An Ensemble Learning based Web Application to predict the risk of Heart Disease

Venna Vinay Ranjan Adithya[a], Kiranmai Battu[b], Dr. Shobana M[c]

[a] Department of Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur, India
Email:- vvraditya01@gmail.com
[b]Department of Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur, India
[c] Assistant Professor,Department of Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur, India

Article History: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 10 May 2021

Abstract: With an approximate 17.9 million annual victims, heart disease stands out as a prominent cause of deaths worldwide, whose fatality can be reduced to a great extent with an expeditious diagnosis. Herein we propose an ensemble learning-based heart disease prediction system. The UCI Heart Disease dataset has been utilized in this work. Relevant data mining methodology has been adopted to create six predictive models. Appropriate hyperparameters were optimized with the help of GridSearchCV along with 5-fold cross-validation. Recall value and ROC score were the performance metrics considered relevant to judge the performance of the models. The best performing models were picked to create a heterogeneous ensemble. The proposed ensemble produced a ROC score of 0.84 and a recall value of 0.94. The suggested ensemble has been found to enhance the predictive capabilities of the classic algorithms

Keywords: Ensemble learning, Heart disease diagnosis, Recall, Receiver Operating Characteristic curve

1. Introduction

Several clinical conditions that emerge due to impedance of the heart and veins are collectively called Heart diseases(also known as Cardiovascular disease CVDs). CVDs account for a significant 31% of all passings across the globe. The statistics of the World Health Organization reveal that a third of all deaths due to various CVDs occur in low-earning and moderate-earning countries. Most of these deaths could have been averted with an early diagnosis and timely administration of medicine. With the advent of COVID-19, experts believe the global burden of cardiovascular disease will grow exponentially in the coming decade. Therefore, research of risks posed by CVDs and their early detection is of great significance.

With the rapid surge in the collection of medical data, researchers have begun to explore new ways to enhance patient prognosis. In the last decade, practitioners have made extensive use of knowledge discovery(popularly called Data Mining) and machine learning in the clinical environment. This approach reveals valuable insights from raw data that help boost the decision-making support. A few sought-after data mining algorithms are K-Nearest Neighbors, Logistic Regression, etc. How well these strategies perform is largely influenced by the kind of data we intend to mine and the inherent features of the algorithms. Customizing our model as per the conditions of operation will improve the results of these models.

Ensemble learning is one such proven technique that improves classification tasks. An ensemble combines the predictive capabilities of several base learners to create a better-performing optimal classifier. To aggregate the predictions made by the base learners, ensembles are equipped with a mechanism like hard voting, soft voting, etc. A heterogeneous ensemble is a blend of various algorithms (also called base learners). Whereas, a homogeneous ensemble is composed of a single algorithm. An ensemble works on the principle of “Wisdom of the crowd” and often outperforms several traditional algorithms.

2. Literature Survey

Commendable work has been done in the application of various knowledge discovery methodologies upon medical datasets to facilitate early prognosis of various ailments e.g., prognosis of several respiratory illnesses, cardiac arrest (Metsker et al., 2018, 351-358), cancer prognosis, etc. In spite of this progress, researchers are working meticulously to develop new strategies and techniques to enhance classification performance. Ensemble learning is one such strategy. Bagging and boosting were the first known application of ensemble learning. Different approaches were followed by various scholars to achieve ensemble learning.

Researchers randomly split the dataset using a mean-based partitioning technique to create homogeneous subsets. The tree stops growing when the height of the tree (H) is equal to the preset maximum tree height (H_{max}) i.e., $H = H_{max}$, or when the quantity of samples within the ith subpartition (N_i) is not more than the specified number of samples (N_{min}) i.e., $N_i \leq N_{min}$. Classification and Regression Tree has been applied to model the several subpartitions. Weights were assigned taking into consideration the accuracy of each CART. An ensemble classifier is then created to make a prediction. (Mienye et al., 2020)

Scholars carried out a detailed study on ensemble classification. Bayes Net, Multilayer Perceptron, and PART were the algorithms that were considered for the study in addition to Naive Bayes, C4.5 algorithm, and Random Forest. The effects of bagging, boosting, and stacking on the performance of these classifiers were thoroughly analyzed. On average, a 7% increase in accuracy has been achieved by the resulting ensemble than any other algorithm. (Latha C & Jeeva S, 2019)

Investigators developed a single-layer feedforward neural network ensemble. A backpropagation learning algorithmic rule was executed within the feedforward. Scaled conjugate gradient (SCG) was the variation employed in the ensemble, in company with Poole–Ribiere conjugate gradient (CGP). Besides Levenberg–Marquardt (LM) was also employed. The ensemble analyzes the probabilities for class targets from the precursor models and scores the data. (Das et al., 2009)

3. Proposed System

1.*An account of the dataset:* The dataset that has been made use of in this study is the "Heart Disease" dataset. The original dataset has 74 predictors from several invasive and non-invasive tests.

Clinical Variables
● Age ● Sex ● CP ● TRESTBPS
Routine Test Data
● CHOL ● FBS ● RESTECG
Electrocardiography Test
● THALACH ● EXANG ● SLOPE ● OLDPEAK
Non-Invasive Test
● THAL ● CA

Table 1. Predictors in the dataset

The output class was labeled as the TARGET.

2.Pre-processing of the dataset:

2.1.The dataset has a few categorical variables, namely CP, SLOPE, CA, THAL, RESTECG. These were handled using One Hot Encoding. This creates dummy variables that represent various categories of the feature. It can take a value of {0,1} that represents either the presence or absence of that particular category of the feature. RESTECG, for instance, can take values {0,1,2}. Upon application of one-hot encoding, three new columns RESTECG_0, RESTECG_1, and RESTECG_2 are created. A RESTECG value of 1 is now represented as 0,1,0 indicating the presence of RESTECG_1 and the absence of all other two columns.

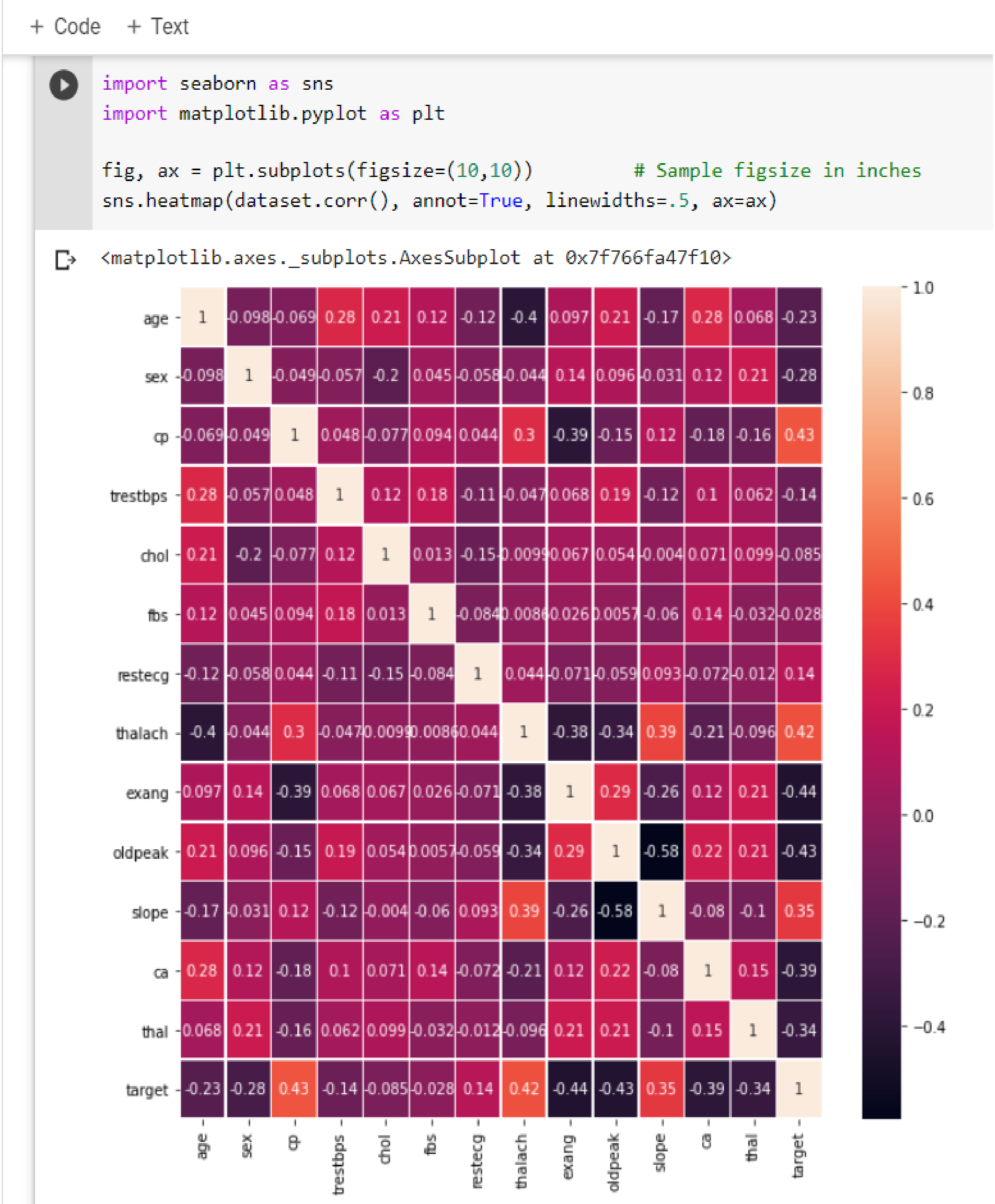
2.2.The numerical attributes were measured using various scales. For example, the maximum value of age is 77 whereas for cholesterol it is as much as 564. All the numerical features were standardized.

$$\hat{x} = \frac{x-\underline{x}}{\sigma} \qquad (1)$$

3.Feature Representation :

3.1.Feature extraction: None.

3.2.Feature selection: A reliable dataset must have independent features (without multicollinearity) to produce optimal results. The correlation matrix was analyzed to ensure the best features were used for representation.



```
import seaborn as sns
import matplotlib.pyplot as plt

fig, ax = plt.subplots(figsize=(10,10))          # Sample figsize in inches
sns.heatmap(dataset.corr(), annot=True, linewidths=.5, ax=ax)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f766fa47f10>
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
age	1	-0.098	-0.069	0.28	0.21	0.12	-0.12	-0.4	0.097	0.21	-0.17	0.28	0.068	-0.23
sex	-0.098	1	-0.049	-0.057	-0.2	0.045	-0.058	-0.044	0.14	0.096	-0.031	0.12	0.21	-0.28
cp	-0.069	-0.049	1	0.048	-0.077	0.094	0.044	0.3	-0.39	-0.15	0.12	-0.18	-0.16	0.43
trestbps	0.28	-0.057	0.048	1	0.12	0.18	-0.11	-0.047	0.068	0.19	-0.12	0.1	0.062	-0.14
chol	0.21	-0.2	-0.077	0.12	1	0.013	-0.15	-0.0099	0.067	0.054	-0.004	0.071	0.099	-0.085
fbs	0.12	0.045	0.094	0.18	0.013	1	-0.084	0.0086	0.026	0.0057	-0.06	0.14	-0.032	-0.028
restecg	-0.12	-0.058	0.044	-0.11	-0.15	-0.084	1	0.044	-0.071	-0.059	0.093	-0.072	-0.012	0.14
thalach	-0.4	-0.044	0.3	-0.047	0.0099	0.0086	0.044	1	-0.38	-0.34	0.39	-0.21	-0.096	0.42
exang	0.097	0.14	-0.39	0.068	0.067	0.026	-0.071	-0.38	1	0.29	-0.26	0.12	0.21	-0.44
oldpeak	0.21	0.096	-0.15	0.19	0.054	0.0057	-0.059	-0.34	0.29	1	-0.58	0.22	0.21	-0.43
slope	-0.17	-0.031	0.12	-0.12	-0.004	-0.06	0.093	0.39	-0.26	-0.58	1	-0.08	-0.1	0.35
ca	0.28	0.12	-0.18	0.1	0.071	0.14	-0.072	-0.21	0.12	0.22	-0.08	1	0.15	-0.39
thal	0.068	0.21	-0.16	0.062	0.099	-0.032	-0.012	-0.096	0.21	0.21	-0.1	0.15	1	-0.34
target	-0.23	-0.28	0.43	-0.14	-0.085	-0.028	0.14	0.42	-0.44	-0.43	0.35	-0.39	-0.34	1

Figure 1. Correlation Matrix

4.The dataset has been partitioned. The ninetieth portion of all instances constitutes the training dataset, whereas the remaining instances make up the testing dataset. While doing so, the proportional data quantities for both the classes have been maintained in both the training and testing datasets.

5.Machine Learning and Model Selection:

5.1.K-Nearest Neighbours: KNN is perhaps the least complex algorithm that assumes the similarity between the new data and the available data. The new data are classified into the category with the most similarity. The optimal value for K has been found using GridSearchCV coupled with 5-fold cross-validation.

5.2.Random Forest: This is a versatile algorithm that can be used for both classification and regression. The number of trees to be built in the forest has been found using GridSearchCV coupled with 5-fold cross-validation.

5.3.Naive Bayes: Naive Bayes is a traditional probabilistic classifier. It makes a classification based on the probability of the object.

5.4.Logistic Regression: Logistic Regression fits an S-shaped curve called the sigmoid curve that predicts the value between 0 and 1.

Prediction = g(z) = $\frac{1}{1+e^{-z}}$ (2)

The variable z is an amalgamation of all input variables utilized in the model.

$Z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$ (3)

Where $\beta_0$is the intercept and $\beta_1, \beta_2, \ldots \beta_n$are the regression coefficients.

5.5.Support Vector Machine: SVM is another popular classification technique that requires less computing power and has a significant accuracy.

5.6.Decision Tree: Every internal node in a decision tree is a representation of a predictor being tested. Each branch that originates from the parent node is a test result. The leaf node stands for the class label.

6.*Ensemble creation:* After analyzing the ROC curves of the models, KNN, SVM, Random Forest, Logistic Regression, and Decision Tree were considered to be used in the ensemble. Every one of these models predicts the output class label, i.e. {0,1} when provided with an instance variable. But in the soft voting technique, the probability of an instance belonging to class 1 as predicted by each precursor is taken into account. An average of all these probabilities is calculated. Class 1 is the predicted output if the calculated mean probability exceeds 0.5. Else class 0 is the output. 0.5 is a default value and is called the threshold value. We can manipulate the type-1 error (otherwise called False Positive) as well as the type-2 error (otherwise called False Negative) by adjusting the threshold by either decreasing or increasing its value. In our model, the threshold value remains as such, i.e., at 0.5.

4. System Architecture

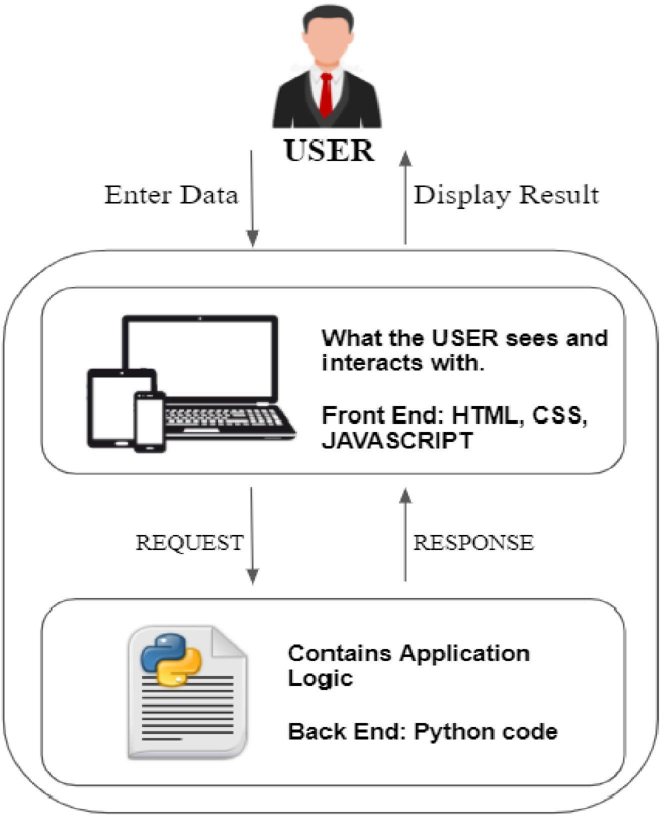

USER
Enter Data
Display Result
What the USER sees and interacts with.
Front End: HTML, CSS, JAVASCRIPT
REQUEST
RESPONSE
Contains Application Logic
Back End: Python code

Figure 2. Web Application Architecture

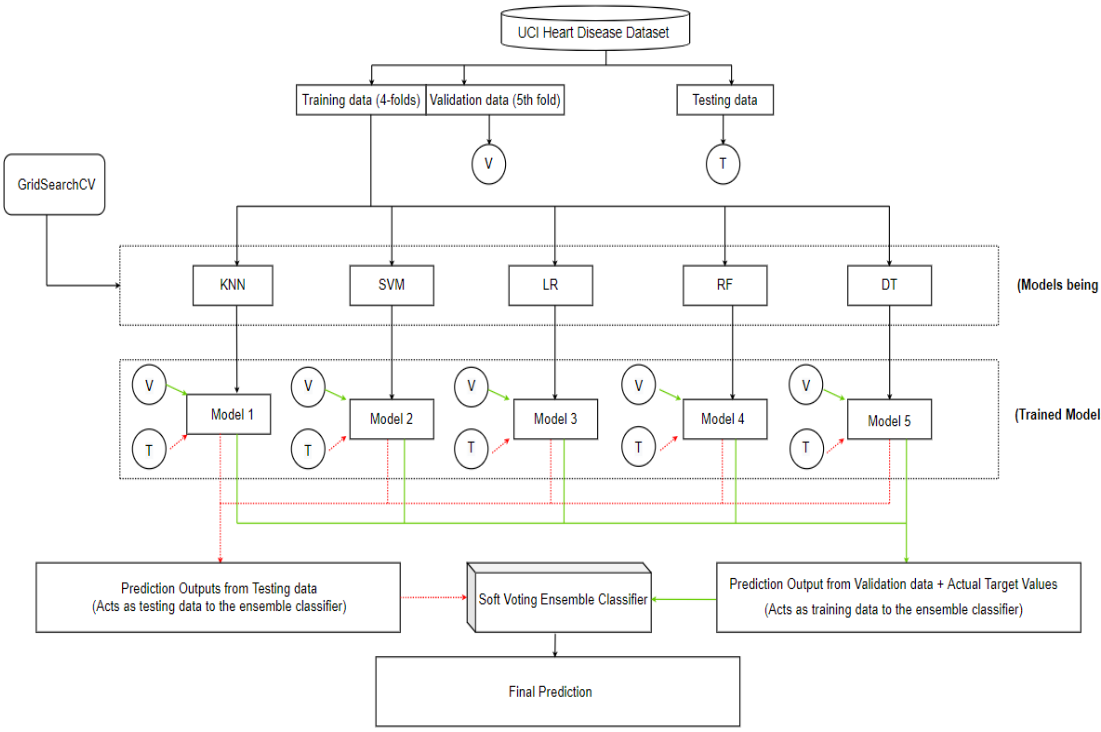

UCI Heart Disease Dataset
Training data (4-folds)
Validation data (5th fold)
Testing data
V
T
GridSearchCV
KNN
SVM
LR
RF
DT
(Models being trained)
V
T
Model 1
Model 2
Model 3
Model 4
Model 5
(Trained Models)
Prediction Outputs from Testing data
(Acts as testing data to the ensemble classifier)
Soft Voting Ensemble Classifier
Prediction Output from Validation data + Actual Target Values
(Acts as training data to the ensemble classifier)
Final Prediction

Figure 3. Ensemble Architecture

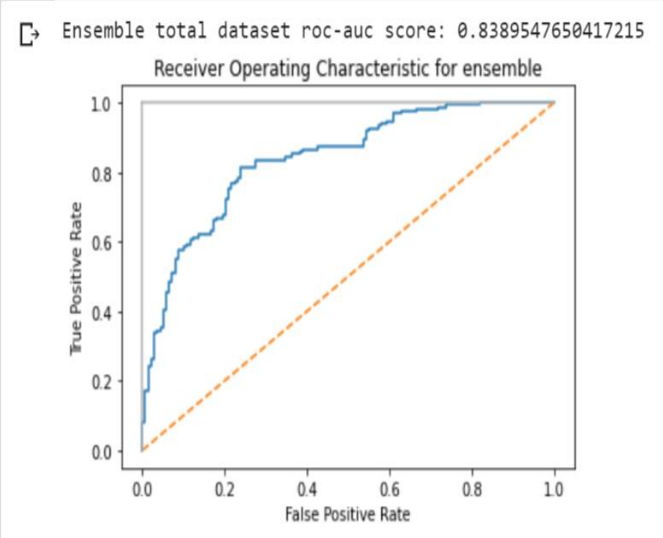
5. Results And Discussions

The prime motive of this project is to develop a first-level diagnosing application that can identify the heart disease risk and minimize the type-2 error, i.e., False Negative.

To evaluate and analyze the efficiency of our model with the precursors, the performance metrics deemed important were recall value and ROC score.

Figure 4 depicts the ROC curve of the ensemble. This is essentially a plot of the various false-positive rates against their corresponding true positive rate. The ROC curve facilitates the selection of an optimal threshold value with a minimum false positive rate and a significantly high true positive rate. The proposed ensemble has achieved a ROC-AUC score of 0.84.

Furthermore, a comparison is carried out among all the models. Various metrics were preferred for the task.



```
Ensemble total dataset roc-auc score: 0.8389547650417215
```

Figure 4. ROC-AUC curve of the ensemble

Figure 5 is a representation of the ROC curves of various models. The proposed ensemble(shown in brown) clearly has a greater area under the curve and outperforms all the traditional algorithms.

Table 2 provides a glimpse of the scores achieved by various models whilst considering several performance metrics.

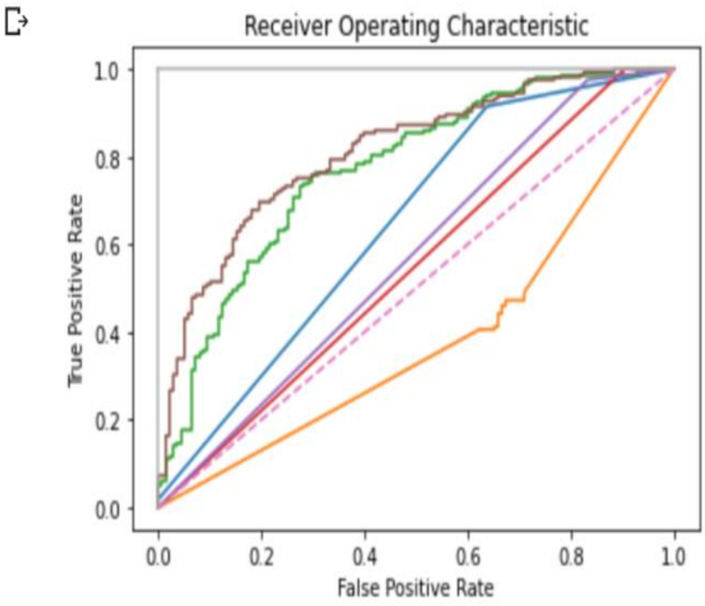


Figure 5. Performance of ensemble (brown curve) as compared to other models

Algorithm	Accuracy	Recall	F1-score	ROC-AUC
KNN	87	80	83	87
RF	74	80	82	85
NB	90	88.8	74	91
LR	80	79	80	84
SVM	80	79	75	84
DT	74	73	63	73
Ensemble	84	94	84	87

Table 2. Performance comparison of various algorithms

Performance Metrics:

True Positive (TP): The user originally suffers from heart disease and is justly classified as positive.

True Negative (TN): The user originally does not suffer from heart disease and is justly classified as negative.

False Positive (FP): The patient is falsely predicted positive wherein reality he does not suffer from heart disease. It is termed a Type-1 error.

False Negative (FN): The patient is falsely predicted negative wherein reality he suffers from heart disease. It is termed a Type-2 error.

Accuracy: The proportion of total correction predictions to the total predictions made (both correct and incorrect).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \qquad (4)$$

Precision: The proportion of true positive forecasts to the total number of positive forecasts.

Precision = $\frac{TP}{TP + FP}$ (5)

Recall value: The proportion of total true positive forecasts to the total number of true positive cases.

Recall = $\frac{TP}{TP + FN}$ (6)

F1-score = 2*$\frac{Precision*Recall}{Precision + Recall}$ (7)

6. Conclusion

The proposed ensemble is quite competitive. It has achieved similar results as some traditional algorithms and, in some cases, outperformed them. The ROC curves further validate the same.

Acknowledgement of Contravening Interests

The authors pronounce that there are no known contending monetary interests or individual connections that might have seemed to impact the work revealed in this paper.

References

1. Abdar, M., Ksiązek, W., U, R. A., Tan, R.-S., Makarenkov, V., & Pławiak, P. (2019). A new machine learning technique for an accurate diagnosis of coronary artery disease. Computer Methods and Programs in Biomedicine, 179. https://doi.org/10.1016/j.cmpb.2019.104992
2. Al-Makhadmeh, Z., & Tolba, A. (2019). Utilizing IoT wearable medical device for heart disease prediction using higher order Boltzmann model: A classification approach. Measurement, 147. https://doi.org/10.1016/j.measurement.2019.07.043
3. Beecy, A. N., Gummalla, M., Sholle, E., Xu, Z., Zhang, Y., Michalak, K., Dolan, K., Hussain, Y., Lee, B. C., Zhang, Y., Goyal, P., Campion Jr, T. R., Shaw, L. J., Baskaran, L., & Al'Aref, S. J. (2020). Utilizing electronic health data and machine learning for the prediction of 30-day unplanned readmission or all-cause mortality in heart failure. Cardiovascular Digital Health Journal, 1(2), 71-79. https://doi.org/10.1016/j.cvdhj.2020.07.004
4. Brunese, L., Martinelli, F., Mercaldo, F., & Santone, A. (2020). Deep learning for heart disease detection through cardiac sounds. Procedia Computer Science, 176, 2202-2211. https://doi.org/10.1016/j.procs.2020.09.257
5. Buenza, J.-J., Puertas, E., García-Ovejero, E., Villalba, G., Condes, E., Koleva, G., Hurtado, C., & F. Landecho, M. (2019). Comparison of machine learning algorithms for clinical event prediction (risk of coronary heart disease). Journal of Biomedical Informatics, 97. https://doi.org/10.1016/j.jbi.2019.103257
6. Das, R., Turkoglu, I., & Sengur, A. (2009). Effective diagnosis of heart disease through neural networks ensembles. Expert Systems with Applications, 36(4), 7675-7680. https://doi.org/10.1016/j.eswa.2008.09.013
7. Garate-Escamila, A. K., El Hassani, A. H., & Andres, E. (2020). Classification models for heart disease prediction using feature selection and PCA. Informatics in Medicine Unlocked, 19. https://doi.org/10.1016/j.imu.2020.100330
8. Latha C, B. C., & Jeeva S, C. (2019). Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. Informatics in Medicine Unlocked, 16. https://doi.org/10.1016/j.imu.2019.100203
9. Metsker, O., Sikorsky, S., Yakovlev, A., & Kovalchuk, S. (2018). Dynamic mortality prediction using machine learning techniques for acute cardiovascular cases. Procedia Computer Science, 136, 351-358. https://doi.org/10.1016/j.procs.2018.08.279
10. Mienye, I. D., Sun, Y., & Wang, Z. (2020). An improved ensemble learning approach for the prediction of heart disease risk. Informatics in Medicine Unlocked, 20. https://doi.org/10.1016/j.imu.2020.100402
11. R, S., & P, C. (2020). Predictive model of cardiac arrest in smokers using machine learning technique based on Heart Rate Variability parameter. Applied Computing and Informatics, 16(1/2). https://www.emerald.com/insight/2210-8327.htm
12. Shah, S. M. S., Shah, F. A., Hussain, S. A., & Batool, S. (2020). Support Vector Machines-based Heart Disease Diagnosis using Feature Subset, Wrapping Selection and Extraction Methods. Computers & Electrical Engineering, 84. https://doi.org/10.1016/j.compeleceng.2020.106628

13. Wang, Z., Zhu, Y., Li, D., Yin, Y., & Zhang, J. (2020). Feature rearrangement based deep learning system for predicting heart failure mortality. Computer Methods and Programs in Biomedicine, 191. https://doi.org/10.1016/j.cmpb.2020.105383.