Indian Institute of Information Technology, Allahabad

Department of Information Technology

M. Tech. Project report

Prediction of Heart Disease using Ensemble Techniques and Multi Layer Perceptron Neural Network

---

Under the Guidance of

Prof. O. P. Vyas

By

Hemant Kumar Lader (*MIT2019007*)

Table of Contents

Prediction of Heart Disease using Ensemble Techniques and Multi Layer Perceptron Neural Network

*Hemant Kumar Lader**
Indian Institute of Information Technology,
Prayagraj, UP, India

Abstract

In medical science, Heart disease is a deadly disease that a large population of people around the world suffers from. Every second most people die due to this problem. For that reason, early diagnosis of such an illness is very important. The traditional way of diagnosing such diseases is not sufficient. Developing machine learning-based medical diagnosis systems for prediction of heart disease gives accurate results than in traditional ways of diagnosing. The predictive modeling solution for the prediction of heart disease is extremely challenging. In this project, I proposed a predictive model using ensemble machine learning technique using different machine learning algorithms as its base model and multi layer perceptron as its final classifier model and compared the individual performances of these base models to the proposed approach. In the end I compared some other ensemble methods to check if they give better results. This project can effectively predict heart disease, so the proposed approach can be further used for developing IOT application and online application for heart disease prediction systems.

Keywords: Heart disease prediction, ensemble learning, machine learning, stacking, multi layer perceptron.

1 Introduction

A large number of people die every year because of heart disease and a large population of people suffers from heart disease. Prediction of heart disease early plays a crucial role in the treatment. If heart disease could be predicted before, lots of patient deaths would be prevented, and also a more accurate and efficient treatment could be provided. According to the World Health Organization, around 17.9 million people die each year. It is expected that the count of deaths due to heart disease may reach up to 23.3 million by 2030 [1]. There are a number of factors that increase the risk of Heart disease.

Some of them are:

- Family history of heart disease
- Smoking
- Cholesterol
- high blood pressure
- Obesity
- Lack of physical exercise

The traditional ways of diagnosing heart disease are not effective for present-day situations. As patients with potential heart disease symptoms are increasing. The diagnosis of the disease is a difficult and tedious task in the medical field. Heart disease prediction depends on various features and symptoms. This multi layered problem can produce false presumptions and unpredictable effects.

However, machine learning techniques are useful to predict the output from existing data. Hence, in this project, I applied, machine learning techniques called classification in order to enhance the performance of predicting the risk of heart disease from the risk factors. Using these techniques I also tried to improve the accuracy of prediction and to reduce the diagnosis time for predicting heart disease risk using the stacking ensemble method by taking Multi Layer Perceptron(MLP) model as the final classifier model.

2 Problem Definition

To develop a system for predicting heart disease from risk factors using machine learning techniques Ensemble Methods and Multi Layer Perceptron model. To predict heart disease efficiently and accurately.

3 Literature Review

There are a lot of studies on the prediction of heart disease as a medical diagnosis system. Machine learning is useful for a diverse set of problems. One of the applications of this technique is in predicting a dependent variable from the values of independent variables. The healthcare field has huge resources which makes this field difficult to handle, data mining can be applied here to manage the resources and handle [1].

Heart disease has been recognized as major root of death even in developed countries [2]. One of the reasons for fatality due to heart disease is due to the fact that the risks are either not identified, or they are identified only at a later stage. However, machine learning methods can be useful for solving this problem and to detect the risk at an early stage. Some of the techniques used for such prediction problems are the Support Vector Machines (SVM), Neural Networks, Decision Trees, Regression and Naïve Bayes classifiers.

Ankita Dewan and Meghna Sharma have discussed various kinds of techniques for developing a heart disease prediction system and proposed using Backpropagation Algorithm as the best classification technique for the targeted system [3]. Subha et al. (2018) [4] have discussed an ensemble based Extreme Learning Machine (ELM) for Cardiovascular disease prediction. However, their experimental results had shown that proposed model outperformed well and produces better classification accuracy than base classification models.

Kamley and Thakur (2019) [5] have presented the comparative study of three important machine learning classification techniques like Support Vector Machine (SVM), Naive Bays (NB)

and K-Nearest Neighbor (KNN) for heart disease prediction. The Kaggle data source is obtained and Matlab R2017a machine learning tool is used for study purposes. Finally, their experimental results had stated that SVM had recorded the highest classification accuracy of 86.12% than other methods.

In [9], the age range was grouped by the K-Nearest Neighbor algorithm into class. The Risk level of each class was identified with the help of ID3 algorithm. The accuracy of prediction was measured by considering different attributes. This gave highest of 80.6% accuracy. In [10], authors investigated the heart disease prediction using KStar, J48, SMO, Bayes Net and Multilayer Perceptron through Weka software. Using k-Fold Cross-Validation resampling procedure. By combining the performance accuracy and ROC curve the overall performance of these techniques is measured.

4 DATASET DESCRIPTION

The Cleveland heart dataset from the UCI machine learning repository has been used for the experiments. The dataset consists of 14 attributes and 303 instances. There are 8 categorical attributes and 6 numeric attributes [13]. Descriptions of each feature are given in Table 1.

Table 1. Dataset attributes and its description

S.No.	Attributes	Brief Description
1	Age	Age in Years
2	Gender	Sex Male and Female
3	Cp	chest pain type ○ Value 1: typical angina ○ Value 2: atypical angina ○ Value 3: non-anginal pain ○ Value 4: asymptomatic
4	Trestbps	resting blood pressure (in mm Hg on admission to the hospital)

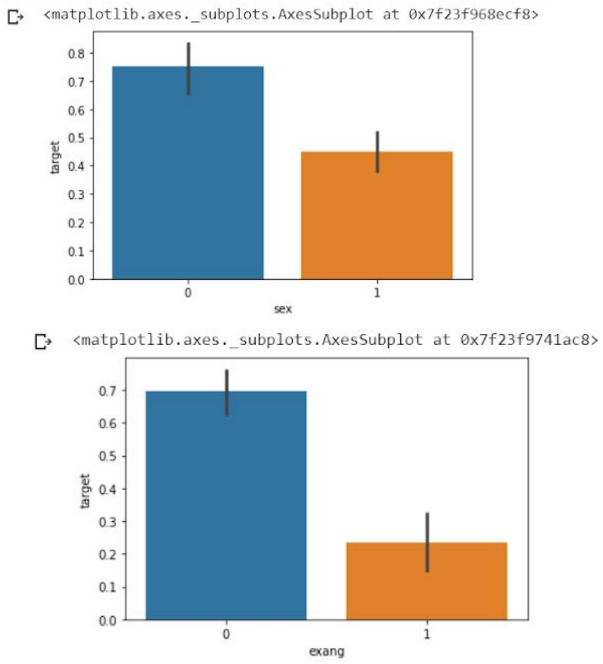
5	Chol	serum cholesterol in mg/dl
6	Fbs	fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
7	restecg	resting electrocardiographic results ○ Value 0: normal ○ Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 005 mV) ○ Value 2: showing probable or definite left ventricular hypertrophy by Estes criteria
8	Thalach	maximum heart rate achieved in beats per minute (bpm)
9	exang	exercise-induced angina (1 = yes; 0 = no)
10	Oldpeak	ST depression induced by exercise relative to rest
11	Slope	the slope of the peak exercise ST segment ○ Value 1: upsloping ○ Value 2: flat ○ Value 3: down-sloping
12	Ca	number of major vessels (0-3) colored by fluoroscopy
13	Thal	Thalassemia; 3 = normal; 6 = fixed defect; 7 = reversible defect
14	Target ()	Disease present or not : 0 = Absent, 1 = Present

The target feature has two classes and hence it is a binary classification problem. To reiterate, the goal is to predict whether a person has heart disease. I split the whole dataset into two parts. One is training dataset which is 70% of the complete dataset and the other 30% as testing dataset. Initially I checked for missing values and found no missing values, also there was no outliers in the dataset.

5 METHODOLOGY

This project is developed with the goal of predicting heart disease and to produce accurate classification accuracy by applying machine learning classifiers. In this project, 7 different machine learning methods are used to compare the results and ensemble them to make a hybrid classifier using the ensemble's Stacking method. Other methods of ensemble methods like Adaboost and Voting classifiers are also used to check if they give better results than the proposed approach.

First I analyzed which attributes of the dataset are strongly correlated. And how it affects the final prediction. By analyzing the dataset it is found that females are more likely to have heart problems than males. Likewise, exang attribute is strongly correlated with the target attribute. And this is how other attributes are correlated. Likewise, exang attribute is strongly correlated with the target attribute. And this is how other attributes are correlated. Figure 1.




```
print(dataset.corr()["target"]

target      1.000000
exang       0.436757
cp          0.433798
oldpeak     0.430696
thalach     0.421741
ca          0.391724
slope       0.345877
thal        0.344029
sex         0.280937
age         0.225439
trestbps    0.144931
restecg     0.137230
chol        0.085239
fbs         0.028046
Name: target, dtype: float64
```

Fig. 1 Correlation of different attributes with target

5.1 STACKING ENSEMBLE METHOD

In [8] Mustafa Jan et.al, used an ensemble approach to develop user interfaced heart disease prediction but they used different base learners in addition to those learners this I used two more base learners. The ensemble approach combines the different classifier models and gives a super classifier model in order to get more accurate prediction performance [14]. This method is well known to data science and machine learning for its capability of merging the power of different techniques to produce a hybrid model. An increase in the performance of this enhanced model is its built-in characteristics. In this project, I proposed a stacking based ensemble classifier that uses Multi layer Perceptron as its final model. This method takes intermediate predicted results as input and then final classifier predicts the final predicted result, hence this approach outperforms better than used individual base learning model. Fig. 2 shows the standard working of ensemble classification approach.

There are other ensemble approaches used in [8] and [5], this project used the stacking ensemble method. Riyaz et al., in [15] use the stacking ensemble method. The structure of the stacking ensemble method is shown in Fig. It uses n different classification model. These classification models are base learners, also known as weak classifiers. The split Training dataset is passed through all the base learners. T final classifier also termed as meta-classifier takes n predictions from these weak classifiers and gives the final predicted result.

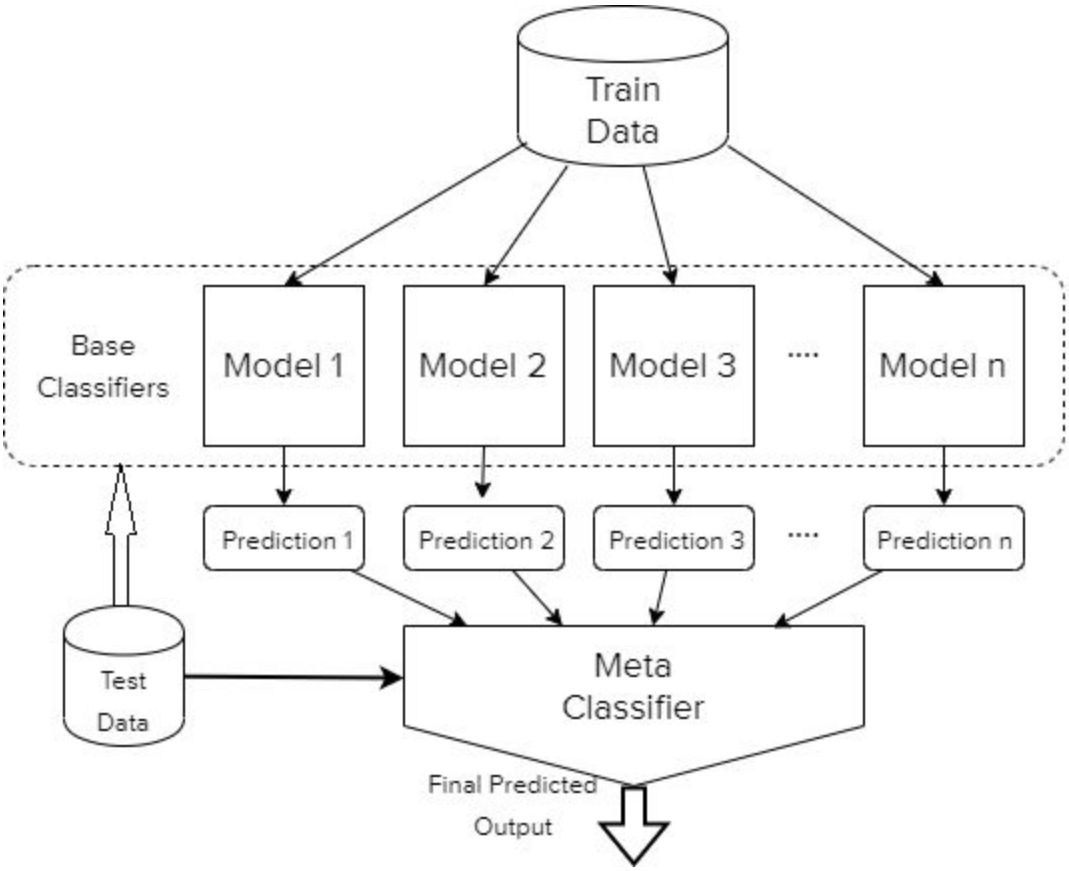


Fig 2. Structure of Stacking ensemble method

5.2 CLASSIFIERS FOR ENSEMBLE METHOD

As in [5] the author used the ensemble of different algorithms. In this project, I used 7 different classifiers to ensemble them using the Stacking ensemble technique. These are Logistic Regression(LR), Naive Bayesian(NB), Support Vector Machine (SVM), K Nearest Neighbour (KNN), Decision Tree(DT), Multilayer Perceptron(MLP) and Random forest (RF). I compared the individual performance of these classifiers Figure 2.

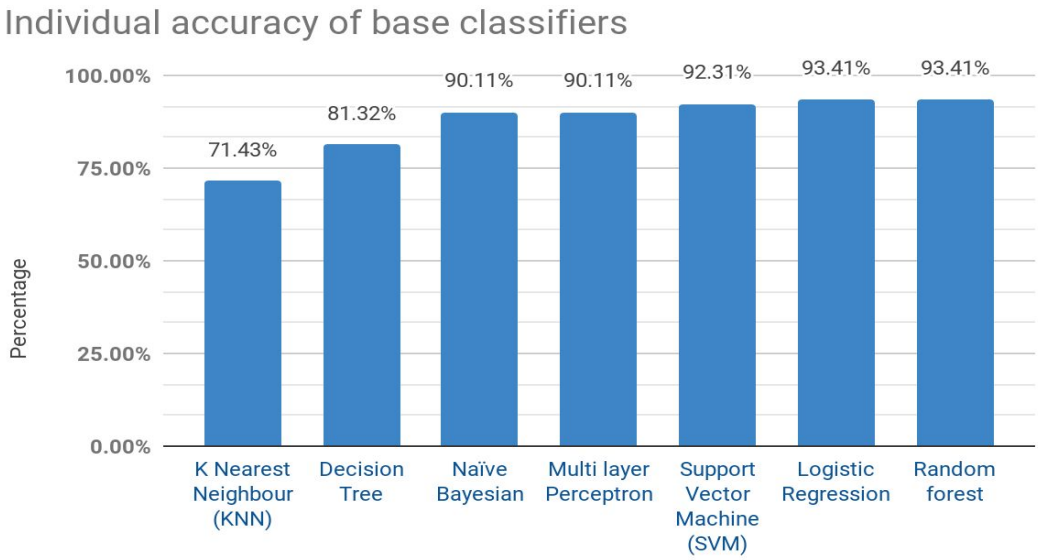


Fig 3. Comparison of different base classifiers

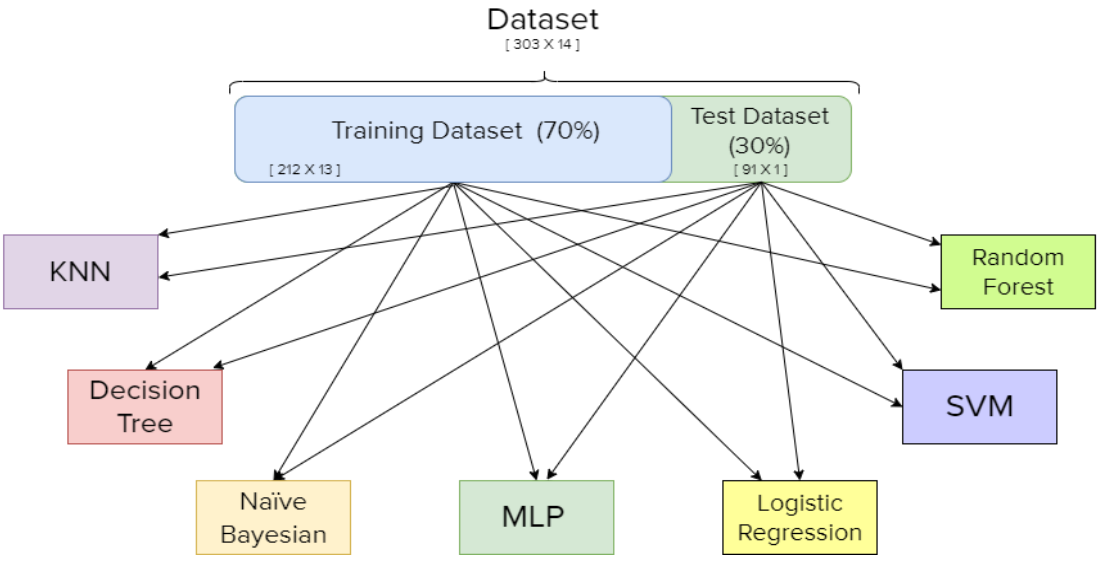


Fig 4. Weak classifiers used in proposed staking ensemble method

5.3 PROPOSED APPROACH

This project work proposed the stacking ensemble model using the seven mentioned classification models. And Multilayer Perceptron as the final model for stacking. I selected

MLP neural network as the meta classifier because in the reviewed paper [2] and [12] uses a neural network and it performed well in predicting heart disease.

All the 13 features of the heart disease dataset are taken as input to the system and each weak learner is trained on these 13 parameters, then the predicted output of these seven models is taken by meta classifier. These predicted outputs are stacked by generating an intermediate dataset. Then base on these intermediate dataset the meta classifier predicts the final prediction result.

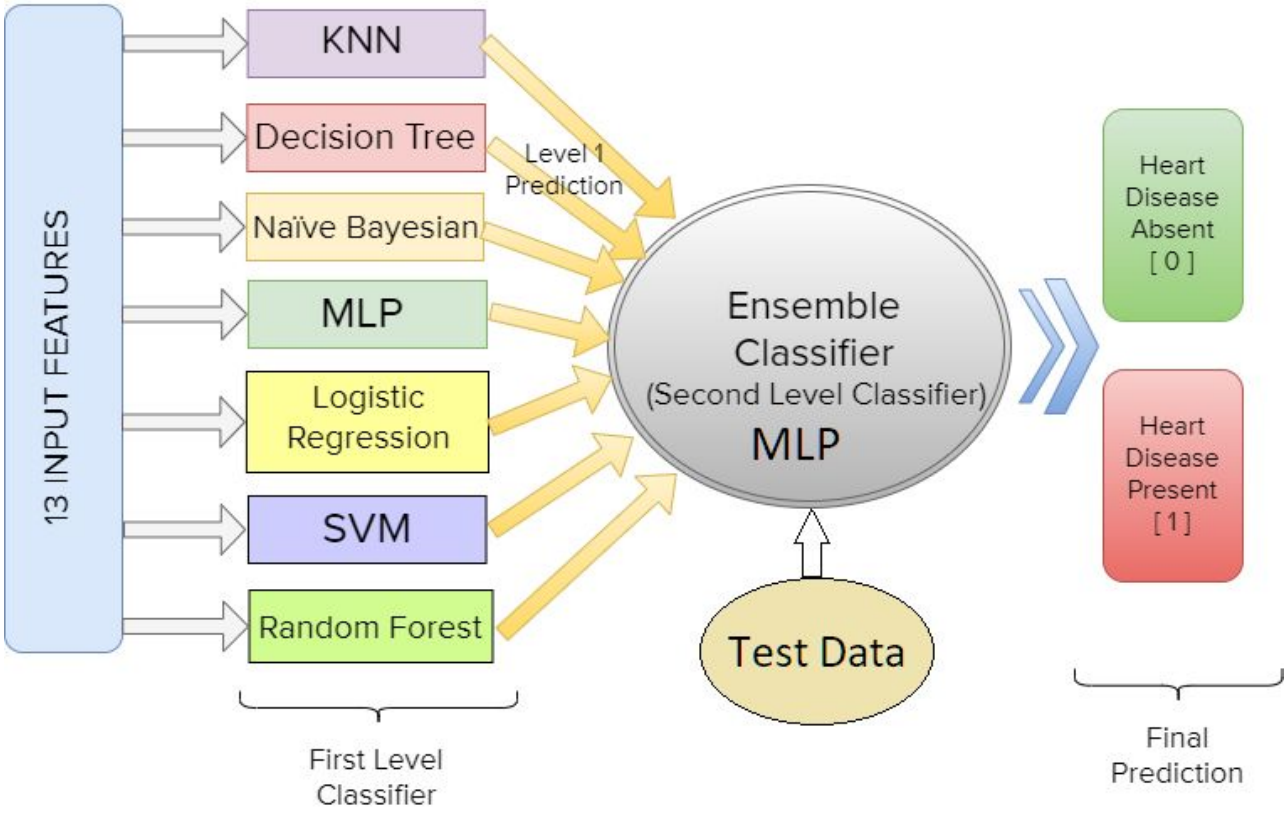


Fig 5. The architecture of proposed system

5.4 MLP AS META CLASSIFIER

One of the most important models in the Artificial Neural Network is Multilayer Perceptron (MLP). The type of architecture used to implement the system is Multilayer Perceptron Neural Network (MLPNN).

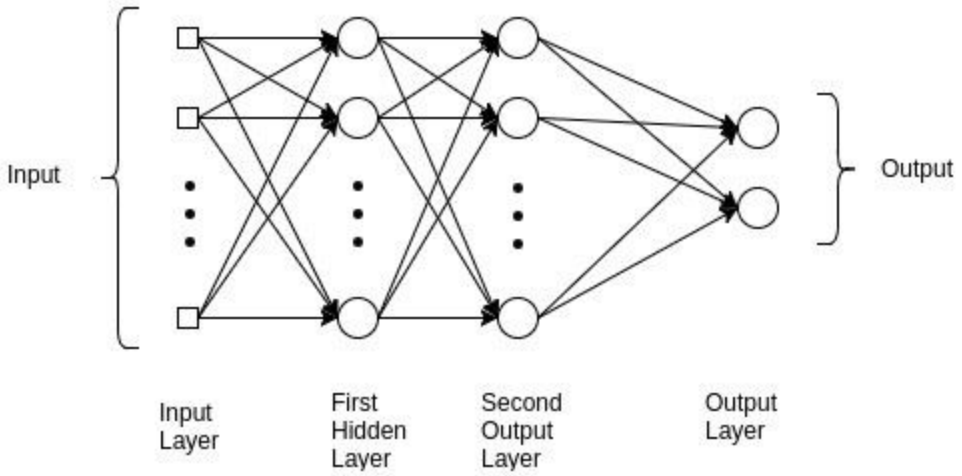


Fig 6: Multilayer Perceptron Neural Network

In this project, two MLP classifiers are used one as a weak learner and one as a meta classifier. The weak classifier has two hidden layers first contains seven neurons and the other hidden layer contains five neurons, similarly, the meta-classifier contains just one hidden layer containing five neurons Fig 7.

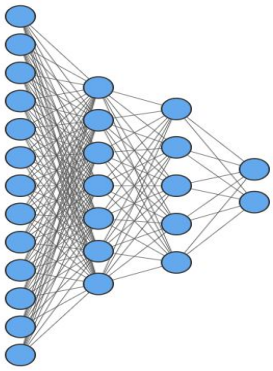


Fig 7. a. MLP of weak classifier

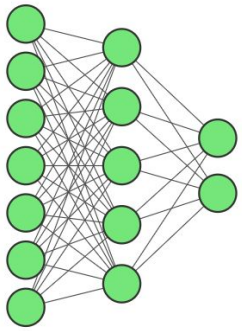


Fig 7.b. MLP of meta classifier

6 RESULT

Using 13 features of the dataset as input the proposed system gives 94.51% accuracy, which is a little higher than the individual performance of the base learner models. In addition to this

proposed approach, I tested for the other ensemble methods like AdaBoost using different classification methods and voting classifier, the performance comparison is shown in Fig 8. Clearly the proposed approach performance is higher than those other methods.

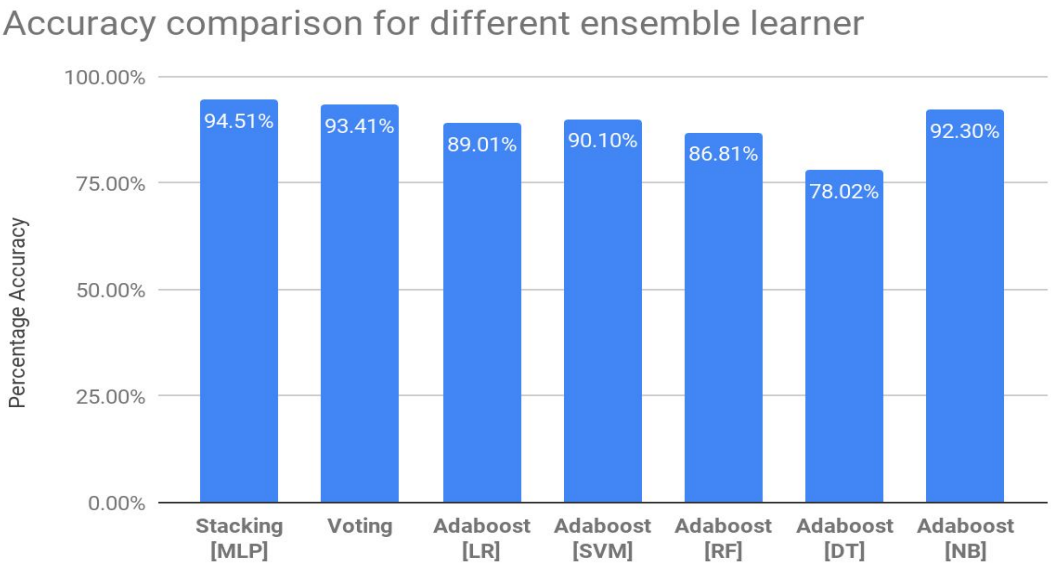


Fig 8. Comparison of other ensemble methods to the proposed method

Heart Disease	Predicted Absent	Predicted Present
Actual Absent	40 TP	3 FN
Actual Present	2 FP	46 TN

Fig. 9 Confusion matrix

Fig. 9 shows the confusion matrix of the predictive model.
True Positive (TP) : Disease actually present, predicted present.
False Negative (FN) : Disease actually present, predicted absent.
True Negative (TN) : Disease actually absent, predicted absent.

False Positive (FP) : Disease actually absent, predicted present.

Accuracy is computed by dividing the number of predictions that are correct by the number of all predictions. The obtained result is multiplied by 100 to get value as a percentage.

$$Accuracy=(TP+TN)/(TP+TN+FP+FN) = 0.9451$$

Precision denotes the ratio of the instances that are predicted as having heart disease actually have heart disease.

$$Precision=TP/(TP+FP) = 0.95$$

Recall denotes the proportion of the instances that actually have heart disease are predicted as having heart disease.

$$Recall=TP/(TP+FN) = 0.93$$

7 CONCLUSION

From the result, it is clear that using ensemble methods gives better performance than that of their individual learning model. Ensemble methods are meta-algorithms that combine several machine learning techniques into one predictive model in order to decrease variance, bias or improve predictions (stacking).

The trained model of this system can be used to combined with web portal or mobile application, It can also be integrated with IoT applications. To alert the doctors prior to the high risk.

Accuracy for this prediction may increase beyond 94.51% by using extreme gradient boosting. It converts the weak model to strong models by adjusting the weights iteratively to reduce bias and increase accuracy.

8 References

1. Karayilan, Tulay, and Ozkan Kilic. “Prediction of Heart Disease Using Neural Network.” *2017 International Conference on Computer Science and Engineering (UBMK)*, 2017, doi:10.1109/ubmk.2017.8093512.
2. K, Vanisree, and Jyothi Singaraju. “Decision Support System for Congenital Heart Disease Diagnosis Based on Signs and Symptoms Using Neural Networks.” *International Journal of Computer Applications*, vol. 19, no. 6, 2011, pp. 6–12., doi:10.5120/2368-3115.
3. M. Sharma, "Prediction of heart disease using a hybrid technique in data mining classification", 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom), pp. 704-706, 2015.
4. R. Subha, K. Anandakumar and A. Bharathi (2018). An Ensemble based Extreme Learning Machine for Cardiovascular Disease Prediction. International Journal of Applied Engineering Research, 13(10), 7903-7912.
5. Kamley, Sachin. “Performance of Hybrid Ensemble Classification Techniques for Prevalence of Heart Disease Prediction.” VOLUME-8 ISSUE-10, AUGUST 2019, REGULAR ISSUE International Journal of Innovative Technology and Exploring Engineering, vol. 8, no. 10, Oct. 2019, pp. 1875–1882., doi:10.35940/ijitee.j9233.0881019.
6. Jan, Mustafa, et al. “Ensemble Approach for Developing a Smart Heart Disease Prediction System Using Classification Algorithms.” *Research Reports in Clinical Cardiology*, Volume 9, 2018, pp. 33–45., doi:10.2147/rrcc.s172035.
7. S. Dangare, et al. “A Data Mining Approach for Prediction of Heart Disease Using Neural Networks”.(2012).
8. Hsieh, Sheau-Ling, et al. “Design Ensemble Machine Learning Model for Breast Cancer Diagnosis.” Journal of Medical Systems, vol. 36, no. 5, Mar. 2011, pp. 2841–2847., doi:10.1007/s10916-011-9762-6.
9. Thomas, J., and R Theresa Princy. “Human Heart Disease Prediction System Using Data Mining Techniques.” 2016 International Conference on Circuit, Power, and Computing Technologies (ICCPCT), 2016, doi:10.1109/iccpct.2016.7530265.
10. Sultana, Marjia, et al. “Analysis of Data Mining Techniques for Heart Disease Prediction.” 2016 3rd International Conference on Electrical Engineering and Information Communication Technology (ICEEICT), 2016, doi:10.1109/ceeict.2016.7873142.
11. Swain, Debabrata, et al. “An Efficient Heart Disease Prediction System Using Machine Learning.” Machine Learning and Information Processing Advances in Intelligent Systems and Computing, 2020, pp. 39–50., doi:10.1007/978-981-15-1884-3_4.
12. Amin, Syed Umar, et al. “Genetic Neural Network Based Data Mining in Prediction of

Heart Disease Using Risk Factors." 2013 Ieee Conference On Information And Communication Technologies, 2013, doi:10.1109/cict.2013.6558288.

13. D. Dua and C. Graff, "Heart Disease Data Set." archive.ics.uci.edu/ml/datasets/Heart Disease
14. Liu, Xiao, et al. "A Hybrid Classification System for Heart Disease Diagnosis Based on the RFRS Method." Computational and Mathematical Methods in Medicine, vol. 2017, 2017, pp. 1–11., doi:10.1155/2017/8272091.
15. Sikora, Riyaz, and Ola Al-Laymoun. "A Modified Stacking Ensemble Machine Learning Algorithm Using Genetic Algorithms." Artificial Intelligence, pp. 395–405., doi:10.4018/978-1-5225-1759-7.ch016.