DOI: 10.15849/IJASCA.220720.03

Int. J. Advance Soft Compu. Appl, Vol. 14, No. 2, July 2022
Print ISSN: 2710-1274, Online ISSN: 2074-8523

Improving the Prediction of Heart Disease Using Ensemble Learning and Feature Selection

Priyanka Gupta[1], D.D. Seth[2]
[1]Computer Science & Engineering, SRM Institute of Science & Technology, SRM University, India
e-mail:priyanka6gupta@gmail.com
[2]Computer Science & Engineering, SRM Institute of Science & Technology, SRM University, India
e-mail:ddgayatri@yahoo.co.in

Abstract

Heart or cardiovascular disease is main cause of mortality. The main objective of developing the proposed model is to increase the accuracy and reliability of predicting the coronary heart disease. This paper attempts in predicting the risk of heart disease more accurately using the techniques of ensemble learning. Moreover, the techniques of feature selection and hyper parameter tuning has been implemented in this work leading to further increase in accuracy. Among the three ensemble techniques, stacking, majority voting and bagging used in this work, the improvement achieved in prediction accuracies is 2.11%, 7.42% and 0.14% respectively. Majority voting has shown the best results in terms of increase in prediction accuracies with an accuracy of 98.38%.

Keywords: *Heart Disease, Ensemble Learning, Feature selection, Machine Learning.*

1 Introduction

Heart disease is the major cause of death in people all over the world. To detect and diagnose the heart disease is a big challenge [1]. Computer Aided Detection (CAD) helps in automatic detection of heart disease. With the advent of machine learning it becomes easy and convenient to analyze medical data. There is need to optimize the performance of machine learning algorithms [2]. Ensemble learning method provides the solution by improving the performance and providing better and more accurate results in early detection of heart disease.

Received 27 February 2022; Accepted 10 June 2022

Ensemble learning is a technique of machine learning which helps in improving the performance of our system by using multiple classifiers. The ensemble learning structure includes bagging, boosting, and stacking methods. Out of these, bagging and boosting are built using the same type of classifiers whereas stacking is constructed using a different type of learners [3]. By combining multiple classifiers, the performance of the model increases to a great extent as compared to the individual classification model. Thus, using ensemble learning enhances the accuracy of prediction for detecting heart disease.
The rest of the paper is structured in the following sequence. Section 2 consists of review of literature related to the work. It contains the existing methods and discusses the techniques available. Section 3 consists of the proposed work and methodology. The results of the experiments are discussed in Section 4. Lastly in Section 5 conclusion is given.

2 Related Work

In [2], Mio et al (2016) have presented a comparative study by implementing the ensemble technique for predicting the coronary heart disease on 4 different datasets like Switzerland University Hospital (SUH), Long Beach Medical Center (LBMC), Hungarian Institute of Cardiology (HIC) and Cleveland Clinic Foundation (CCF). In [4], Yekkala et al (2017) have presented a comparative analysis of various ensemble methods like bagging, boosting (AdaBoost), and random forest. Particle swarm optimization (PSO) was used for feature selection. The bagging tree has achieved the best results. In [5], Mohan et al (2019) have proposed a hybrid model combining random forest along with the linear model (HRFLM) for prediction of heart disease. The author has performed a comparative analysis of various algorithms with the hybrid model. The proposed model has shown the highest accuracy when compared with other individual classifiers. In [6], Latha et al (2019) have focused on enhancing the accuracy of prediction of weak classifiers by using ensemble techniques like bagging, boosting, stacking and majority voting. In [7], Sarkar et al (2019), have proposed the hybrid model development (integrating GA and PRISM learner). The accuracy of prediction of the proposed hybrid model exceeds around 6 percent than that of the sequential GA-based hybrid model. In [8], Mieyne et al (2020) have proposed an ensemble model which was formed using different CART models based on the weighted ageing ensemble (WAE) classifier for the prediction of heart disease. Datasets used were Cleveland and Framingham datasets. Dataset was partitioned randomly using the mean splitting-based approach. In [9], Kumar et al (2020) presented an ensemble classifier using k nearest neighbor (KNN), support vector machine (SVM), modified k nearest neighbor (M-KNN), and CART decision tree algorithm for prediction of heart disease. The ensemble method achieved the highest accuracy when compared with individual classifiers. In [10], Ali et al (2020) have proposed a monitoring system for smart healthcare for prediction of heart disease. The techniques of feature fusion and ensemble learning were used.

Features extracted from sensor data and Electronic Medical Record (EMR) were combined using feature fusion. Information technique was used for selecting important features and removing the irrelevant features. They have compared proposed ensemble model with traditional machine learning models and achieved higher accuracy. In [11], Sri (2020) has presented a comparative study of various algorithms like support vector machine, naive bayes, logistic regression, neural network, and voting classifier for heart disease prediction. The voting classifier has achieved the highest accuracy. The author also proposed the development of GUI. In [12], Mienye et al (2020), have proposed a sparse autoencoder-based artificial neural network for detecting the heart disease. The authors have compared the performance measures of the proposed model with the artificial neural network (ANN). The proposed model has higher accuracy and better results. The optimizer used was adam and batch normalization was applied. In [13] Xiao et al (2020), have proposed a deep residual neural network (DRNN). The accuracy approached 95%, better than the various machine learning algorithms like decision tree 68%, logistic regression 87%, naive bayes 80%, k-nearest neighbors (KNN) 60%, and random forest 83%. Similarly, in [14], Chowdhary and Singh (2020) have used the decision tree (DT) algorithm. The ada-Boost algorithm was utilized to optimize the output of the decision tree. At max_depth = (6), model was showing maximum accuracy. The accuracy of the ada-boost algorithm was 0.89. In [15], Bhatia et al (2020), have proposed model that works on ensemble learning that is stacking, boosting and bagging algorithms. Stacking, boosting and bagging works on a hybrid algorithm which test and trains the dataset. In this system, an application is developed consisting of 2 modules: patient and doctor's login. The doctor module records the case history and case details. In patient login, the medical history of the patient is visible. The ensemble technique uses the combination of weak learners for implementing the hybrid model. In [16], Lakshmanrao et al (2021) have proposed an ensemble model for heart disease prediction. Three Sampling techniques were used to balance the imbalanced dataset. The selection of features was done using two methods, ANOVA and mutual information methods. Experiments are performed on 2 different datasets, UCI and Kaggle dataset. In [17], Yuan et al (2020) have proposed the technique of hybrid gradient boosting decision tree along with logistic regression (HGBDTLR) for prediction of heart disease. In [19]-[20] authors have proposed a majority voting technique for predicting heart disease. In [21] Kannan and Vasanthi (2019) have compared the prediction accuracies and ROC of different machine learning classifiers like random forest, logistic regression, stochastic gradient boosting and support vector machines for predicting heart disease. In [22] Pillai et al (2019) have proposed the use of recurrent neural network to predict heart disease. In [23] Bhat et al (2019) have employed the use of artificial neural network with backpropagation for prediction of heart disease. In [24] Repaka et al (2019) have proposed the smart heart disease prediction by utilizing naïve bayes algorithm for the prediction of heart disease and AES for providing the security. In [25] Ricciardi et al (2020) have employed

the use of a combination of linear discriminant analysis and principal component analysis for predicting the heart disease.

By reviewing the works performed by various authors it has been found that the performance of the existing systems is comparatively less. So, in this work, we presented a model which will try to increase the efficiency as well as the performance of the system using the techniques of ensemble learning, feature selection, and hyperparameter tuning.

3 Proposed Work and Methodology

In the proposed work we have used an ensemble learning approach for increasing the prediction accuracy of predicting the risk of coronary heart disease. In our proposed system we have created an ensemble of various individual classifiers (Support Vector Machine, Decision Trees, K Nearest Neighbors, Random Forest, and Gradient Boosting). We have used various ensemble techniques (majority voting, stacking, and bagging). The performance of these techniques has been compared. Moreover, in this paper, the performance of the classifiers is enhanced further by utilizing the technique of feature selection and hyperparameter tuning.

3.1 Dataset Description

The dataset used in this study is the Framingham heart study dataset from Kaggle. The dataset has 4240 instances and 15 attributes [18] which are described in the Table 1. Table 1 shows the dataset attributes with their definitions.

Table 1: Description of dataset

S No.	Attributes	Description
1	Sex	0: male,1: female
2	Age (In Years)	Age of Patient in years
3	Current Smoker	1: If the patient is a Current Smoker 2: If the patient is a Non-Smoker
4	Cigs Per Day	Number of Cigarettes the person smokes in a day
5	BPMeds	Patient on BP Medication 0: If the patient not on BP Medication 1: If the patient is on BP Medication
6	Prevalent Stroke	The patient had a previous stroke 0: If the patient is not having a previous stroke 1: If the patient is having a previous stroke
7	Prevalent Hyp	The patient is Hypertensive or not 0: If the patient is not Hypertensive 1: If the patient is Hypertensive
8	Tot Chol	Total Cholesterol level
9	Sys BP	Systolic Blood Pressure
10	Dias BP	Diastolic Blood Pressure

11	Diabetes	The patient is Diabetic or Non-Diabetic
12	BMI	Body Mass Index
13	Heart Rate	Rate of Heart
14	Glucose	Glucose level
15	Education	Education of Person
16	Ten Year CHD	Target-10 Year CHD Risk 0: No Risk of CHD 1: Risk of CHD

The Ten Year CHD signifies the target attribute. It is categorized into two classes '0' denotes the absence of risk of coronary heart disease (CHD) and class '1'denotes the presence of coronary heart disease (CHD).

3.2 Data handling

3.2.1 Data collection

In this work the dataset utilized is the Framingham dataset.

3.2.2 Data Preprocessing

The data was pre-processed first in which data was cleaned, null and missing values were handled. Any inconsistencies present in the data were removed. The Framingham dataset utilized in this work comprises of missing values that needs to be handled. The dataset was checked for any duplicate values.

3.2.3 Outlier detection and elimination

Outlier detection is a technique to identify the observation in the dataset that do not follow the normal pattern and lies far away from the other values. These are called outliers. The accuracy of the model gets improved by eliminating the outliers. In this work, box plots were used for detecting the outliers. Outliers were observed in two columns, total Cholesterol, and systolic BP columns. The outliers from these columns were removed.

3.2.4 Data balancing

The dataset was now checked whether it is balanced or not. It was observed that the number of negative cases were much larger as compared to positive cases. Therefore, the dataset was highly imbalanced and needs to be balanced otherwise it will cause a problem when the fitting of the model is done. The Random Over Sampling technique was utilized here to balance the dataset. The number of instances in the minority class were increased so that they become equal to the majority class.

3.3 Chi-square test based feature selection

Feature selection is important as it leads to an increase in the prediction accuracy of the model. The reduced number of features makes the model lightweight. Moreover, it requires less time for training and requires less memory.

In this paper, the chi-square test method was utilized for selecting the features from the dataset. The dataset contains 15 attributes. Chi-square test scores were found for the 10 best features in the dataset. The best features were selected based on these scores which are displayed in the descending order. Out of 10 features, 8 best features were considered. Fig. 1 has shown the various features according to their chi-square test scores in descending order.

Feature	Score
sysBP	2121.922128
glucose	1232.342416
age	1006.482991
cigsPerDay	788.788750
totChol	769.066248
diaBP	486.662514
prevalentHyp	221.104823
sex	66.775610
BPMeds	66.216216
diabetes	54.258065

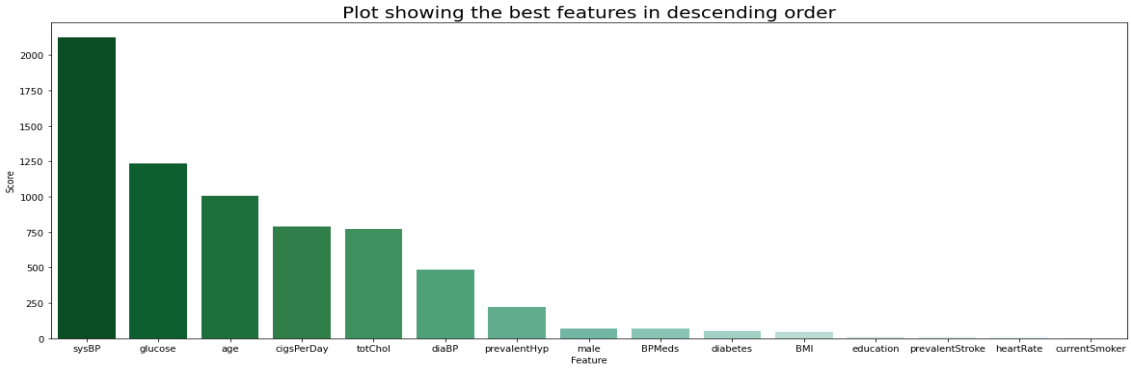

Fig. 1 Feature selection using chi square test

3.4 Ensemble methods

Various Ensemble Algorithms [19]-[20] are described below as:

3.4.1 Stacking

It is an ensemble method which uses meta classifier for combining multiple classifiers. Stacking consists of several layers and each layer passes its prediction to the above and finally the topmost layer model makes the decision based on models in the below layers. The bottom layer gets the input from the dataset.

3.4.2 Majority Voting

Majority voting accumulates the findings of each classifier and output class is predicted on the majority of votes. Voting classifier is trained on various models and predicts the output based on class having highest probability. In this, single ensemble model is created which is trained by individual classifiers to predict output for each output class based on their combined majority of voting.

3.4.3 Bagging

Bagging is an ensemble method that takes random subset of data from original dataset. It accumulates the performance of individual classifiers to give final output. This technique is used for weak learners which has high variance and low bias. It consists of 3 steps: bootstrapping, parallel training & aggregation. First different subsets of data are created by selecting the data points randomly and with replacement. These datasets are referred to as bootstrap replicates. Next these data subsets are trained in parallel and independently. Finally, the result from each of the classifiers is aggregated based on averaging or majority voting. Bagging can be implemented easily and helps to reduce variance.

3.5 Methodology

The first step was to collect the data. The dataset used here was the Framingham dataset. The pre-processing of data was done so that missing and invalid data are taken care of. The dataset was checked for any duplicate values. If duplicates are present, they should be removed. Then the outliers present in the data were detected and handled. The dataset used in this work was not balanced. So, the dataset was balanced using the technique of oversampling in which the minority class instances were increased.

The three-fold cross-validation was used. The dataset utilized for training was 70% and for testing was 30%.

First, all the 15 features of the dataset were used. The accuracies of individual classifiers, Random Forest, Support Vector Machine, Decision Tree, K Nearest Neighbors, and Gradient Boosting were calculated. Ensemble techniques were used for improving the performance of the classifiers. Various ensemble methods utilized were majority voting, stacking, and bagging. The individual classifiers were ensembled together using majority voting and stacking which improves the performance of the classifiers. The stacking algorithm combines the individual classifiers using meta classifier. The bagging algorithm utilizes the decision tree classifier as the base classifier for improving the performance.

After that the feature selection was performed. The features selection took place based on the importance of features. Eight best features were selected. Then the hyperparameter tuning of the four best performing classifiers, K Nearest Neighbors, Gradient Boosting, Random Forest and Decision Tree was performed. Now ensemble techniques were utilized again in a similar manner. This led to a further increase in prediction accuracies. The highest increase in prediction accuracy was observed in the Majority voting technique.

The architecture of the proposed system is given in Fig. 2.

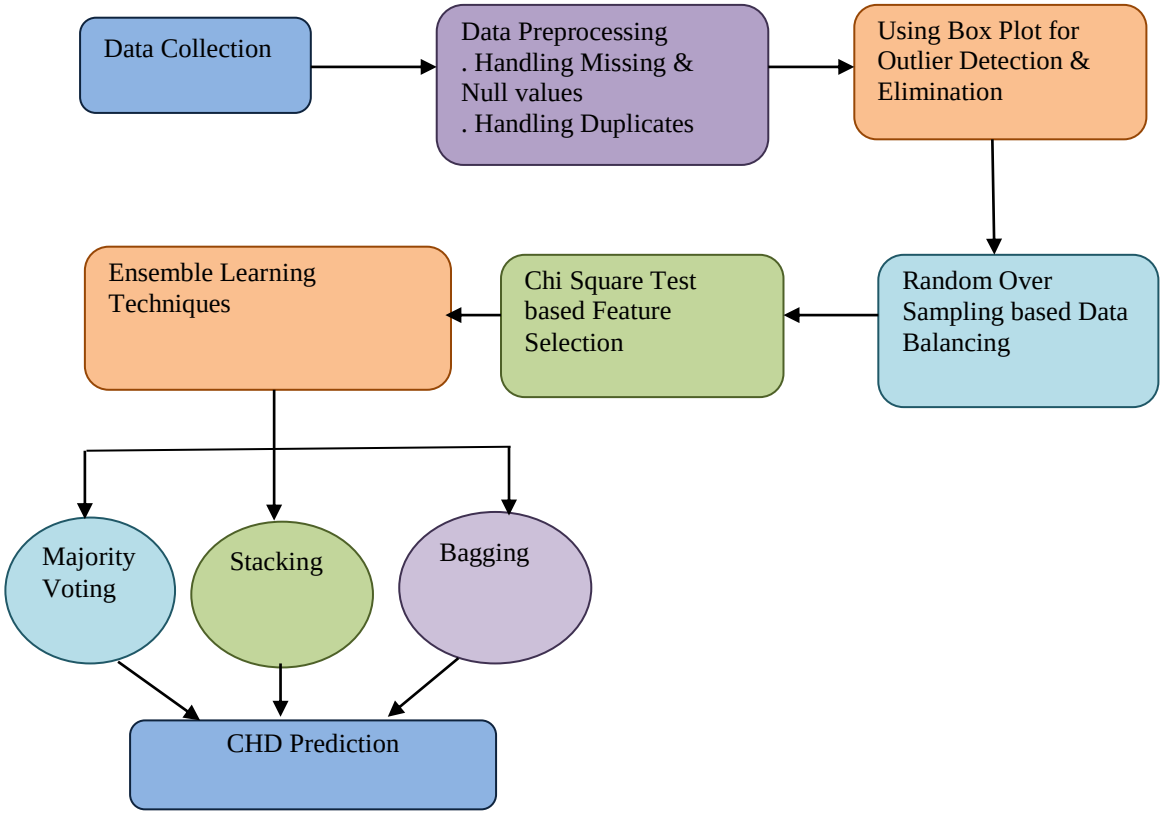

Fig. 2 Architectural diagram of proposed system

4 Results and Discussion

4.1 Performance of the classifiers with ensemble techniques

Various classification algorithms were analyzed and compared for their performance on the Framingham dataset. Some classifiers have shown good performance while other classifiers have shown poor performance.

Table 2: Performance measures of classifiers without ensemble techniques and feature selection

Algorithms	Precision	Recall	F1-Score	Accuracy
Support Vector Machine	0.67	0.70	0.69	67.45
Naïve Bayes	0.66	0.51	0.57	61.76
K Nearest Neighbor	0.87	0.99	0.92	91.70
Decision Tree	0.83	0.99	0.90	89.297
Random Forest	0.88	0.97	0.92	91.95
Gradient Boosting	0.71	0.75	0.73	72.36

Ensemble techniques were used for improving the performance of the algorithms. The algorithms used were support vector machine, random forest, decision tree, k nearest neighbors, and gradient boosting. The performance metrics are shown. (Table 3, Fig. 3 & 4)

Table 3: Performance measure of classifiers using ensemble techniques and without feature selection

Algorithms	Precision	Recall	F1-Score	Accuracy
Stacking	0.94	0.98	0.96	96.27
Majority Voting	0.86	0.99	0.92	90.96
Bagging	0.87	0.97	0.92	91.16

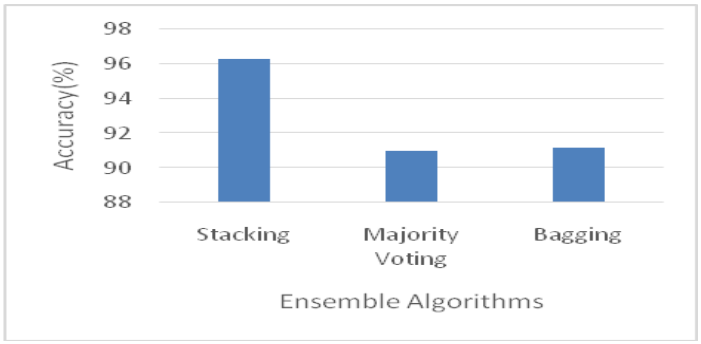

Fig. 3 Comparison of ensemble classifiers for their prediction accuracies without feature selection

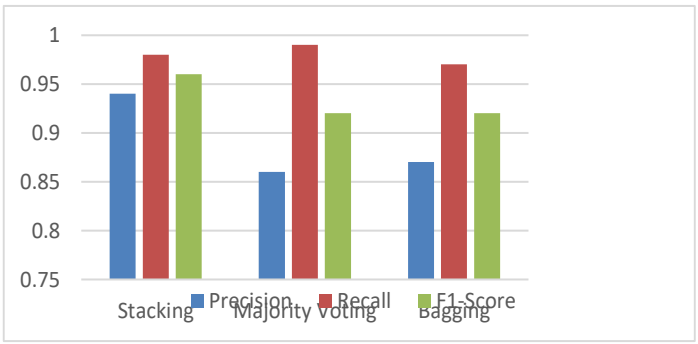

Fig. 4 Comparison of Ensemble Classifiers for their Performance Measures without Feature Selection

4.2 Performance enhancement with feature selection

The performance of the classifiers was further enhanced by using the technique of feature selection and hyper parameter tuning. The selection of the features takes place according to their importance. The importance of various attributes was found using Chi Square Test method and the attributes which have more impact on heart disease prediction were taken into consideration. The feature set was constructed, and the performance measures were evaluated. Features such as age

and gender signify the personal information of the patient and the remaining 13 features represents the physiological information about the patient.
The feature set was created using the eight best features. These features in the feature set are described as follows:
Feature set = { sysBP, glucose, age, CigsPerDay, totChol, diaBP, prevalentHyp, sex }
The improvement in the performance of ensemble techniques (stacking, majority voting, and bagging) by using the feature selection technique and hyperparameter tuning is given. (Table 4, Fig. 5 & 6)

Table 4: Performance measure of classifiers using ensemble techniques and with feature selection

Algorithms	Precision	Recall	F1-Score	Accuracy
Stacking	0.98	0.98	0.98	98.379
Majority Voting	0.98	0.98	0.98	98.379
Bagging	0.87	0.98	0.92	91.3

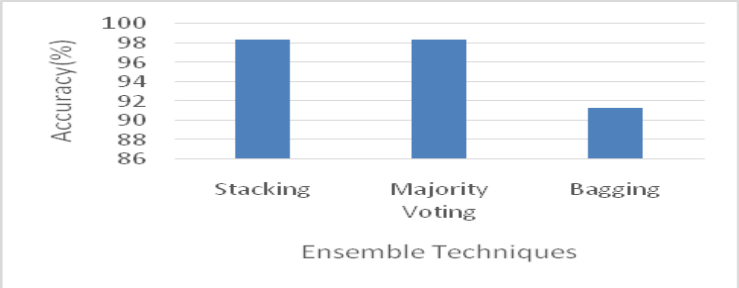

Fig. 5 Comparison of ensemble classifiers for their prediction accuracies using feature selection

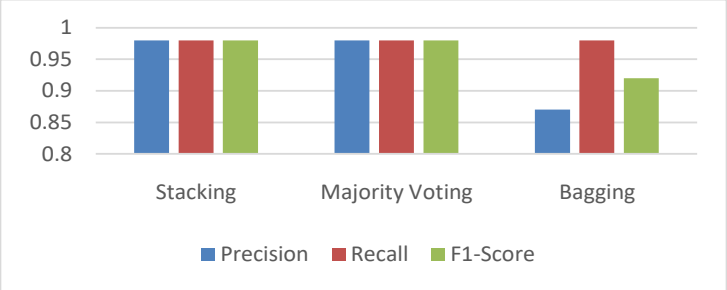

Fig. 6 Comparison of ensemble classifiers for their performance measures using feature selection

Ensemble techniques helped in predicting the risk of heart disease with higher accuracy as compared to individual classifiers. The accuracy of the model was further enhanced by using the techniques of feature selection and hyperparameter tuning. Out of three ensemble classifiers, stacking, majority voting, and bagging, the Majority voting has shown the highest improvement of 7.42% in prediction accuracy with feature selection. Whereas, in the case of stacking the rise in the

accuracy of 2.11%, and for bagging the rise in the accuracy of 0.14% was achieved using the technique of feature selection.

5 Conclusion

Heart or cardiovascular disease is the main cause of mortality. The purpose of developing the presented model was to increase the reliability, robustness, and accuracy of predicting the risk of coronary heart disease. In this work, an ensemble learning approach was used for accurate, reliable, and early detection of coronary heart disease. Various ensemble techniques were used for the prediction of heart disease like majority voting, stacking, and bagging for improving the performance of the classifiers. It has been shown that the performance of the classifiers has improved further by using the techniques of feature selection. Out of 3 Ensemble classifiers i.e., stacking, majority voting and bagging. Majority voting has shown the best results in terms of increase in prediction accuracies with an accuracy of 98.38%**.**

References

[1] Kim, J.K. & Kang, S. (2017). Neural Network-Based Coronary Heart Disease Risk Prediction Using Feature Correlation Analysis. *Journal of Healthcare Engineering*," vol. 2017, Article ID 2780501, pp.1-13.

[2] Miao, K. H., Miao, J. H. and G. J. Miao. (2016). Diagnosing Coronary Heart Disease Using Ensemble Machine Learning. *International Journal of Advanced Computer Science and Applications, (IJACSA)*, 7(10), 30-39.

[3] Li H, et al. (2018). Ensemble learning for overall power conversion efficiency of the all-organic dye-sensitized solar cells. *IEEE Access* 2018; 6:34118–26.

[4] Yekkala, I., Dixit ,S. and Jabbar, M. A. (2017). Prediction of heart disease using ensemble learning and Particle Swarm Optimization. In *Smart Technologies for Smart Nation (SmartTechCon). 2017 International Conference On* (pp. 691-698). IEEE.

[5] Mohan, S., Thirumalai, C. and Srivastava, G. (2019). Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. *IEEE Access*, 7, 81542-81554.

[6] 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, 1-9

[7] Sarkar, B.K. (2020). Hybrid model for prediction of heart disease. *Soft Computing* 24, 1903–1925.

[8] 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(100402),1-5.

[9] Kumar, R.S., Fatima, S.S., Thomas, A. (2020, May). Heart Disease Prediction using Ensemble Learning Method. *International Journal of Recent Technology and Engineering (IJRTE)*, ISSN: 2277-3878, 9 Issue-1, 2612-2616.

[10] F. Ali, S.P. Sappagh, S.M.R. Islam, D. Kwak, A. Ali, M. Imran, K.S. Kwak, (2020, November). A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion. *Information Fusion (2020)*, 63, 208-222.

[11] Sri, B.U. (2020). Effective Heart Disease Prediction Model Through Voting Technique. *International Journal of Engineering Technology and Management Sciences (IJETMS)*, Issue:5, 4, 10-13, September.

[12] Mienye, I. D., Sun, Y., Wang, Z. (2020). Improved sparse autoencoder based artificial neural network approach for prediction of heart disease. *Informatics in Medicine Unlocked*, 18(100307), 2-5.

[13] Xiao, N., Zou, Y., Yin, Y., Ziu, P., Tang, R. (2020). DRNN: Deep Residual Neural Network for Heart Disease Prediction. In *Machine Learning and Computer Application Journal of Physics: Conference Series 1682 (2020) 012065, 2020 International Conference on* (pp.1-5), IOP Publishing.

[14] Choudhary G. and Singh, S.N. (2020). Prediction of Heart Disease using Machine Learning Algorithms. In *Smart Technologies in Computing, Electrical and Electronics (ICSTCEE 2020 International Conference on* (pp. 197-202). IEEE.

[15] Bhatia, M. and Motwani, D. (2020). Use of Ensemblers Learning for Prediction of Heart Disease. In *Trends in Electronics and Informatics (ICOEI)(48184) 2020. 4th International Conference on*, (pp. 1016-1023), IEEE.

[16] Lakshmanarao, A., Srisaila, A. and Kiran, T. S. R. (2021). Heart Disease Prediction using Feature Selection and Ensemble Learning Techniques, In *Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), 2021.Third International Conference on* (pp. 994-998). IEEE. doi: 10.1109/ICICV50876.2021.9388482.

[17] Yuan, K., Yang, L., Huang, Y., Li, Z. (2020). Heart Disease Prediction Algorithm Based on Ensemble Learning," In *Dependable Systems and Their Applications (DSA), 2020. 7th International Conference on* (pp. 293-298), IEEE.

[18] Link for Framingham dataset - https://www.kaggle.com/datasets/aasheesh200/framingham-heart-study-dataset

[19] Atallah, R. and Al-Mousa, A. (2019). Heart disease detection using machine learning majority voting ensemble method. In *New Trends in Computing Sciences (ICTCS), Proceedings of the 2019 2nd International Conference on* (pp. 1–6), IEEE, October.

[20] Raza, K. (2019). Improving the prediction accuracy of heart disease with ensemble learning and majority voting rule. *UHealthcare Monitoring Systems*, (pp. 179–196), Jamia Millia Islamia, New Delhi.
[21] Kannan, R. and Vasanthi, V. (2019). Machine learning algorithms with roc curve for predicting and diagnosing the heart disease. *Soft Computing and Medical Bioinformatics*, 63– 72.
[22] Pillai, N. S. R., Bee, K. K. and Kiruthika, J. (2019). Prediction of heart disease using rnn algorithm. *International Research Journal of Engineering and Technology*, 6(3), 4452-4458.
[23] Bhat, R. Chawande, S. and Chadda, S. (2019). Prediction of test for heart disease diagnosis using artificial neural network. *Indian Journal of Applied Research*, 9, 48-50.
[24] Repaka, A. N., Ravikanti, S. D. and Franklin, R. G. (2019). Design And Implementing Heart Disease Prediction Using Naive Bayesian. In *Trends in Electronics and Informatics (ICOEI), 2019 3rd International Conference on* (pp. 292-297), IEEE. doi: 10.1109/ICOEI.2019.8862604.
[25] C. Ricciardi, A. S. Valente, K. Edmund et al. (2020). Linear discriminant analysis and principal component analysis to predict coronary artery disease, *Health Informatics Journal*, 26(3), 2181–2192.

Notes on contributors

Priyanka Gupta is currently working as an Assistant Professor in Computer Science & Engineering Department in SRM Institute of Science & Technology, Delhi-NCR Campus, Ghaziabad, India. Her research interests include Artificial Intelligence, Machine Learning and Deep Learning.

Dr. D.D. Seth, Senior Member IEEE is currently working as Professor in Computer Science and Engineering in SRM Institute of Science and Technology, Delhi-NCR Campus. His field of research includes Cognitive Radio Network, Software defined Network, Applications of AI/ML in wireless Networks.