tree, and Support Vector Machine models produced the highest accuracy of 95.12%. That model is finalized.

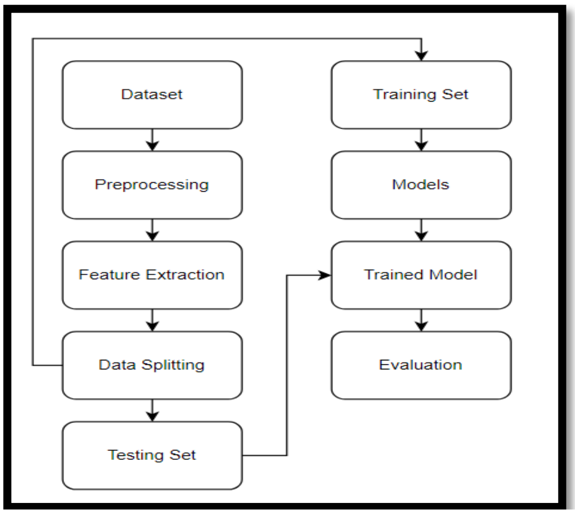
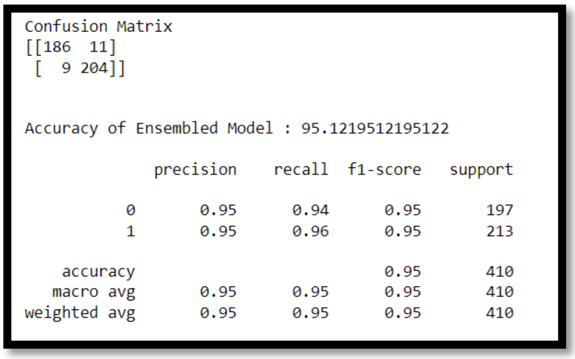


Figure 4. Flowchart of Supervised Learning



```
Confusion Matrix
[[186  11]
 [  9 204]]


Accuracy of Ensembled Model : 95.1219512195122

              precision    recall  f1-score   support

           0       0.95      0.94      0.95       197
           1       0.95      0.96      0.95       213

    accuracy                           0.95       410
   macro avg       0.95      0.95      0.95       410
weighted avg       0.95      0.95      0.95       410
```

Figure 5. Accuracy and Confusion matrix of the Ensemble

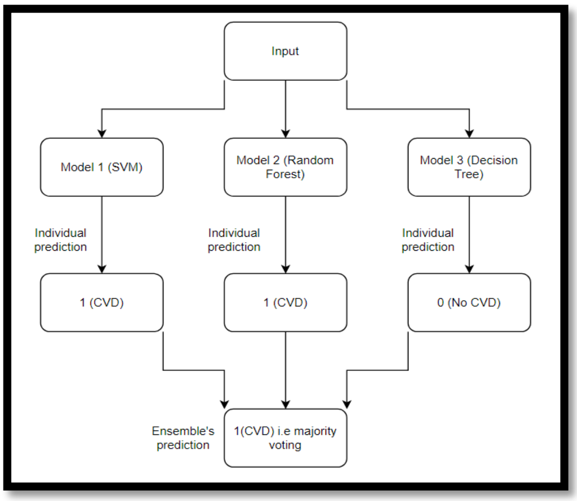


Figure 6. Hard Voting Ensemble technique

6. Conclusion

In this paper, the Hard Voting Ensemble method is used to predict heart diseases. The precision of the model was greatly increased by careful feature selection. Of all other methods, the SVM, Decision Tree, and Random Forest model assembly produced the best accuracy. the accuracy of the ensembled model is calculated using the values from the confusion matrix, It concludes that the Ensembled model produced using SVM, Decision Tree, and Random Forest gives 95.12%. accuracy.

References

1. A. Singh and R. Kumar, "Heart Disease Prediction Using Machine Learning Algorithms," 2020 International Conference on Electrical and Electronics Engineering (ICE3), Gorakhpur, India, 2020, pp. 452-457, doi: 10.1109/ICE348803.2020.9122958.
2. S. K. J. and G. S., "Prediction of Heart Disease Using Machine Learning Algorithms," 2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT), Chennai, India, 2019, pp. 1-5, doi: 10.1109/ICIICT1.2019.8741465.
3. AvinashGolande, Pavan Kumar T "Heart DiseasePrediction Using Effective Machine Learning Techniques", International Journal of Recent Technology and Engineering (IJRTE) ISN: 2277-3878, Volume-8, Issue-1S4, June 2019.
4. Marimuthu, M. et al. "Analysis of heart disease prediction using various machine learning techniques." Advances in Computerized Analysis in Clinical and Medical Imaging. Chapman and Hall/CRC, 2019. 157-168.
5. Kishore, Abhay, et al. "Heart attack prediction using deep learning." International Research Journal of Engineering and Technology (IRJET) 5.04 (2018): 2395-0072.
6. A. Newaz and S. Muhtadi, "Performance Improvement of Heart Disease Prediction by Identifying Optimal Feature Sets Using Feature Selection Technique," 2021 International Conference on Information Technology (ICIT), Amman, Jordan, 2021, pp.446-450, doi:10.1109/ICIT52682.2021.9491739.
7. R. Atallah and A. Al-Mousa, "Heart Disease Detection Using Machine Learning Majority VotingEnsemble Method," 2019 2nd InternationalConference on new Trends in Computing Sciences(ICTCS), Amman, Jordan, 2019, pp. 1-6, doi:10.1109/ICTCS.2019.8923053.