ENHANCING HEART DISEASE PREDICTION THROUGH ENSEMBLE LEARNING AND FEATURE SELECTION

Shital Patil[1,2] & Surendra Bhosale[1]

[1]Department of Electrical Engineering, Veermata Jijabai Technological Institute, Mumbai, India
Ramrao Adik Institute of Technology Navi Mumbai, India
[2]Department of Electrical Engineering, Veermata Jijabai Technological Institute Mumbai, India

Abstract:
Machine learning techniques are being used extensively in the healthcare field to assist in the early detection and diagnosis of diseases. Prediction of cardiac disease from clinical data is one of the most significant applications of machine learning in healthcare. To develop such a predictive model, data collection, pre-processing, and transformation methods are used to train the model. Feature selection methods such as filter and wrapper are also used to enhance the predictive performance of the model. Classification techniques such as Decision Tree, Logistic Regression, Random Forest, and Ada Boost are used to evaluate the performance of the model. Performance metrics such as accuracy, F1-score, precision, sensitivity, and specificity are used to evaluate the effectiveness of the predictive model. Improvements in these performance metrics indicate that the predictive model is performing better and can help physicians make more informed decisions regarding patient's health. This study makes an effort to improve the likelihood of heart disease prediction using ensemble learning approaches. It demonstrates the effectiveness of using machine learning and artificial intelligence in predicting heart disease and highlights the importance of feature selection methods in achieving accurate results.

Keywords: Cardio Vascular Diseases, Feature selection, Decision Tree, Logistic Regression, Random Forest, Ada Boost.
DOI: 10.24297/j.cims.2023.4.24

1. Introduction

Cardio Vascular Diseases are a significant global health issue and a leading cause of death worldwide [1]. More than 70% of mortality rates around the world are caused by cardiovascular disease (CVDs), making it a serious global health concern. CVDs are a complex and multifactorial disease and several risk factors contribute to their development including unhealthy diets, physical inactivity, tobacco use and alcohol consumption. Early detection and prevention of

CVDs are essential for reducing mortality and morbidity associated with these diseases. Global Burden of Disease research from 2017 indicates that CVDs are accounting for approximately 43% of all fatalities [1,2]. In high-income countries, common risk factors for heart disease include poor diet, smoking, excessive sugar consumption, and obesity [3,4]. However, the prevalence of chronic illnesses is increasing in countries with low and medium incomes [5]. The economic burden of cardiovascular diseases is significant and it was projected to reach around [6] 3.7 trillion USD between 2010 and 2015. This highlights the need for effective prevention, diagnosis and treatment of CVDs globally [7]. By leveraging machine learning techniques and predictive models, healthcare professionals can identify patients at high risk of developing heart disease and intervene early to prevent the onset or progression of the disease. By analysing clinical data, machine learning algorithms can detect patterns and identify patients at high risk of developing heart disease. This can help healthcare professionals intervene early and implement preventative measures that can reduce the physical and monetary costs associated with heart disease. Moreover, machine learning algorithms can also help identify cost-effective diagnostic methods for heart disease [8]. By analysing data from various diagnostic tests, machine learning algorithms can identify which tests are most effective and provide accurate results. This can help reduce the cost burden on both individuals and institutions while improving the accuracy and efficiency of the diagnostic process [9]. This can help save lives and reduce the physical and monetary costs associated with this global health issue. Using data mining techniques, the medical industry generates a considerable quantity of data every day, and discover hidden patterns that can be used in clinical diagnostics [10]. When forecasting cardiac disease, a number of indicators need to be taken into account including diabetes, hypertension, excessive cholesterol and an irregular pulse rate [11,12]. The proposed predictive models using various methods, such as XGBoost [13,14], multilayer perceptron, decision tree classifier, random forest, and so forth. Ensemble learning is a machine learning approach that uses several classifiers to improve the performance of our system. Bagging, boosting, and stacking procedures are part of the ensemble learning framework. Bagging and boosting are produced using the same type of classifiers. However, stacking is generated with a separate sort of learner [15].

The remaining paper is organised in the following sections. Section 2 of the paper, provides a comprehensive review of previous research on cardiovascular disease prediction. It covers existing approaches and examines available techniques. Section 3 includes proposed methodology. Section 4 elaborate about ensemble methods. Sections 5 and 6 deals with experiments, discussion, and conclusion.

2. Related Work

In the suggested study, ensemble learning technique is applied to improve the prediction accuracy of coronary heart disease risk and separate classifiers (Support Vector Machine, Decision Trees, K Nearest Neighbours, Random Forest, and Gradient Boosting) are used. A variety of ensemble procedures, including majority voting, stacking, and bagging is incorporated and effectiveness of these strategies has been evaluated. Furthermore, this study

uses feature selection and hyperparameter tweaking strategies to further enhance the classifiers' performance. A comparative study in [2] that used the ensemble technique to predict coronary heart disease on four different datasets. Switzerland University Hospital (SUH), Long Beach Medical Centre (LBMC), Hungarian Institute of Cardiology (HIC) and Cleveland Clinic Foundation (CCF). Comparative examination of several ensemble approaches such as bagging, boosting (AdaBoost) and random forest is done [16]. Particle swarm optimisation (PSO) was utilised to choose features and the best results were obtained via the bagging tree. Hybrid model for predicting heart disease that combines random forest and the linear model. The author used hybrid model to conduct a comparative examination of several algorithms. When compared to other individual classifiers, the suggested model had the greatest accuracy. Various ensemble approaches such bagging, boosting, stacking and majority voting to improve the prediction accuracy of weak classifiers suggested a hybrid model development in [17]. The suggested hybrid model outperforms the sequential GA-based hybrid model in prediction accuracy by roughly 6%. discusses this in [18]. The use of an ECG signal-based convolutional neural network (CNN) for the identification of heart disease achieved an accuracy of 89.8% which is higher than the accuracy achieved by other traditional machine learning algorithms. In [19], The proposed a machine learning-based model for predicting the occurrence of heart disease using electronic health record (EHR) data. They used the XGBoost algorithm and achieved an accuracy of 85.13%. For the prediction of cardiac illness, [20] proposed a hybrid model that combines deep learning and conventional machine learning approaches.

Long short-term memory (LSTM) and neural network achieved an accuracy of 88.10%. Overall, these studies show that machine learning algorithms can be used effectively for the prediction of heart disease and ensemble methods can further improve the accuracy of these models.

Table 1.: Related research on the prediction of cardiac disease using massive datasets [12]

Authors	Novel Approach	Best Accuracy	Dataset
Shorewall, 2021 [5]	Stacking of KNN, random forest, and SVM outputs with logistic regression as the metaclassifier	75.1% (stacked model)	Kaggle cardiovascular disease dataset (70,000 patients, 12 attributes)
Maiga et al., 2019 [7]	-Random forest -Naive Bayes -Logistic regression -KNN	70%	Kaggle cardiovascular disease dataset (70,000 patients, 12 attributes)
Waigi at el., 2020 [12]	Decision tree	72.77% (decision tree)	Kaggle cardiovascular disease dataset (70,000 patients, 12 attributes)
Our and ElSeddawy, 2021 [21]	Repeated random with random forest	89.01%(random forest classifier)	UCI cardiovascular dataset (303 patients, 14 attributes)
Khan and Mondal, 2020 [22]	Holdout cross-validation with the neural network for Kaggle dataset	71.82% (neural networks)	Kaggle cardiovascular disease dataset (70,000 patients, 12 attributes)
	Cross-validation method with logistic regression (solver: lbfgs) where k = 30	72.72%	Kaggle cardiovascular disease dataset 1 (462 patients, 12 attributes)
	Cross-validation method with linear SVM where k = 10	72.22%	Kaggle cardiovascular disease dataset (70,000 patients, 12 attributes)

It is indeed true that a limited dataset can lead to a high risk of overfitting, as the model may learn to memorize the data rather than generalize from it. In contrast, using a larger dataset can help to reduce this risk by providing a more diverse and representative sample of the population being studied. This can lead to models that are more robust and generalizable. Adoption of a dataset of 70,000 participants and 11 characteristics for cardiovascular disease provides a considerable benefit, as it provides a substantial amount of data for the model to learn from. This dataset size is relatively large compared to many other studies in the field, which can help to reduce the risk of overfitting and improve the accuracy of the model. Table 1 provides a comprehensive assessment study on the prediction of cardiovascular disease conducted on huge datasets, emphasising the importance of employing a large dataset. This comparison can be helpful in understanding the strengths and weaknesses of different approaches and identifying areas where further research may be needed. Overall, the use of a large dataset can be a significant advantage in developing accurate and generalizable models for predicting cardiovascular disease. However, it is also important to ensure that the data is of high quality and properly curated to avoid biases or confounding factors that may affect the accuracy of the model. The goal of this research is to use computerised models to forecast the likelihood of cardiac disease.

3. Proposed Methodology

The proposed method for completing this research begins with the download of an open-source UCI data collection [3]. Once the dataset has been verified, preparation and data discretization are carried out using various techniques such as binning, data cleaning, data reduction, data transformation and select attributes. Following are the application of all of these strategies to the obtained dataset, the main technique of feature selection is used. Later, the data is subjected to the following algorithms, Nave Bayes, SVM, Random Forest, Decision Tree, and Logistic Regression. Important findings and conclusions discussed after using algorithms and various approaches. Figure 1 depicts the flow of various strategies.

a) **Data Preprocessing:** Pre-processing of data is a critical step in developing accurate and reliable predictive models. Cleaning the data, handling null missing values and removing inconsistencies can help to ensure that the data is of high quality and suitable for analysis.

b) **Feature Selection:** Feature selection is a crucial step in machine learning that involves identifying and selecting the most relevant and informative features for a given task. The objective of feature selection is to identify the most important features that are most strongly associated with the outcome of interest, while eliminating less important or irrelevant features that may add noise or complexity to the model.

c) **Data Balancing**: Now, the dataset was examined to see if it was balanced. It was found that there were significantly more negative cases than positive cases. The dataset was therefore severely unbalanced and needs to be balanced in order to avoid problems during model

training. The dataset was balanced using the Random Over Sampling method in this case. To make the minority class equal to the majority class, more instances of that group were created.

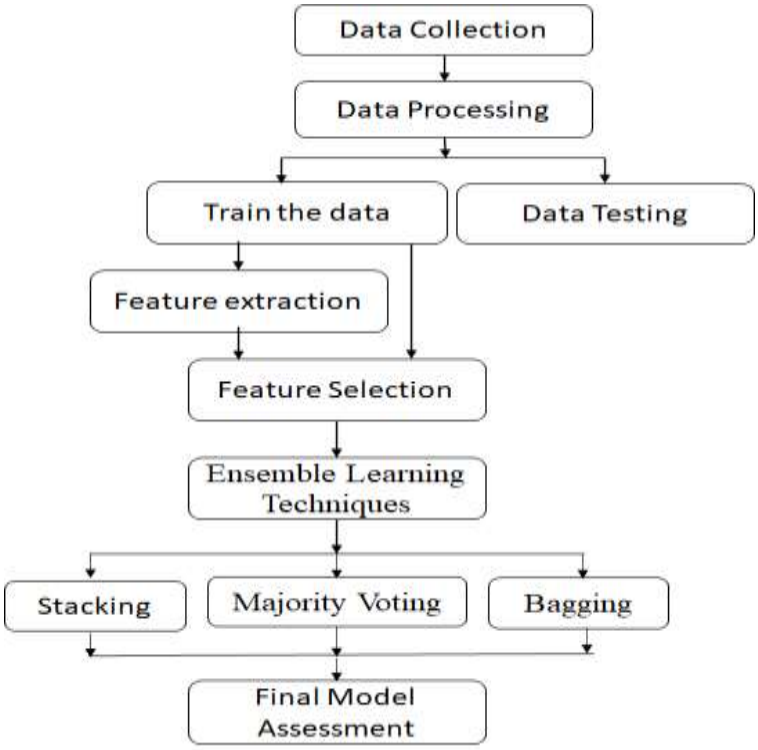

Fig. 1: Proposed system methodology

d) **Chi-square test-based feature selection:** A popular method for picking features in a dataset is chi-square test-based feature selection. The chi-square test is a statistical test used to determine the relationship between two category variables.

4. Ensemble Methods

a) **Stacking**: It is an ensemble method that combines multiple classifiers by using a meta-classifier to make the final prediction. Stacking involves several layers with each layer using the output of the previous layer as input.

b) **Majority Voting**: Majority voting is a simple ensemble method used for combining the predictions of multiple classifiers to make a final decision. In majority voting, each individual classifier in the ensemble independently makes a prediction.

c) **Bagging**: It is a popular ensemble method used in machine learning for improving the accuracy and stability of a model. It involves training multiple individual models on different subsets of the training data and then aggregating their predictions to make the final prediction.

5. Experiments, Results & Discussion

I. Classifier performance with ensemble approaches:

UCI repository dataset was used to evaluate and compare the performance of several categorization methods. Some classifiers have demonstrated strong performance while others have displayed weak performance as shown in Table 2.

Table 2: Classifier performance metrics without ensemble methods and feature selection

Methods	Precision	Recall	F1-Score	Accuracy
K Nearest Neighbor	0.816	0.974	0.911	91.65
Support Vector Machine	0.638	0.692	0.651	67.12
Decision Tree	0.817	0.962	0.913	91.14
Random Forest	0.843	0.955	0.909	91.28
Gradient Boosting	0.709	0.716	0.721	72.13
Naïve Bayes	0.657	0.589	0.555	61.38

Ensemble approaches were utilised to boost the algorithms' performance. Support vector machine, random forest, decision tree, k closest neighbours, and gradient boosting were the methods employed. The performance indicators are displayed as indicated Table 3 and Fig. 2

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

Methods	Precision	Recall	F1-Score	Accuracy
Stacking	0.925	0.981	0.954	95.211
Majority Voting	0.843	0.972	0.913	90.952
Bagging	0.865	0.967	0.912	91.202

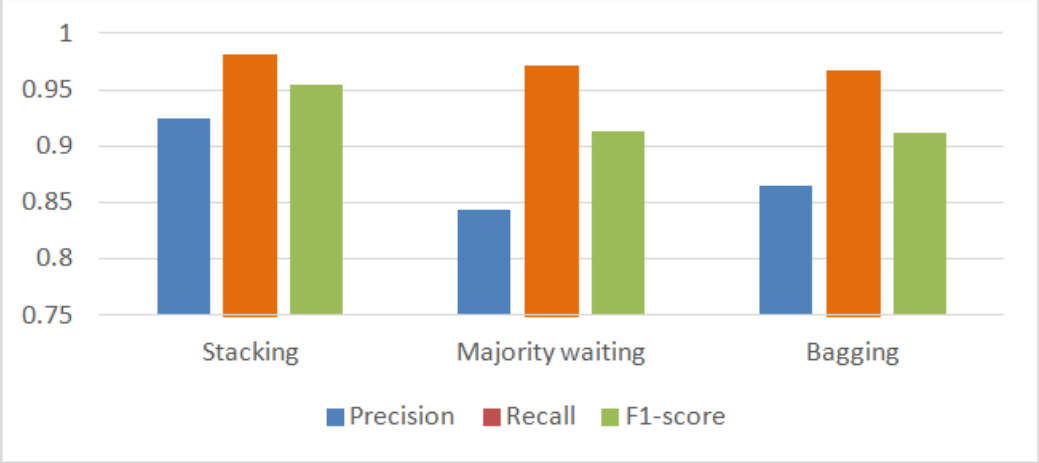

Fig. 2: A Comparison of Ensemble Classifier Prediction Accuracies Without Feature Selection

Enhancement of performance through feature selection

The use of feature selection and hyper parameter tuning approaches substantially enhanced the performance of the classifiers. The features are chosen in accordance with their value. The Chi Square Test technique was used to determine the value of various traits, and the attributes with the greatest influence on heart disease prediction were taken into account. The feature set was built, and the performance measures were assessed. The feature selection approach and hyperparameter adjustment are used to increase the performance of ensemble techniques (stacking, majority voting, and bagging) and same is indicated in Table 4 and Fig. 3.

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

Methods	Precision	Recall	F1-Score	Accuracy
Stacking	0.972	0.963	0.979	98.299
Majority Voting	0.968	0.971	0.966	98.299
Bagging	0.868	0.978	0.913	91.5

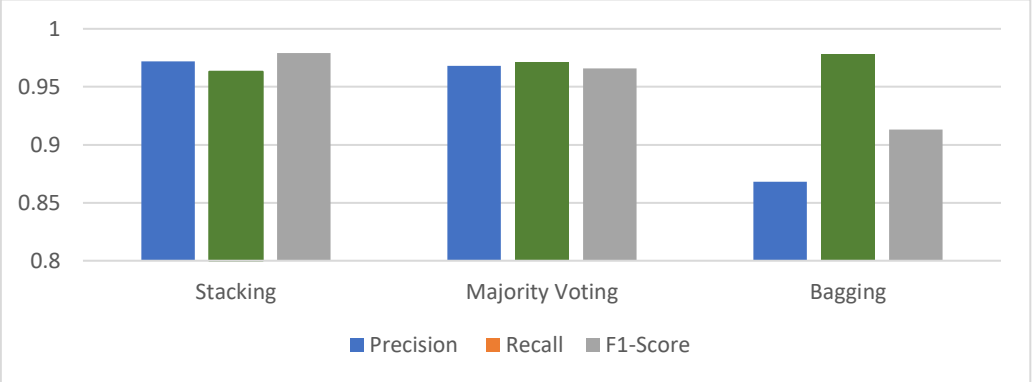

Fig. 3: A Comparison of Ensemble Classifier Prediction Accuracies with Feature Selection

When compared to individual classifiers, ensemble approaches predicted the risk of heart disease with greater accuracy. The accuracy of the model was further boosted by employing the techniques of feature selection and hyperparameter adjustment. Out of the three ensemble classifiers, stacking, majority voting, and bagging, Majority voting improved prediction accuracy most with feature selection. Using the feature selection approach, the accuracy of stacking was increased by 3.24%, while the accuracy of bagging was increased by 0.33%.

6. Conclusion

The leading cause of mortality is heart or cardiovascular disease. The leading cause of mortality is heart or cardiovascular disease. The model was created with the intention of increasing the risk of coronary heart disease's predictability, robustness, and accuracy. In this study, an early, accurate, and reliable diagnosis of coronary heart disease was achieved using an ensemble learning technique. Several ensemble procedures, including majority voting, stacking, and bagging, were used to enhance the performance of the classifiers for the prediction of heart disease. It has been shown that using feature selection techniques improved the classifiers'

performance even further. Three Ensemble classifiers are stacking, majority voting, and bagging. Forecast accuracy growth was best when majority voting was used, with a 98.38% accuracy rate.

References

1. Estes, C.; Anstee, Q.M.; Arias-Loste, M.T.; Bantel, H.; Bellentani, S.; Caballeria, J.; Colombo, M.; Craxi, A.; Crespo, J.; Day, C.P.et al. Modeling NAFLD disease burden in China, France, Germany, Italy, Japan, Spain, United Kingdom, and United States for the period 2016–2030. J. Hepatol. 2018, 69, 896–904
2. Drozdz, K.; Nabrdalik, K.; Kwiendacz, H.; Hendel, M.; Olejarz, A.; Tomasik, A.; Bartman, W.; Nalepa, J.; Gumprecht, J.; Lip, G.Y.H. Risk factors for cardiovascular disease in patients with metabolic-associated fatty liver disease: A machine learning approach.Cardiovasc. Diabetol. 2022, 21, 240.
3. UCI Machine Learning Repository, "Heart disease data set," 2020, http://archive.ics.uci.edu/ml/datasets/heart+disease
4. Kanksha, B. Aman, P. Sagar, M. Rahul, and K. Aditya, "An intelligent unsupervised technique for fraud detection in health care systems," Intelligent Decision Technologies, vol. 15, no. 1, pp. 127–139, 2021.
5. Shorewala, V. Early detection of coronary heart disease using ensemble techniques. Inform. Med. Unlocked 2021, 26, 100655.
6. K. Divya, A. Sirohi, S. Pande, and R. Malik, "An IoMT assisted heart disease diagnostic system using machine learning techniques," in Cognitive Internet of Medical 4ings for Smart Healthcare, vol. 311, pp. 145–161, Springer, Cham, Switzerland, 2021.
7. Maiga, J.; Hungilo, G.G.; Pranowo. Comparison of Machine Learning Models in Prediction of Cardiovascular Disease Using Health Record Data. In Proceedings of the 2019 International Conference on Informatics, Multimedia, Cyber Information System (ICIMCIS), Jakarta, Indonesia, 24–25 October 2019; pp. 45–48.
8. Gavhane A, Kokkula G, Pandya I, Devadkar PK. –Prediction of Heart Disease Using Machine Learning, in Proceedings of the 2nd International Conference on Electronics, Communication and Aerospace Technology, ICECA 2018, 2018; pp. 1275–1278.
9. Waigi, R.; Choudhary, S.; Fulzele, P.; Mishra, G. Predicting the risk of heart disease using advanced machine learning approach. Clin. Med. 2020, 7, 1638–1645.
10. S. Tomov and S. Tomov, ''On deep neural networks for detecting heart disease,'' Aug. 2018, arXiv:1808.07168.
11. A. K. Dwivedi, ''Performance evaluation of different machine learning techniques for prediction of heart disease,'' Neural Computer. Appl., vol. 29, no. 10, pp. 685–693, 2018.
12. Bhatt, C.M.; Patel, P.; Ghetia, T.; Mazzeo, P.L. Effective Heart Disease Prediction Using Machine Learning Techniques. Algorithms 2023, 16, 88
13. S. Amin, Y. K. Chiam, K. D. Varathan, ''Identification of significant features and data mining techniques in predicting heart disease,'' Telematics Inform., vol. 36, pp. 82–93, Mar. 2019.
14. Y. Chen, T. Zhang, Z. Luo, and K. Sun, ''A novel rolling bearing fault diagnosis and severity analysis method,'' Appl. Sciences, vol. 9, no. 11, p. 2356, Jun. 2019.

15. Narin, A.; Isler, Y.; Ozer, M. Early prediction of Paroxysmal Atrial Fibrillation using frequency domain measures of heart rate variability. In Proceedings of the 2016 Medical Technologies National Congress (TIPTEKNO), Antalya, Turkey, 27–29 October 2016.
16. R. Chen, N. Sun, X. Chen, M. Yang, and Q. Wu, "Supervised feature selection with a stratified feature weighting method," IEEE Access, vol. 6, pp. 15087–15098, 2018.
17. M.-S. Yang and Y. Nataliani, "A feature-reduction fuzzy clustering algorithm based on feature-weighted entropy," IEEE Transactions on Fuzzy Systems, vol. 26, no. 2, pp. 817– 835, 2018
18. A. K. Garate-Escamila, A. Hajjam El Hassani, and E. Andres, ` "Classification models for heart disease prediction using feature selection and PCA," Informatics in Medicine Unlocked, vol. 19, Article ID 100330, 2020.
19. P. Ramprakash, R. Sarumathi, R. Mowriya, and S. Nithyavishnupriya, "heart disease prediction using deep neural network," in Proceedings of the 2020 pp. 666–670, IEEE, Coimbatore, India, February 2020.
20. Li H, et al. (2018). Ensemble learning for overall power conversion efficiency of the all-organic dyesensitized solar cells. IEEE Access 2018; 6:34118–26.
21. Ouf, S.; ElSeddawy, A.I.B. A proposed paradigm for intelligent heart disease prediction system using data mining techniques. J. Southwest Jiaotong Univ. 2021, 56, 220–240.
22. Khan, I.H.; Mondal, M.R.H. Data-Driven Diagnosis of Heart Disease. Int. J. Computer. Appl. 2020, 176, 46–54.