**EFFECTIVE HEART DISEASE PREDICTION USING HYBRID MACHINE LEARNING TECHNIQUES**

*Report submitted to SASTRA Deemed to be University*

*as the requirement for the course*

**BCSCCS801: MINI PROJECT**

**Submitted by**

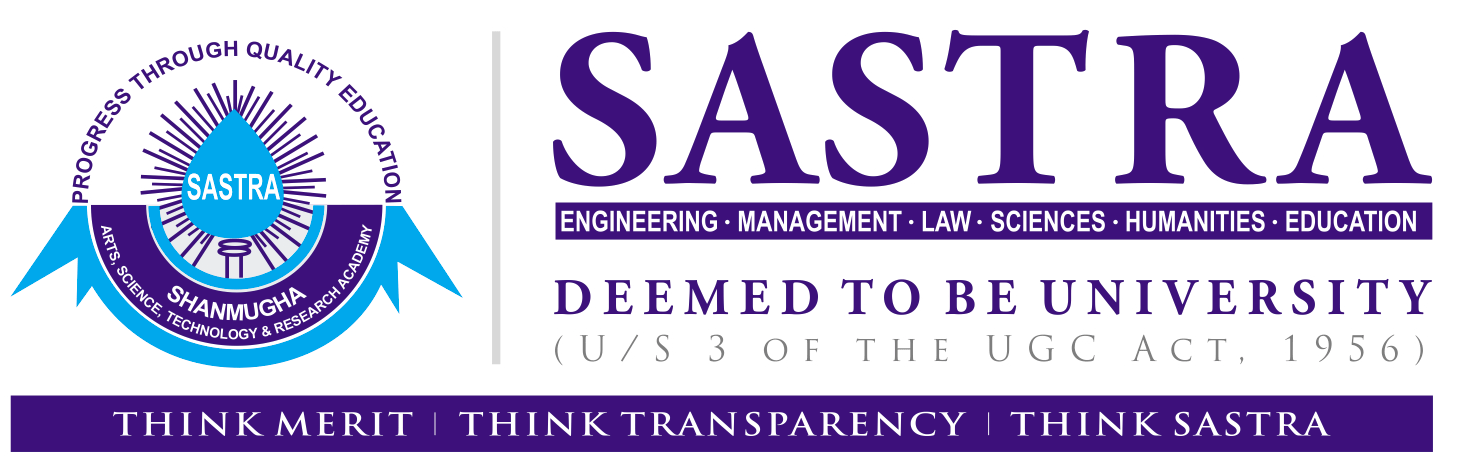
**Pabba Rohith**

**(Reg No. 123003172)**

**Gumpula Hruthik Kumar**

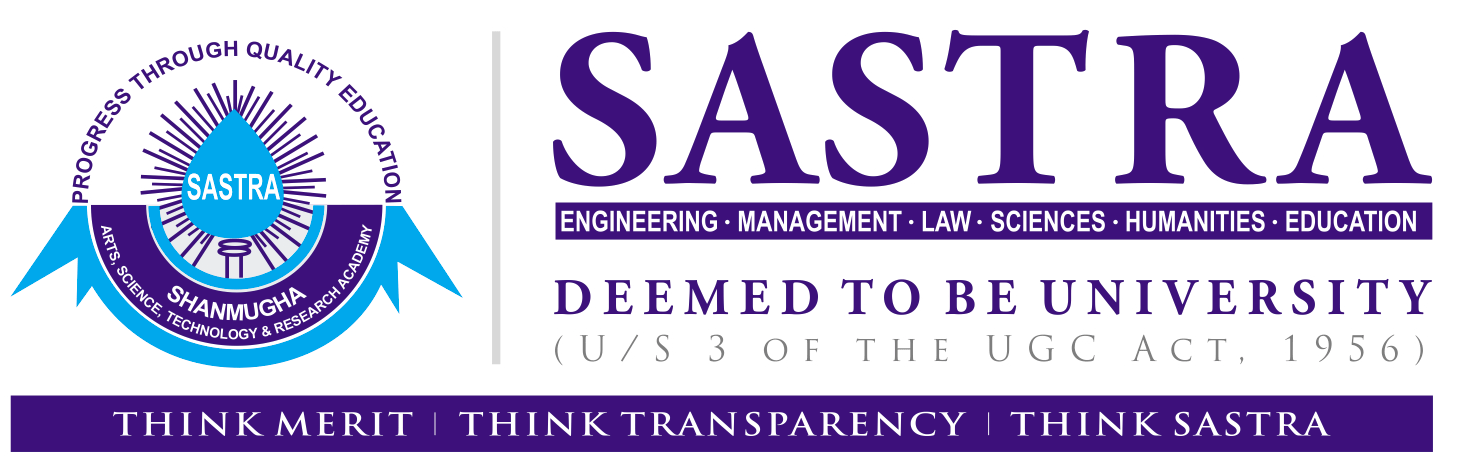
**(Reg No.123003072)**

**Marella Vishnu Vardhan Reddy**

**(Reg No.123003133)**

**SCHOOL OF COMPUTING**

**THANJAVUR, TAMIL NADU, INDIA-613401**

  
**SCHOOL OF COMPUTING**

**THANJAVUR, TAMIL NADU, INDIA-613401**

**Bonafide certificate**

This is to certify that the thesis titled “**Heart Disease Prediction using Hybrid Machine Learning Techniques”** submitted in partial fulfilment of the requirements for the award of the degree of B.Tech Computer Science and Engineering to the SASTRA Deemed to be University, is a bona-fide record of the work done by **Mr. PABBA ROHITH**(Reg.No. 123003172), **Mr. GUMPULA HRUTHIK KUMAR** (Reg.No.123003072), **Mr. MARELLA VISHNU VARDHAN REDDY**(Reg.No. 123003133) during the final semester of the academic year 2021-22, in the **School of Computing**, under my supervision. This thesis has not formed the basis for the award of any degree, diploma, associateship, fellowship or other similar title to any candidate of any University.

**Signature of Project Guide with affiliation:**

**Date:**

Mini Project Viva-voice held on:

**Examiner -I Examiner-II**



**DECLARATION**

We declare that the thesis titled “**Heart Disease Prediction using Hybrid Machine Learning Techniques”** submitted by us is an original work done by us under the guidance of **Mrs. Sudha N**, **Assistant Professor-III, School of Computing**, **SASTRA Deemed to be University** during the 6th semester of the academic year 2021-2022, in the **School of Computing**. The wok is original and wherever we have used materials from other sources. We have given due credit and cited them in the text of the thesis. This thesis has not formed the basis for the award of any degree, diploma, associateship, fellowship or other similar title to any candidate of any University.

**Signature of the Candidate(s):** 1.

## 3.

**Name of the Candidate(s) :** 1. PABBA ROHITH

2.GUMPULA HRUTHIK KUMAR

1. MARELLA VISHNU VARDHAN REDDY

**Date :**

**Acknowledgements**

We would like to acknowledge each and every one who had a role to play in making our humble efforts an out-to-out success.

We would like to thank **Dr. S. Vaidhyasubramaniam**, **our Honorable Vice - Chancellor** for providing us with an opportunity and the necessary infrastructure for carrying out this project as a part of our curriculum.

We find very heartening the encouragement and strategic support offered at every step of our college life by **Dr. A. Umamakeswari, Dean School of Computing, SASTRA Deemed University** and for that we would like to thank them time and again.

We extend our heartfelt thanks to **Dr. V. S. Shankar Sriram, Associate Dean (Computer Science and Engineering, School of Computing)** and **Dr. Muthaiah. R, Associate Dean (Information Technology, School of Computing)** for the motivation and support offered by them in materializing this project.

## Our Guide and Mentor Dr. Sudha N, Assistant Professor, School of Computing was the driving force behind this whole idea from the start.

Last but not the least, I am very much thankful to my parents for guiding me in every step which I took.

**List of Figures**

|  |  |  |
| --- | --- | --- |
| **Figure No.** | **Title** | **Page No.** |
| 1.1 | Methodology for implementation of the project | 1 |
| 4.1 | Heart data before pre-processing | 14 |
| 4.2 | Features that containing the null values | 14 |
| 4.3 | Heart data after the pre-processing | 14 |
| 4.4 | Confusion matrix and performance metrices of Logistic Regression and Naive Bayes | 15 |
| 4.5 | Confusion matrix and performance metrices of SVM and Random Forest | 15 |
| 4.6 | Confusion matrix and performance metrices of DT and GBT | 15 |
| 4.7 | Confusion matrix and performance metrices of VOTE and HRFLM | 15 |
| 4.9 | Graph showing the performance of the machine learning models with the HRFLM model | 16 |
| 4.10 | GUI showing the data to be entered to predict the heart disease using HRFLM | 17 |
| 4.11 | GUI showing the output for the entered data | 17 |

**List of Tables**

|  |  |  |
| --- | --- | --- |
| **Table No.** | **Table Name** | **Page No.** |
| 4.8 | Table showing the performance of machine learning models with the proposed model HRFLM | 16 |

**Abbreviations**

**LR -** Logistic Regression

**NB -** Naive Bayes

**RF –** Random Forest

**GBT –** Gradient Boosted Trees

**DT –** Decision Trees

**SVM –** Support Vector Machine

**HRFLM –** Hybrid Random Forest using Linear Model

**ML –** Machine Learning

**Abstract**

Heart disease has become common cause for death in the modern world. This disease is caused irrespective of age in these days and if not controlled or predicted at the earlier stage then this can cause a great damage and even fatality of the person. There are various kinds of cases that lead to these kinds of diseases. Some of the most commonly caused are high blood pressure where it is the situation in which the blood flows through the heart valves with a great pressure, heart valve disease where it is a situation in which due to some kind of abnormality any one valve out of the four valves couldn’t open or close in a right way which leads to blockage of blood or leakage of blood and many other kinds of diseases that leads to heart attack where the heart stops functioning suddenly and the person not even able to breathe which leads to death of the person.

There are many techniques which are used to predict the heart diseases such as Genetic Algorithm, Decision Trees, Naive Bayes and even Neural Networks which gave valuable results and they have also given results of Atrial fibrillation, Normal Sinus Rhythm, Premature Ventricular Contraction which helped us to predict the disease at the earlier stage. Since the disease is related to the heart it is more complex and more sensitive part, so the disease has to be handled very carefully and if the severity is more the treatment has to be done immediately. So, this lead to implement an algorithm that is even more efficient than the above algorithms with such as accuracy etc., parameters.

So, the main objective of the current proposed model uses all the features without any restrictions on using the feature selection. The technique we used here is the hybrid model named HRFLM (Hybrid Random Forest with Linear Model). We have shown that our proposed hybrid method has a stronger capability than the previous like the Decision Trees, Naive Bayes etc., in some of the parameters such as accuracy, specificity and in error control.

**Keywords**: Logistic Regression, Naive Bayes, Decision Tree, Hybrid Random Forest with Linear Model, Specificity.

**Table of Contents**

**Title Page No.**

Bonafide Certificate ii

Declaration Form iii

Acknowledgements iv

List of Figures v

List of Tables vi

Abbreviations vii

Abstract viii

1. Summary of the base paper 1
2. Merits and Demerits of the base paper 4
3. Source Code 5
4. Snapshots 14
5. Conclusion and Future Plans 18
6. References 19
7. **SUMMARY OF BASE PAPER**

**Base paper URL:** [**https://ieeexplore.ieee.org/document/8740989**](https://ieeexplore.ieee.org/document/8740989)

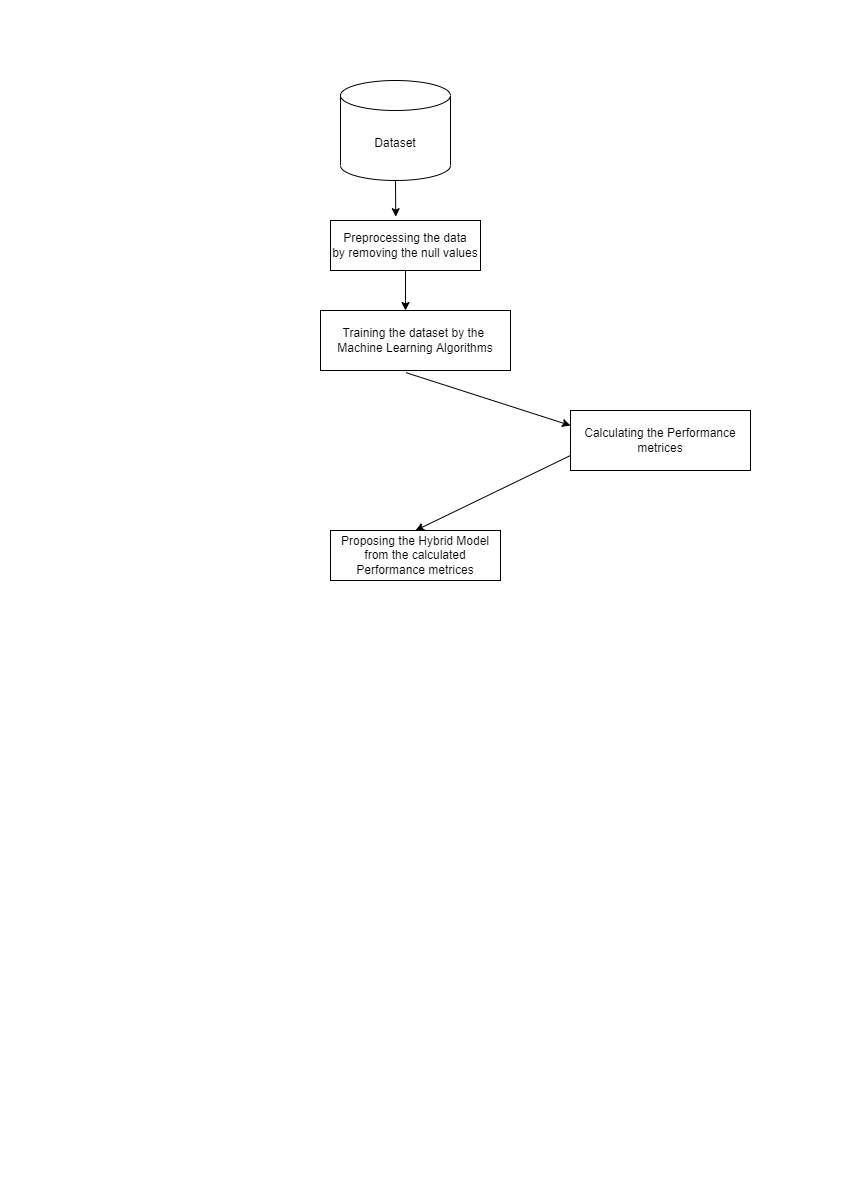
**Title of the Base Paper:** Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques

**Journal:** IEEE Access

**Page no:** 81542-81554

**Volume:** 7

**Year of publication:** 2019

**METHODOLOGY:**

**3.SOURCE CODE**

**Fig 1.1 Methodology for implementation of the project**

The proposed methodology for the Effective Heart Disease Prediction using Hybrid Machine Learning Techniques involves the above given steps. Initially it involves Data collection and preprocessing of the data. We preprocess the collected data by removing the null values or the empty values in the dataset. After preprocessing the data, we train the machine learning models by using the cleaned dataset.

**DATASET COLLECTION AND PREPROCESSING**

This dataset is taken from the UCI website. The dataset contained 303 different data out of which 6 data contained the null values. These data are removed from the dataset in the preprocessing stage. To obtain the precise solution, the dataset is divided into two sets namely the Training and Test set. The test dataset is 15% of the data of the original dataset.

**MACHINE LEARNING TECHNIQUES**

After preprocessing the data by applying the above steps, the model is now trained by the following machine learning algorithms with the help of the training dataset and then the trained model is tested using the test data. The ML techniques that are implemented in the current paper are as follows.

1. **Logistic Regression**

It comes under the category of Supervised learning. In Logistic Regression we use a function called sigmoid function/sigmoid curve. Logistic regression predicts the output of a categorical dependent variable using a set of independent variables. That is, it can be either Yes or No, 0 or 1, true or False. Instead of giving the exact value as 0 and 1, it gives the probabilistic values that lie between 0 and 1.

1. **Naive Bayes**

Naive bayes classifier is also supervised learning algorithm, it works based on Bayes theorem and used for solving classification problems. It predicts the output based on probability of an object; hence it is known as probabilistic classifier. If the occurrence of a certain feature is independent of the occurrence of other features, then it is called as Naive and it is called as naive bayes because it depends on the bayes theorem.

1. **Random Forest Classifier**

Random Forest is a popular Machine Learning algorithm, it comes under the supervised machine learning technique. The process of combining multiple classifiers to solve a complex problem and to improve the performance of the model is called ensemble learning. Random forest works on the concept of ensemble learning. To improve the predictive accuracy of a dataset, the random forest classifier takes the average of decision trees on various subsets of the given dataset. Based on the majority votes of predictions that are taken from each decision tree, it predicts the final output.

1. **Support Vector Machine**

Support Vector Machine is a popular Machine Learning algorithm, it comes under the supervised machine learning technique. In SVM, we use decision boundary that separates the target variable into two different classes with best plane that plane is called Hyperplane. In SVM we use Linear SVM.

1. **Decision Tree**

Decision tree is also a supervised learning technique where we use the concept of model trees and we build the trees based on combining two or more features which contribute to the disease prediction and using this we build n number of model trees and predict the result for the given test data from the constructed trees. The depth of the tree can also be mentioned so that we can improve the accuracy.

1. **Gradient Boosted Trees**

In the field of machine learning we all come across of errors in predicting the output, the errors may include Bias error or Variance error, so gradient boosting is such a powerful machine learning algorithm that it is used to reduce the bias error of the model. Gradient boosting works by constructing simple (weak) predictions models, where each model tries the error left over by the previous model.

1. **Voting Classifier**

A voting classifier is the one of machine learning model that trains on ensemble of various combination of models and predicts the output based on the highest probability. In this it simply collects all the outputs of different classifier models and finalize the output based on highest majority of voting. In this project we have implemented hard voting, in hard voting we will predict the output class based on the highest majority of votes i.e, the class which has highest probability of being predicted.

1. **Stacking Classifier**

Stacking classifier is also a machine learning model that train on ensemble classification. It is classified as two layered estimators; the first layer will consist of all the baseline classification models which are used to predict the output for the given test data set and second layer is considered as Meta-classifier which takes the outputs of prediction of all the baseline classification as input. Here we give these predicted results to another classifier model so that it generates new prediction.

1. **Merits and Demerits of the Base Paper**

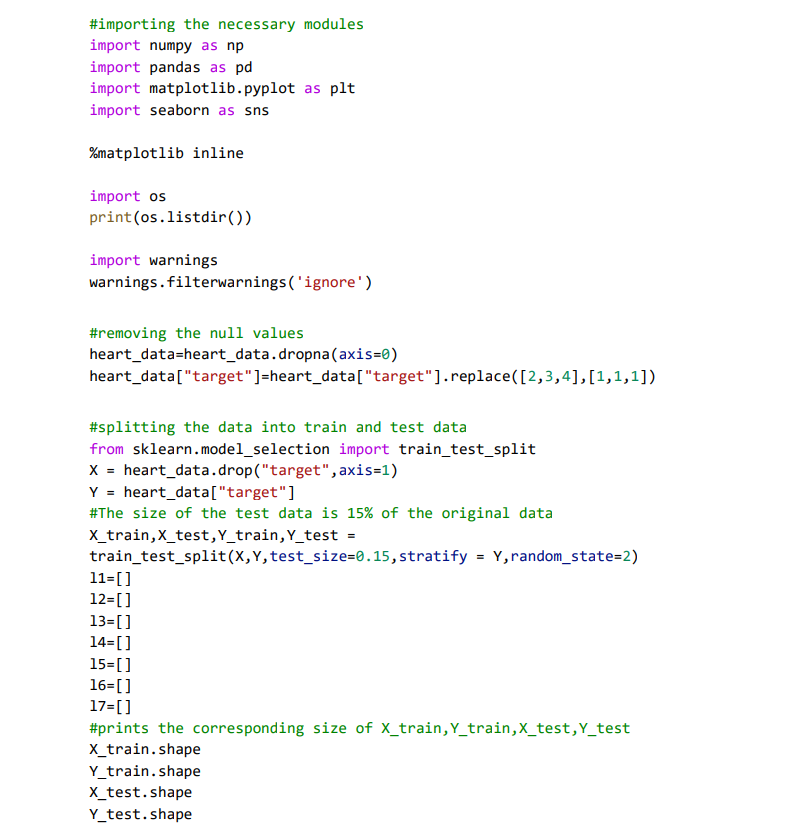
|  |  |  |
| --- | --- | --- |
| **S. No** | **Research Paper** | **Methodology proposed** |
| **1.** | Using PSO Algorithm for producing the best rules in diagnosis of the heart disease | In this base paper they used Particular Swarm Optimisation method in predicting the heart disease.  But the algorithm has low convergence. |
| **2.** | Radial Basis function neural network for prediction of cardiac arrhythmias based on heart rate time series | In this base paper they used Radial function in predicting the disease but the classification in this function is low. |
| **3.** | Hybrid Intelligent modelling schemes for heart disease classification | In this base paper they used Logistic Regression and ANN but the accuracy is low in ANN. |
| **4.** | A computational intelligence method for effective diagnosis of heart disease using genetic algorithm | In this base paper they used Genetic algorithm in the prediction of the heart disease but the efficiency of the algorithm is less. |
| **5.** | Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks | In this base paper they have used the concept of fuzzy neural networks but they work on inaccurate inputs and the accuracy is too low |

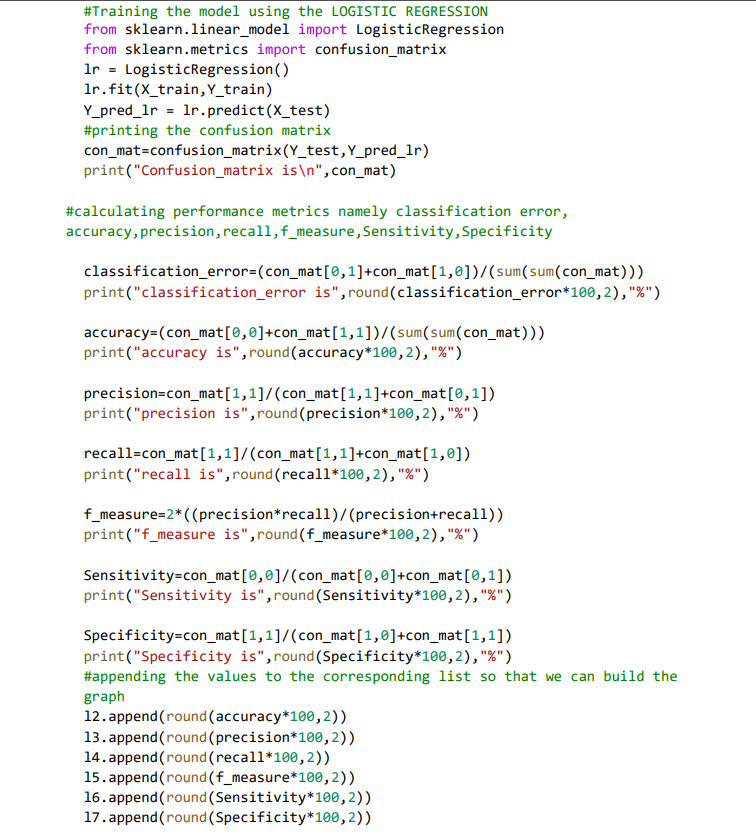
The merits that we have achieved from the base paper are we have improved the accuracy of many algorithms namely Naive Bayes, Logistic Regression, Gradient Boosted Trees, Voting classifier and the proposed hybrid model.

We are able to achieve these merits because of the better data pre-processing techniques.

1. **SOURCE CODE**

**PREPROCESSING THE DATA**



**LOGISTIC REGRESSION**

**NAIVE BAYES**



**RANDOM FOREST**



**GRADIENT BOOSTED TREES**



**DECISION TREE**



**SUPPORT VECTOR MACHINE**



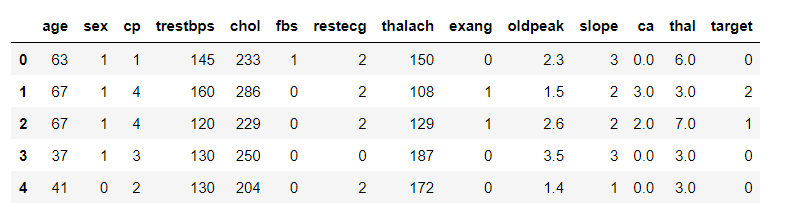
**VOTING CLASSIFIER**



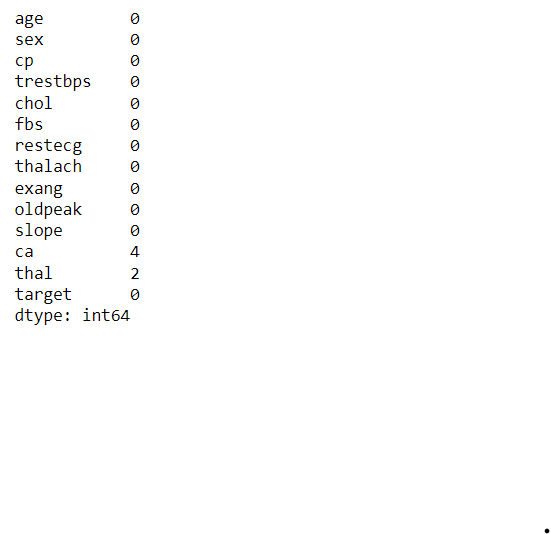
**HRFLM (HYBRID RANDOM FOREST USING LINEAR MODEL)**



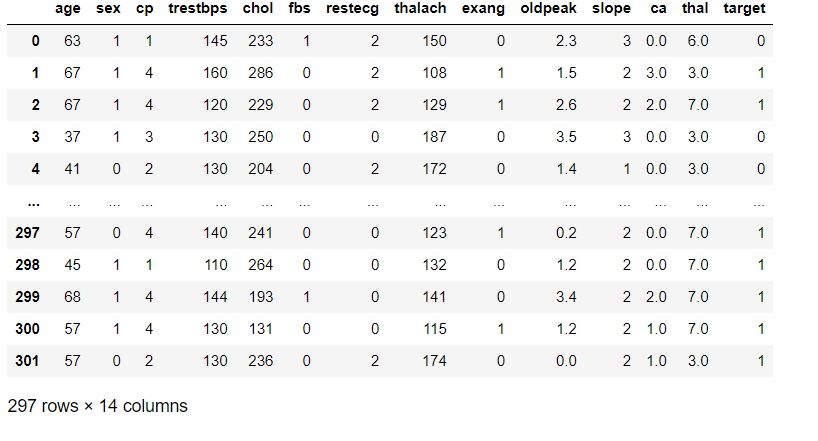
1. **SNAPSHOTS**

****

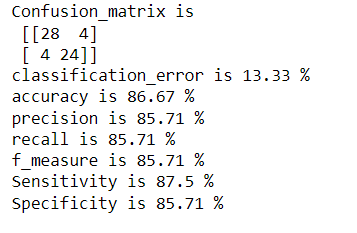
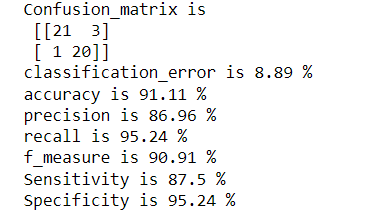
**Fig 4.1 Heart data before pre-processing**



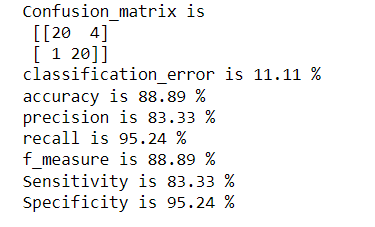
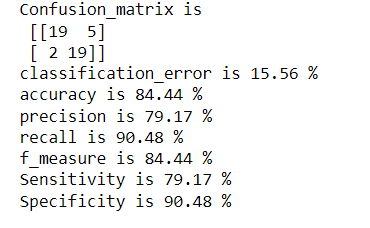
**Fig 4.2 Features that containing the null values**

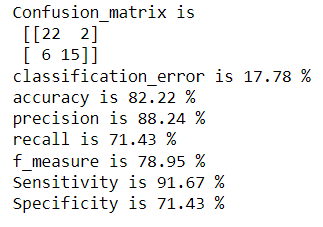
****

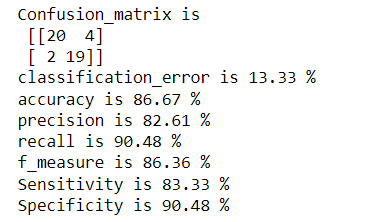
**Fig 4.3 Heart data after the pre-processing**

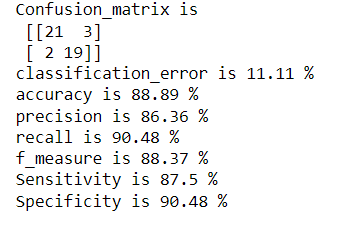


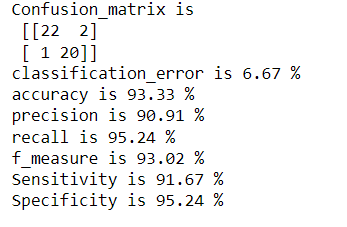
**Fig 4.4 Confusion matrix and performance metrices of Logistic Regression and Naive Bayes**

****

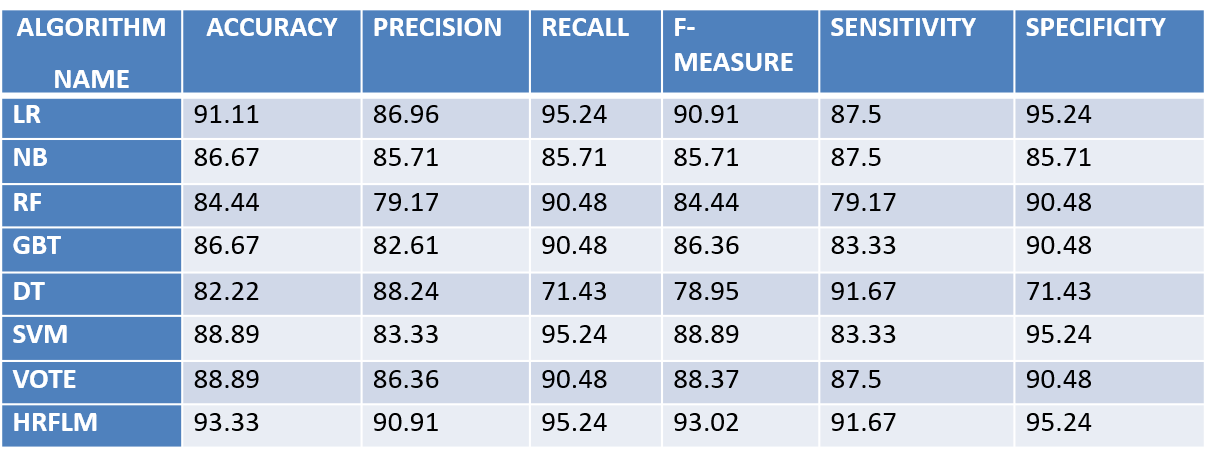
**Fig 4.5 Confusion matrix and performance metrices of SVM and Random Forest**

****

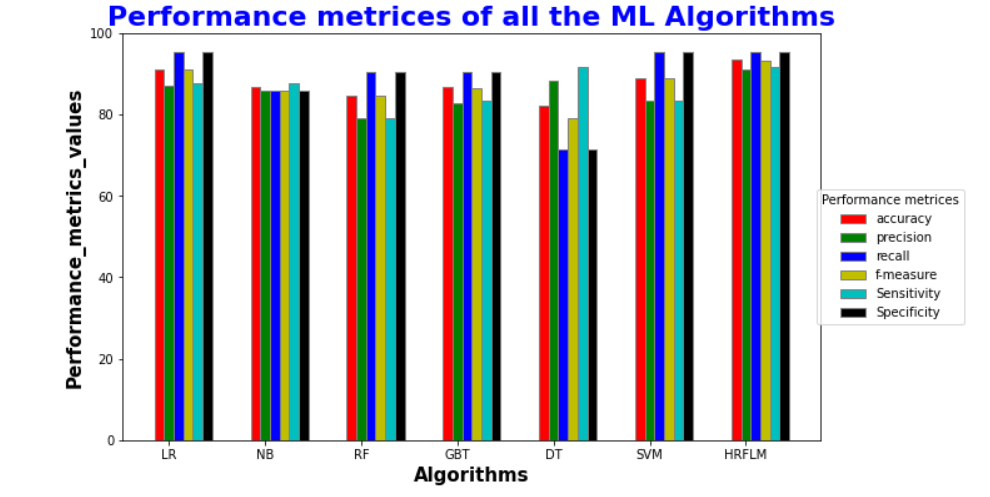
**Fig 4.6 Confusion matrix and performance metrices of DT and GBT**

****

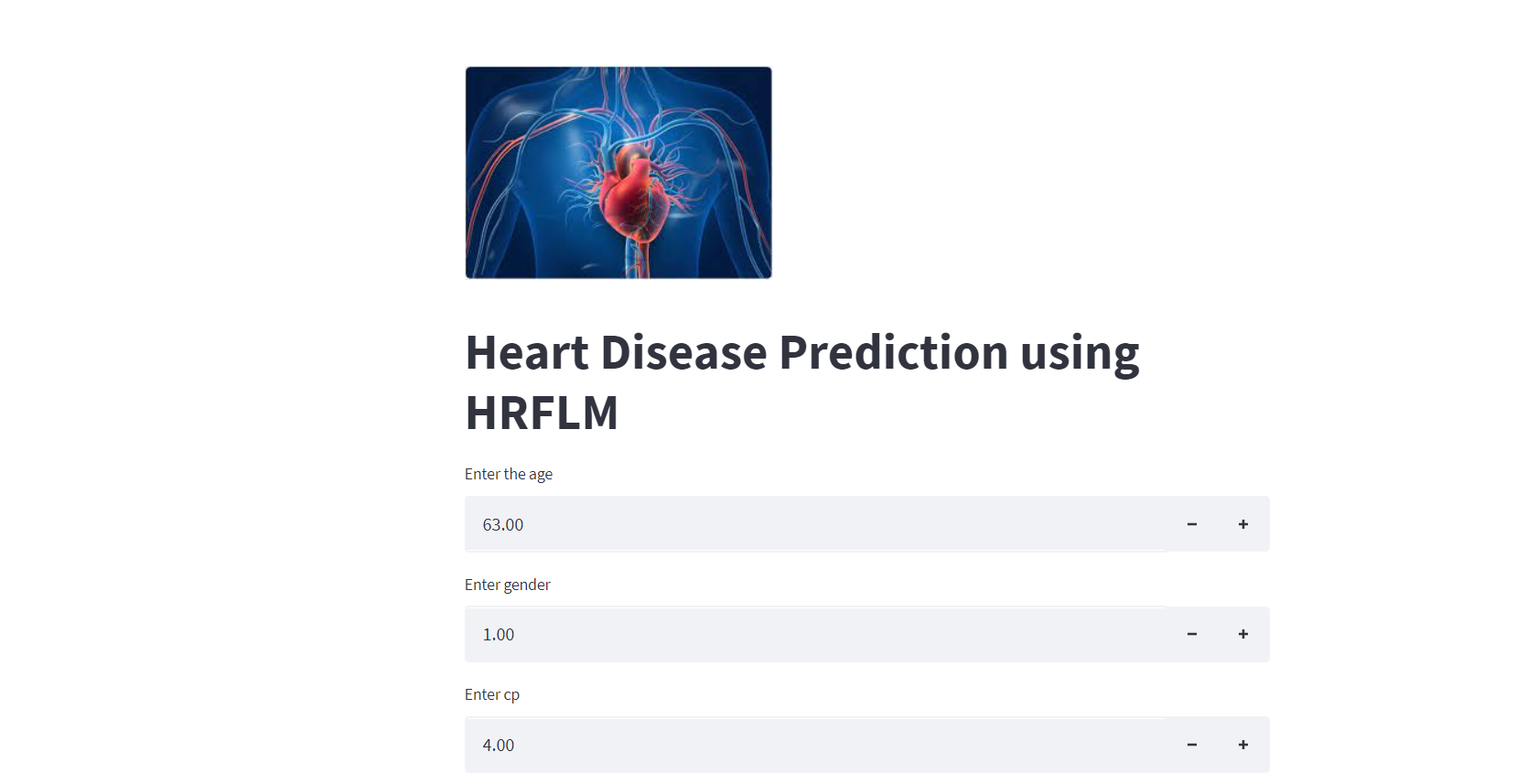
**Fig 4.7 Confusion matrix and performance metrices of VOTE and HRFLM**

****

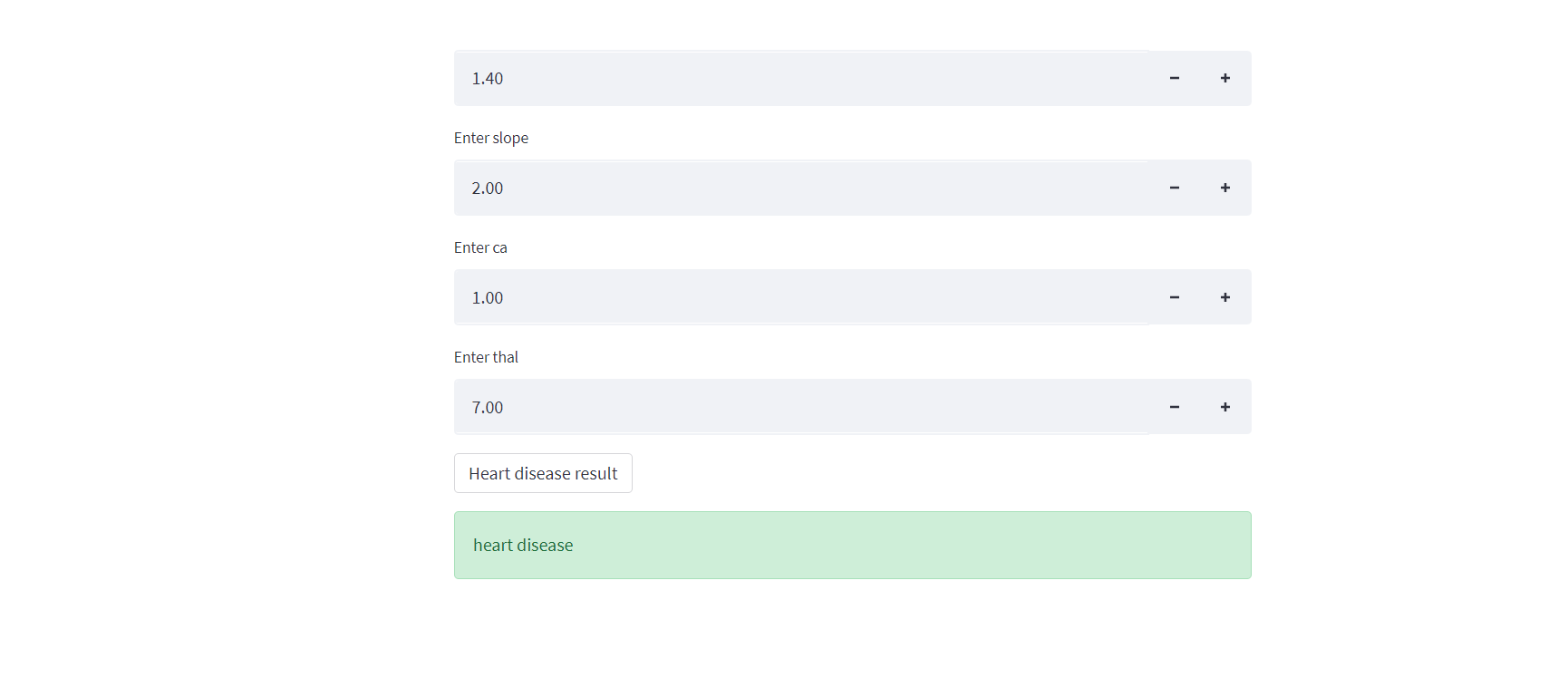
**Fig 4.8 Table showing the performance of machine learning models with the proposed model HRFLM**

****

**Fig 4.9 Graph showing the performance of the machine learning models with the proposed HRFLM**

****

**Fig 4.10 GUI showing the data to be entered to predict the heart disease using HRFLM**

****

**Fig 4.11 GUI showing the output for the entered data**

1. **CONCLUSION AND FUTURE PLANS**

The proposed framework illustrates about the Heart disease prediction using the Hybrid Machine Learning Techniques. In this framework, initially we pre-processed the data and then we used the ML classifier algorithms and trained the model with this pre-processed dataset. The ML classifier algorithms that we used in this framework are Logistic Regression, Naive Bayes, Random Forest, Gradient boosted trees, Decision Tree, Support Vector Machine and Voting Classifier.

From the above ML classifier algorithms, we calculated the performance metrices of each model and from the results, we proposed a Hybrid model using the Logistic Regression, Random Forest and Decision Tree. After proposing the Hybrid model the calculated experimental results show that the accuracy has improved a lot when compared to that of the normal ML classifier algorithms. The proposed HRFLM framework has achieved an accuracy of 93.33%. We have also used other performance metrics in this framework namely precision, recall, F-measure, Sensitivity and Specificity.

Later, this proposed Hybrid model can be tested on large datasets and we can also combine various other algorithms in this hybrid model like the ConvolutionalNeuralNetworks (CNN) in order to implement an algorithm and to get the model with the less classification error rate and also with increased specificity and sensitivity.

1. **REFERENCES**

* A. S. Abdullah and R. R. Rajalaxmi, ‘‘A data mining model for predicting the coronary heart disease using random forest classifier,’’ in Proc. Int. Conf. Recent Trends Comput. Methods, Commun. Controls, Apr. 2012, pp. 22–25.
* C. A. Devi, S. P. Rajamhoana, K. Umamaheswari, R. Kiruba, K. Karunya, and R. Deepika, ‘‘Analysis of neural networks-based heart disease prediction system,’’ in Proc. 11th Int. Conf. Hum. Syst. Interact. (HSI), Gdansk, Poland, Jul. 2018, pp. 233–239.
* H. A. Esfahani and M. Ghazanfari, ‘‘Cardiovascular disease detection using a new ensemble classifier,’’ in Proc. IEEE 4th Int. Conf. Knowl.- Based Eng. Innov. (KBEI), Dec. 2017, pp. 1011–1014.
* F. Dammak, L. Baccour, and A. M. Alimi, ‘‘The impact of criterion weights techniques in TOPSIS method of multi-criteria decision making in crisp and intuitionistic fuzzy domains,’’ in Proc. IEEE Int. Conf. Fuzzy Syst. (FUZZ-IEEE), vol. 9, Aug. 2015, pp. 1–8.
* A. Gavhane, G. Kokkula, I. Pandya, and K. Devadkar, ‘‘Prediction of heart disease using machine learning,’’ in Proc. 2nd Int. Conf. Electron., Commun. Aerosp. Technol. (ICECA), Mar. 2018, pp. 1275–1278.