International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-11 Issue-2, July 2022

Applying and Improving Accuracy of Heart Disease Prediction Model using Meta-classifiers and Ensemble Learning Methods with Feature Selection

Uma K, M Hanumanthappa

Abstract: ***Healthcare industry is a significant sector for producing an enormous amount of data daily. The lack of helpful information is the primary motive for introducing machine learning or data mining techniques for extracting the required pattern needed to make a decision. Globally, heart disease is the leading cause of death. Prediction of heart disease early may help the survival of the patient life. This paper explores the machine learning technologies, ensemble learning, and meta-classifier to predict heart disease with feature selection methods to improve the accuracy. It presents a performance comparison between classifiers, ensemble learning methods, and meta-classifier.***

Keywords: Machine Learning, Ensemble Method, Heart Disease, Meta-classifier, Feature Selection Methods.

I. INTRODUCTION

Since 1999, coronary artery disease has been one of the top causes of death worldwide. It is increasing and has become a deadly disease with no limit. Compared to other diseases, coronary heart disease has caused many fatalities [1]. Coronary heart illness is a recent arrival on the global scene. Various countries' key risk factors impact heart disease death rates, particularly cholesterol levels, BP, smoking, exercise, and diet. 80 to 90 percent of persons who die from coronary heart disease have one or more main risk factors influenced by lifestyle, even though genetic factors play a role. The number of persons suffering from heart disease is currently increasing. Early and accurate diagnosis, followed by the appropriate course of treatment, can significantly prolong life. Heart disease is more likely to occur in certain age groups. According to a recent study, if the cardiac condition is detected early on, it can be effectively managed [1][2]. However, performing a flawless analysis is challenging due to numerous challenging elements associated with cardiac disorders. For instance, many symptoms may be present in addition to cardiac and heart diseases, which occur frequently.

Manuscript received on 30 June 2022 | Revised Manuscript received on 17 July 2022 | Manuscript Accepted on 15 July 2022 | Manuscript published on 30 July 2022.
***** Correspondence Author
Uma K*, Research Scholar, Department of Computer Science and Applications, Bangalore University, Bangalore (Karnataka), India.
Dr. M Hanumanthappa, Professor, Department of Computer Science and Applications, Bangalore University, Bangalore (Karnataka), India.

Due to this complexity, it is necessary to automate the process of medical diagnosis so that medical professionals can be helped during the diagnostic process. In the meantime, a reliable diagnostic procedure is required to find undiscovered and vital information in medical data, and data mining is becoming increasingly popular in the healthcare industry. Patient records can be analyzed using data mining to evaluate causes and symptoms and develop appropriate therapies. Additionally, it can discover clinical best practices to support the creation of norms and standards of care. In recent years, researchers have been more interested in one of the healthcare applications called treatment effectiveness, which identifies the disease pattern and symptoms and gives better treatment. For example, research on finding the causes of heart diseases and suggests the correct treatment. The diagnosis is the process of determining the disease or illness by examining the patients physically or through laboratory test centers [3]. The diagnosis is based on the history of the patient. In the medical field, the physicians make the diagnosis process every time the patient is admitted to the hospital, which is time-consuming and will affect the patient's condition. Developing reliable and effective medical decision support systems to assist the diagnosis decision process has become a more challenging task to decrease diagnosis time and increase diagnostic accuracy. Generally speaking, medical diagnosis is a complicated process, so the solution is to create an intelligent system. Ensemble learning methods are used, such as bagging, stacking, and boosting. It has demonstrated considerable promise for use in developing and implementing a heart disease prediction system [4].

II. ENSEMBLE LEARNING METHODS

Machine learning techniques combine the predictions from several individual models to produce an optimal model. Often called ensemble learning, this is a widely-used and preferred technique. There are three main types of ensemble learning methods are,

A. Bagging

Bootstrap aggregation is another name for bagging. Bagging can produce a diverse group of ensemble members by changing the training data to recruit a diverse group of ensemble members [4].

The machine algorithm uses a single unpruned decision tree, and each model is trained using a different sample from the same training dataset. The predictions generated by the ensemble members are then combined using basic statistics like voting or average. An influential part of this method is the way each sample of the dataset is prepared for training ensemble members. All models are trained using their own set of samples. The critical elements of bagging will work in each of the following steps.

- The training dataset is bootstrapped with small samples.
- Each model is fitted with unpruned decision trees.
- A simple vote or average of predictions is used.

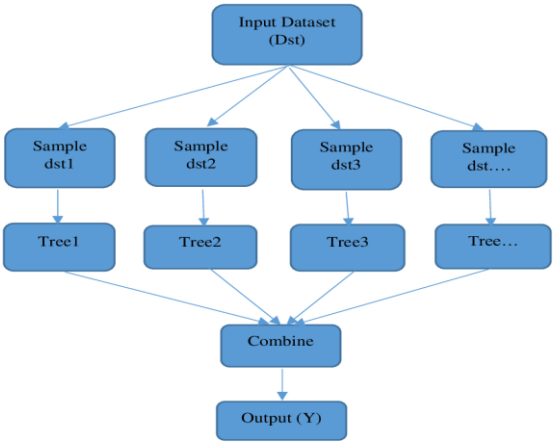

Fig.1. Bagging

Random forest is one example of bagging taking to build the ensemble method for predicting heart disease.

B. Boosting

The Boosting technique allows ensembles to learn from mistakes and make better predictions. The predictability of models is significantly increased when a weak base learner is paired with numerous other weak base learners to create a strong learner. As weak learners are grouped in sequence, weak learners learn from the next weak learners in the series, resulting in better prediction models [4].

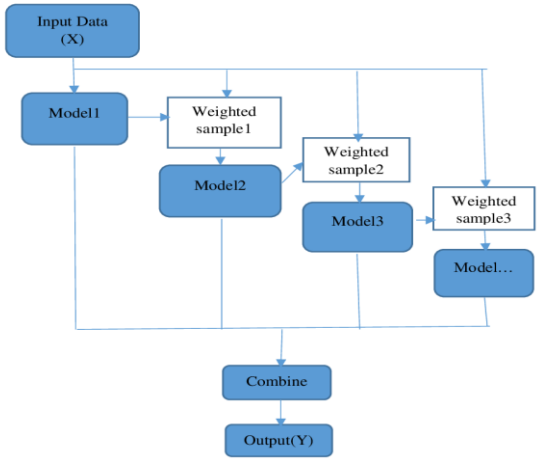

Fig.2. Boosting

Different kinds of boosting exist, including Extreme Gradient Boosting, which is also named as XG Boost, and Adaptive Boosting, also known as AdaBoost & Gradient Boosting.

AdaBoost: It employs weak learners in the form of decision trees, which typically have one split and are known as "decision stumps" or weak learners. Observations with comparable weights make up the central decision stump in AdaBoost.

Gradient Boosting: Predictors are successively added to the ensemble with Gradient Boosting, where the predictors preceding their successors correct their errors, thus enhancing the model's accuracy. Errors in the previous predictors are corrected with new predictors. It assists the gradient booster in identifying and countering problems in predictions made by learners.

XGBoost: It uses boosted gradient decision trees to improve performance. A large portion of the performance depends on the computing speed and performance of the target model. Gradient boosted machines are implemented in a sequential manner, which makes them slow.

C. Stacking

Stacking generalizations is another ensemble method. This technique can be grouped in a training algorithm with various similar learning algorithms. In addition to Regression, density estimations, distance learning, and classification, stacking has been used in multiple applications. Stacking can also be used to determine bagging error rates.

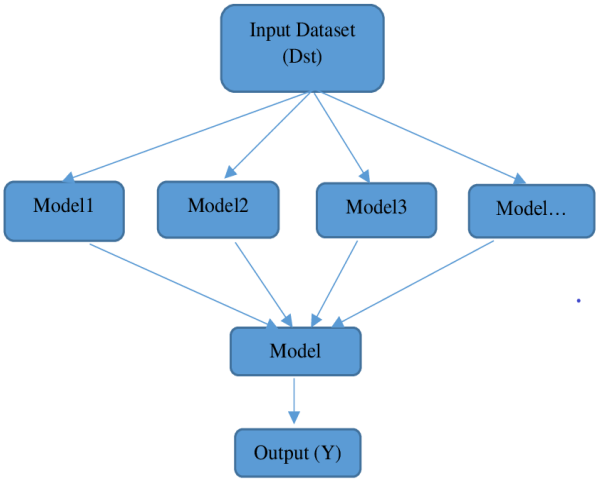

Fig.3.Stacking

III. RELATED WORK

C. Beulah Christalin Latha and S. Carolin Jeeva [2019] have worked on predicting the risk of heart disease and improvisation of the accuracy of the model using the ensemble method [5]. Initially, the authors used machine learning algorithms to perform the prediction task and moved on to employing ensemble machine learning methods to improve prediction accuracy. Such as bagging and boosting with the feature selection method, and they achieved 85.48%. Asma Baccouche et al. [2020] has been worked on a Mexican heart disease case study using ensemble deep learning models [6].

Retrieval Number: 100.1/ijrte.B71890711222
DOI: 10.35940/ijrte.B7189.0711222
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

The researchers used the Mexican heart disease dataset, which consists of 800 records with 141 attributes. The authors selected the Neural Network, CNN, Deep learning, feature selection, and Sampling techniques to improve the model performance. Finally, they achieved 91 to 94% accuracy by applying different methods. Mohammed Nasir Uddin, and Rajib Kumar Halder [2021], proposed machine learning ensemble methods based multilayer dynamic system to predict the risk of cardiovascular disease [7]. The researchers combined the four similar datasets into a single and trained dataset and then split the data into five different ratio types to perform the task. The authors applied three techniques, Random Forest, Naïve Bayes, and Gradient Boosting, by using the voting method and accompanied 94% accuracy by their MLDS method.

IV. APPROACH METHODOLOGY

The research aims to increase the model's predictive power for heart disease. Ensemble learning approaches will be more accurate when predicting cardiac disease than individual classification methods.

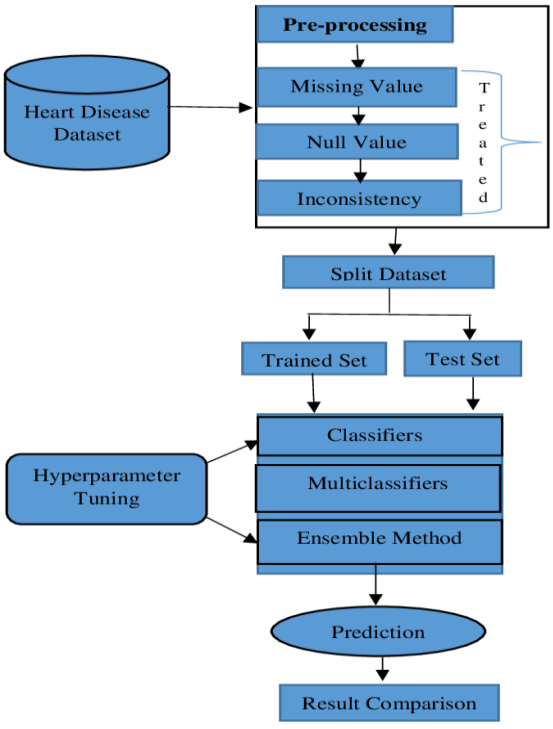

Fig.4.Approach Methodology.

Data Collection: The heart disease dataset was collected online from UCI (University of California, Irvin), Machine Learning Repository. The instance format and attributes in the dataset are the same. These datasets have one predicted characteristic in addition to 76 natural features. Only 14 of the 76 qualities are crucial to predicting the disease. After data collection, the data should be put into a data warehouse as target data that are domain specific. The final dataset consists of 303 records with 13 and one class attribute.

Data Pre-processing: This data mining step entails data cleansing, data reduction, data integration & transformation. This heart disease dataset contains noisy, inconsistent, and missing values. Missing values, inconsistency, and noisy values are treated and cleansed.

Sl No.	Attribute	Values	Description
1	Age	29-62(Numerical value)	Age of patient in years
2	Gender	0-Male ; 1-Female	Gender of Patient
3	Cpt	1 - typical angina ; 2 - atypical angina ; 3 - non-anginal pain ; 4 – asymptomatic	Type of chest pain
4	Restbp	Numerical value In mm/ Hg(140mm/Hg)	Resting blood pressure
5	Chol	Numerical value inmg/dl (289mg/dl)	Cholesterol in mg/dl
6	Fbs	0– false ; 1– true	Fasting blood sugar >120 mg/dl (1 = true; 0 = false)
7	Restecg	0 - normal; 1- having ST-T2 – hypertrophy	Resting electrocardiographicresults
8	thalach	Numerical value (140, 173)	Maximum heart rate achieved
9	Exang	0 – no; 1– yes	Exercise induced angina (1 = yes; 0 =no)
10	oldpeak	Numerical value	ST depression induced by exercise relative torest
11	Slope	1 – upsloping; 2- flat ; 3– downsloping	The slope of the peakexercise ST segment
12	Ca	0 – 3 vessels (numeric value)	Number of major vessels (0-3) colored by fluoroscopy
13	Thal	3– normal ; 6 - fixed defect ; 7 - reversable defect	Thalassemia

Table.1. Heart disease data attribute description.

The data is then separated into training and test data sets, which are then applied to different algorithms to get accuracy score results. Five classification techniques were applied in the first stage to predict heart disease. The five classifiers are Logistic Regression (LR), K-Nearest Neighbour (KNN), Naive Bayes (NB), Decision Tree (DT) & Support Vector Machine (SVM). The model achieves high accuracy using logistic regression classifier. In the second stage, experiment by combining the classifiers to improve accuracy among five classifiers. Finally, ensemble methods are used to predict heart disease with the feature selection method to improve accuracy.

V. EXPERIMENTAL SETUP AND RESULT DISCUSSION

Python pandas were used with importing libraries to experiment with heart disease prediction, and hyperparameter tuning was applied to select the best features. The experiment was conducted in three stages.

Stage 1: In the first stage, individual classifier techniques are applied to predict heart disease. The classification techniques SVM, LR, DT, KNN, and NB were used with and without feature selection methods.

Table.2. Performance of Classifiers.

Classifiers	Accuracy %		Accuracy % (hyperparameter tuning)	
	Train	Test	Train	Test
LR	93.24	92.86	-	
KNN	91.32	88.52	86.36	86.88
NB	83.88	81.96	85.24	83.88
SVM	90.12	84.35	88.03	87.32
DT	100	70.00	77.38	62.49

In the first stage of the experiment, various machine learning techniques are used with and without the hyperparameter tuning strategy to predict cardiac disease. From the above table value, the Decision tree gives the cent percent on trained data and 70% on test data, typically with hyperparameter tuning, and the same DT gives 77.38% and 62.49% for train and test data.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

LR achieves 93.24% and 92.86% for train and test data without feature selection. Logistic Regression does not needed hyperparameter tune. KNN reaches 91.32% and 88.52% for train and test data. With hyperparameter tuning, KNN comes 86.36% and 86.88% accuracy for train and test data. The accuracy of KNN without hyperparameter tuning is only better than hyperparameter tuning. Next, NB gives improved accuracy with hyperparameter tuning, i.e., 85.24% and 83.88% for train and test data compared to 83.88% and 81.96%. In SVM, hyperparameter tuning data gives improved accuracy, i.e., 87.32% compared to 84.35% for test data. For the train, data SVM gives 90.12% and 88.03% without and with hyperparameter tuning.

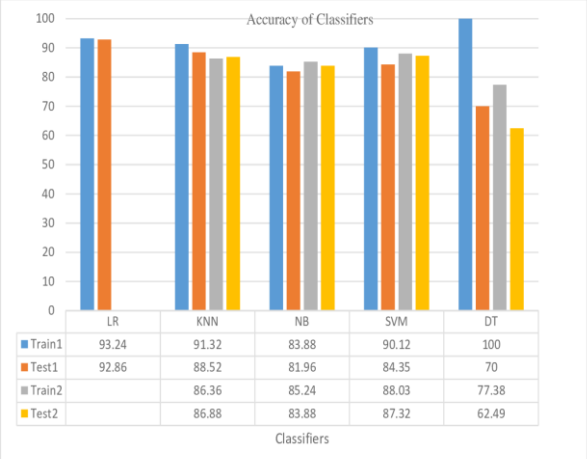

Fig.5. Classifiers performance chart.

In the given chart, without tweaking the classifier's hyperparameters, train1 and test1 demonstrate their accuracy. It is evident from Train 2 and Test 2 how accurate hyperparameter tuning works.

Stage 2: In the second stage, Meta classifier techniques are used based on hard voting and soft voting to predict heart disease. The classification techniques SVM, LR, DT, KNN, and NB are applied to the experiment. LR is combined for every classifier to get better accuracy than a single classifier. Finally, all classifiers are blended to achieve the highest accuracy.

Table.3. Performance of meta-classifiers.

Meta_classifiers	Accuracy% (Soft Vote)		Accuracy % (Hard Vote)	
	Train	Test	Train	Test
LR + KNN	89.13	86.88	90.15	88.52
LR + NB	84.86	83.60	86.35	85.20
LR + SVM	85.04	82.13	88.90	83.48
LR +DT	100	80.32	91.23	85.24
LR + KNN + NB + SVM +DT	90.24	81.96	91.05	86.88

In this experiment, meta-classifiers are used to predict heart disease with two types of voting techniques: hard voting and soft voting. A meta-classifier is a classifier that uses all previous predictions as features to determine a single final prediction. Consequently, it chooses the last class as the outcome that requires from among those predicted by different classifiers. Hard voting involves choosing the prediction that has received the most support. In contrast, the soft voting method gives the highest overall probability by adding each prediction's possibilities in each model. From the given table, the accuracy of the hard voting method leads to the soft voting method.

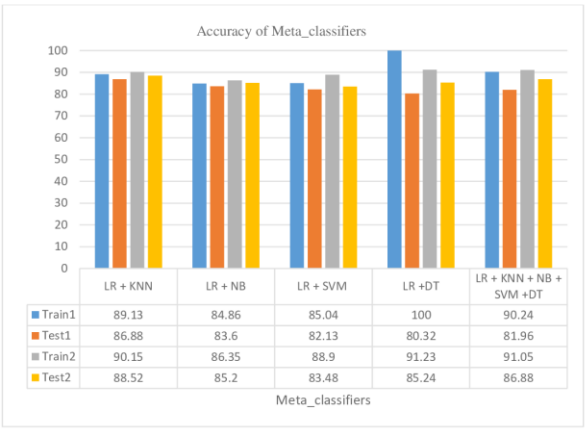

Fig.6. Meta-classifier performance chart.

The chart, train1, and test1 show the accuracy of the meta-classifiers with soft voting type. Train2 and test2 show the accuracy with the hard voting type.

Stage 3: In the final stage, to increase the model's accuracy, ensemble learning techniques, including stacking, boosting & bagging, are used.

Table.4. Performance of Ensemble methods.

Ensemble Method	Algorithms	Accuracy %		Accuracy % (Hyperparameter tuning)	
		Train	Test	Train	Test
Bagging	Random Forest	75	75	89	84
Boosting	AdaBoost	93	75	88	84
	GradientBoost	100	77	100	80
	XGBoost	98	80	92	89
Stacking	Logistic, SVM, KNN, NB, DT	87	84	-	

This experiment demonstrates the improved accuracy of the training dataset compared to test data. The Gradient boosting technique takes 100% to train the data in both tests. XGBoost also performs well compared to all other ensemble learning methods.

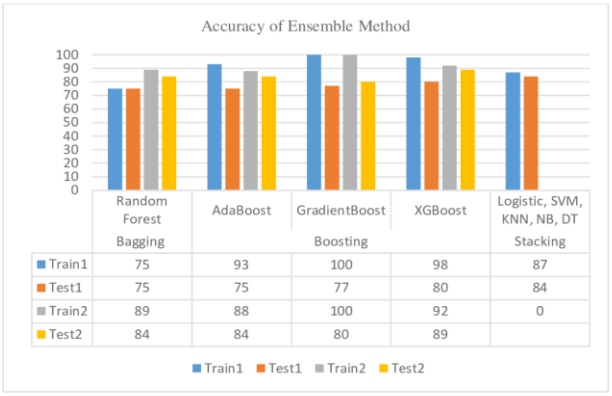

Fig.7. Ensemble learning method performance chart.

In the given chart, train1 and test1 indicate the accuracy of the ensemble learning methods without any feature selection. Train2 and test2 show accuracy with a soft voting method. Train2 and test2 demonstrate the accuracy with hypermeter tuning.

Retrieval Number: 100.1/ijrte.B71890711222
DOI: 10.35940/ijrte.B7189.0711222
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

VI. CONCLUSION

Globally, heart disease is the primary cause of death. Effective disease treatment depends on early disease identification. This study proposes an approach to predict heart disease more effectively. The research was carried out in three stages, initially with individual classifiers and a hyperparameter tuning method. Then, meta-classifiers are used with hard and sift voting types to improve the accuracy. Lastly, ensemble learning methods are used with and without hyperparameter tuning. Individual classifier Decision Tree, meta-classifier LR and DT, and Gradient Boosting give the cent percentage on a trained dataset. Logistic Regression achieves 93% accuracy on the test dataset. This method also can serve as a tool for predicting heart disease risk and providing clinical advice efficiently.

REFERENCE

1. World Health Organization 2016 [online], http://www.who.int/cardiovascular_diseases/en/
2. Deepali Chandra, "Diagnosis of Heart disease using Data Mining Algorithm" International Journal of Computer Science and Information Technologies, Vol(2), 2014, 1678-1680.
3. Hian Cbye Kob et al., “Data mining Applications in Healthcare” Journal of Healthcare Information management-Vol.19, No.2.
4. Cha Zhang and Yunqian Ma, “Ensemble Machine Learning Methods and Applications”, Springer New York, NY, 2012, https://doi.org/10.1007/978-1-4419-9326-7. [CrossRef]
5. C. Beulah Christalin Latha and S. Carolin Jeeva,” Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques”, Informatics in Medicine Unlocked, Volume 16, 2019, 100203, https://doi.org/10.1016/j.imu.2019.100203. [CrossRef]
6. Asma Baccouche, , Begonya Garcia-Zapirain , Cristian Castillo Olea and Adel Elmaghraby, “Ensemble Deep Learning Models for Heart Disease Classification: A Case Study from Mexico”, Information 2020, 11, 207; doi: 10.3390/info11040207 www.mdpi.com/journal/information. [CrossRef]
7. Mohammed Nasir Uddin, Rajib Kumar Halder, “An ensemble method based multilayer dynamic system to predict cardiovascular disease using machine learning approach”, Informatics in Medicine Unlocked, Volume 24, 2021, 100584, https://doi.org/10.1016/j.imu.2021.100584. [CrossRef]

AUTHORS PROFILE

Uma K, Research Scholar, pursuing a Ph.D. in the Department of Computer Science and Applications, Bangalore University, and also completed her master’s degree from Bangalore University. Currently working as a research scholar. Her area of research interests is data mining and machine learning.

Dr. M Hanumanthappa, working as a Professor, in the Department of Computer Science and Applications, Bangalore University, Bangalore, India. Pursuing research in the area of Information and Retrieval, Data Mining, Network Security, and Natural Language Processing. He is having more than 18 years of Teaching, Administration, and Industry Experience.