International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Enhancing Heart Disease Risk Prediction Accuracy through Ensemble Classification Techniques

Rupali Atul Mahajan[1], Dr. Balasaheb Balkhande[2], Dr. Kirti Wanjale[3], Dr. Abhijit Chitre[4], Tushar Ankush Jadhav[5], Dr. Sheela Naren Hundekari[6]

Submitted: 25/05/2023 **Revised**: 07/07/2023 **Accepted**: 26/07/2023

Abstract- A crucial part of handling different data science problems is machine learning, a branch of artificial intelligence. One of its common uses is making predictions based on past data. Classification is a powerful machine learning technique that is frequently used to produce accurate predictions. On the other hand, some categorization algorithms may have a maximum level of accuracy. In this study, the ensemble classification technique which combines multiple classifiers is investigated as a means of enhancing the precision of less accurate algorithms. Accurate diagnosis and classification of cardiovascular diseases are essential for selecting the most effective treatment and reducing mortality. Machine learning has developed into a crucial tool in the medical sector by leveraging data patterns for improved diagnosis. This project aims to reduce misdiagnosis and improve patient outcomes by developing a predictive model for cardiovascular illnesses using machine learning techniques. In this study, machine learning is used to address the problem of accurate classification of cardiovascular diseases. The developed methodology can help diagnosticians make informed choices that lead to quick and targeted therapies. This study shows how machine learning has a significant impact on medicine and has the potential to reduce cardiovascular disease-related mortality. On a dataset of heart disease patients, the work focuses on employing ensemble classification to increase prediction accuracy. The goal is to demonstrate the algorithm's value in predicting diseases early using medical data and to raise the accuracy of weaker classifiers. Experimental comparisons were done to determine the impact of the ensemble technique on the accuracy of heart disease prediction.

Keywords: *Machine Learning, Ensemble method, heart disease, Classification Techniques, prediction model*

I. Introduction

The CHDD is used by researchers to investigate and examine risk factors for heart disease, create prognostic models, and assess the efficacy of various diagnostic procedures and therapeutic approaches. The dataset has considerably advanced our knowledge of and ability to treat cardiac disease. 17 million people have tragically passed away as a result of the high cost and unfeasibility [5]. Additionally, cardiovascular disease has major financial repercussions for businesses because it accounts for 25 to 30 percent of annual medical costs for employees [8]. Therefore, to lessen the financial and physical toll on people and organizations, early identification of heart disease is essential.

The heart disease and stroke will continue to be the two main causes of cardiovascular disease fatalities, with the overall number of deaths from these illnesses expected to reach 24.6 million by 2035 [9]. These figures show how critical it is to take quick action to identify and treat heart disease in order to lessen its catastrophic effects on the health of the entire world. More than 70% of all fatalities are caused by cardiovascular disease (CVDs), including heart disease, which continues to be a primary cause of sickness and death worldwide. Obesity, tobacco use, excessive sugar consumption, and unhealthy eating habits are all prominent risk factors for heart disease, which are more common in high-income countries.

However, a concerning rise in chronic diseases is also being seen in low- and middle-income nations. According to estimates, from 2010 to 2015, CVDs had an economic impact of almost USD 3.7 trillion on a global level. This significant financial burden is a reflection of the price of healthcare services, the cost of treatment, lost productivity, and other relevant expenditures related to managing CVDs. It is essential to address the prevention, early identification, and management of heart disease in order to improve both

[1]*Associate Professor, Department of Computer Engineering Department, Vishwakarma Institute of Information Technology Pune, Maharashtra, India*
[2]*Associate Professor, Vasantdada Patil Pratishthan College of Engineering and Visual Arts, Mumbai, Maharashtra, India*
[3]*Associate Professor, Department of Computer Engineering, Vishwakarma Institute of information technology Pune, Maharashtra, India*
[4]*Associate professor, Department of Electronics and Telecommunication Engineering, Vishwakarma Institute of Information Technology, Pune, Maharashtra, India*
[5]*Department of Mechanical Engineering, Sinhgad College of Engineering, Pune, Maharashtra, India*
[6]*MIT ADT University, Loni kalbhor, Pune, Maharashtra, India*
rupali.mahajan@viit.ac.in1, balkhandeakshay@gmail.com2, kirti.wanjale@viit.ac.in3, abhijit.chitre@viit.ac.in4, tajadhav.scoe@sinhgad.edu5, sheela.hundekari@mituniversity.edu.in6

public health and the economic burden on societies around the world.

By enabling the examination of enormous amounts of data and revealing hidden patterns that can help with clinical diagnosis, data mining techniques have transformed the medical industry [10]. Numerous studies undertaken over the past few decades have shown the importance of data mining in healthcare [11].

II. Review of Literature

Techniques like machine learning and data mining are extremely helpful for solving a variety of issues, including forecasting dependent variables based on independent factors. Due to its enormous and complicated data resources, which are impossible to handle manually, the healthcare industry, in particular, greatly benefits from the deployment of these methodologies. Data mining and machine learning provide potent methods to extract significant insights from the wealth of healthcare data, including electronic health records, medical imaging, genetics, and patient demographics. With the aid of these tools, healthcare professionals are able to rapidly evaluate massive datasets, spot trends, and assign varied medical problems and outcomes to categories or forecasts. Healthcare professionals may increase diagnostic precision, forecast disease development, identify risk factors, optimize treatment regimens, and even improve patient outcomes by utilizing machine learning algorithms.

With an outstanding accuracy of 92.1% in heart disease risk prediction, Support Vector Machines (SVM) emerged as the best predictor among these methods. With an accuracy of 91%, neural networks came in second place, while decision trees came in third with an accuracy of 89.6% [23]. These machine learning algorithms examine pertinent patient data, including medical history, demographic data, and results of diagnostic tests, to find trends and create predictive models for the early identification of heart disease risks. Healthcare practitioners can improve their capacity to recognize people who are at early risk of acquiring heart disease by utilizing these cutting-edge procedures. This makes it possible for preemptive interventions, individualized treatment programs, and lifestyle adjustments that may lower the morbidity and mortality linked to heart disease.

Shah et al.'s work [18] aimed to create a model utilizing machine learning methods to predict cardiovascular disease. They used the Cleveland heart disease dataset from the UCI machine learning repository, which contained 303 occurrences and 17 characteristics. Numerous supervised classification techniques were used by the researchers, including naive Bayes, decision trees, random forests, and k-nearest neighbour (KNN). According to the study's conclusions, the KNN model had the greatest accuracy rate among the examined models, at 90.8%. This result demonstrates the potential utility of machine learning methods for cardiovascular disease prediction. The study also stresses how crucial it is to choose the right models and methods in order to forecast such diseases with the best accuracy.

The study's machine learning approaches performed effectively, correctly classifying 114 out of 144 (79.17%) low-risk patients and 40 out of 47 (85.11%) high-risk patients. The model's area under the curve (AUC) score, which was 0.87, demonstrated its effectiveness in identifying extensive CVD in MAFLD patients using straightforward patient criteria. These results show the effectiveness of machine learning techniques in identifying MAFLD patients at risk for cardiovascular disease. By using these methods, medical personnel may be able to better risk stratify patients with MAFLD, carry out prompt interventions, and manage their cardiovascular health.

Table 1: Related survey for prediction CVD disease from large dataset

Paper	Method Used	Maximum Accuracy	Experimental Dataset Used
[5]	Random Forest, Support Vector Machine, and Stacking of KNN	76.1%	Cardiovascular disease dataset (CVD) from kaggle
[7]	Naive Bayes, Random forest, KNN and Logistic regression	72%	CVD form kaggle
[12]	Decision tree	74.77%	CVD form kaggle
[21]	Repeated random with random forest	88.01%	CVD from UCI

[22]	Holdout cross-validation integration with neural Network	73.82%	CVD from kaggle
[22]	Cross-validation method with logistic regression (solver: lbfgs) where k = 30	73.72%	CVD1 from kaggle
[22]	Cross-validation method with linear SVM where k = 10	73.22%	CVD from kaggle

III. Dataset Description

Cleveland heart dataset (CHD) dataset from the UCI machine learning repository was used in the study. There are 303 occurrences in this collection, and there are a total of 14 attributes. Eight of these characteristics are categorized, and the remaining six are numerical. To guarantee a thorough representation of the factors connected with the study's goal, the dataset was carefully chosen. The description of dataset represent in table 2.

Table 2: Dataset feature description

Sr. No.	Feature Variable	Minimum Value	Maximum Value	Classes
1	Gender	-	-	2 categories Male and female
2	Cholesterol	1	3	3 categories Low, Normal, High
3	Glucose	1	3	3 categories Normal, Above Normal, High
4	Smoking	0	1	2 categories Yes, No
5	Alcohol Intake	0	1	2 categories Yes, No
6	Physical Activity	0	1	2 categories Yes, No
7	Presence or Absence of Cardiovascular Disease (Cardio)	0	1	2 categories Yes, No
8	Age	10,798	23,713	-
9	Height	55	250	-
10	Weight	10	200	-
11	Systolic Blood Pressure	-150	16,020	-
12	Diastolic Blood Pressure	-70	11,000	-

Patients between the ages of 29 and 79 are included in the dataset utilized in this investigation. The value 1 for male patients and 0 for female patients serves as a gender indicator. There are four distinct types of chest discomfort that are regarded as signs of heart disease.

The attribute "trestbps" denotes the measurement of resting blood pressure, and the attribute "chol" denotes the level of cholesterol. A value of 1 is given if the fasting blood sugar is under 120 mg/dl and a value of 0 if it is higher. "Fbs" stands for fasting blood sugar level. The electrocardiogram result for resting is shown as "Restecg," the maximal heart rate is shown as "Thalach," and exercise-induced angina is shown as "Exang," with pain levels of 1 or 0 denoting the existence of pain. The terms "Oldpeak" and "slope" relate to the activity-induced ST depression and the slope of the peak exercise ST segment, respectively. The variable "ca" denotes the quantity of main vessels fluoroscopically colored, while the attribute "thal" denotes the number of minutes the exercise test lasted.

IV. Classification Techniques

The classification method of supervised learning is used to forecast outcomes using historical data. In this study, a novel method for identifying heart disease using classification algorithms is provided. Additionally, by using a group of classifiers, the classification accuracy is improved. The dataset is split into two subsets: a training set and a test set—to evaluate the suggested methodology. The training dataset is used to train the individual classifiers so they can recognize patterns and connections in the data. The test dataset is then used to evaluate the efficiency and performance of these classifiers.

1. Naive Bays:

A probability-based classification technique built on the ideas of the Bayes theorem is the Naive Bayes classifier, commonly referred to as the Bayesian classifier. It is regarded as a particular instance of the Bayesian network. The Naive Bayes classifier bases one of its main presumptions on the idea that all features are conditionally independent. Accordingly, modifications to one feature have no impact on modifications to other features' probabilities.

When classifying high-dimensional datasets, the Naive Bayes method excels. The classifier effectively manages datasets with a lot of attributes by utilizing the notion of conditional independence. Based on the individual probabilities of each feature given a specific class, the algorithm can treat each feature individually and generate predictions as a result.

A set of training data with appropriate class labels is called D. Each tuple in the collection has n attributes that are denoted by the expression X = A1, A2,..., An. The collection has m classes identified by the letters C1, C2,..., Cm. If and only if the posterior probability of class Ci, conditioned on X, is the highest among all the classes, the Nave Bayes classifier predicts that tuple X belongs to class Ci.

$$P\left(\frac{Ci}{X}\right) > P\left(\frac{Cj}{X}\right) \quad for\ 1 \le j \le m, i \ne j$$

By reducing the combined probability of X and Ci, denoted as P(Ci | X), by the likelihood of the tuple X, denoted as P(X), one can determine the probability of class Ci given tuple X, denoted as P(Ci | X), according to Bayes' theorem. This can be stated as follows:

$$P(Ci \mid X) = (P(X \mid Ci) * P(Ci)) / P(X)$$

The posterior probability P(Ci | X) must be maximized in order for us to conclude that tuple X belongs to the class Ci in the Naive Bayes classifier. The maximum posteriori hypothesis is applied to the class Ci for which P(Ci | X) is maximal. We can ascertain the most likely class to which tuple X belongs by contrasting the posterior probability of all classes.

$$P(Ci \mid X) = \prod_{k=1}^{m} (P(X\,k \mid Ci)$$

If attribute Ak is categorical, then P(X | C)k=i denotes the likelihood, given the dataset D, of finding the value xk for attribute Ak in tuples of class Ci. This probability is determined by dividing the total number of class Ci-belonging tuples in Ci (|Ci ∩ D|).by the number of tuples in D that have the value xk for attribute Ak.

$$P(X \mid Ci)P(Ci) > P(X \mid Cj)\,P(Cj) \qquad for \quad 1 \le j \le n \quad , \quad i \ne j$$

If and only if the probability P(Ci | X) is maximized over all classes, the classifier will predict that tuple X belongs to class Ci. In other words, the classifier assigns the class label Ci to tuple X if the posterior probability of class Ci, given tuple X, is the highest. To identify the most likely class given tuple X, the posterior probabilities of all classes are computed and compared.

2. Random Forest:

The term "random forest" denotes that the algorithm creates a structure like a forest made up of numerous individual trees. Given that it integrates the results of several algorithms, it is referred to as an ensemble algorithm.

Algorithm Random Forest:

Let D be a training set $D = \{(x1, y1), \ldots, (xn, y,)\}$

Let $h = h1(x), h2(x), \ldots,$ for an ensemble of weak classifier

If each h_k is a decision tree, the parameters of the tree are defined as

$\theta = (\theta_{k1,}\, \theta_{k2,}, \ldots\ldots, \theta_{kp,})$

Each decision tree k leads to a classifier $h_k(x) = h(X \mid \theta_k)$

Final Classification $f(x) =$ Majority of $h_k(x)$

Each tree in a random forest utilizes a random selection of features and is trained on a distinct subset of the training data. With less overfitting, the model performs better overall and is more robust thanks to this randomness. To arrive at a final categorization determination during prediction, the random forest integrates the forecasts of each individual tree.

The random forest approach can handle complicated datasets and capture subtle correlations between attributes by utilizing the diversity of numerous decision trees. It is renowned for its capacity to manage high-dimensional data, deal with missing values, and offer perceptions into the significance of features. Overall, the ensemble nature of the random forest method and its capacity to deliver accurate and dependable results make it a strong and well-liked option for classification problems.

3. Decision Tree:

Large datasets are managed using decision trees, structures that resemble trees. They are frequently represented as flowcharts, where the inner nodes stand in for the characteristics or attributes of the dataset and the outer branches reflect the results. Decision trees are admired for their effectiveness, dependability, and readability.

The value of the property at each node is compared to the relevant data in the record to determine the next step in the tree. The method follows the corresponding branch that leads to the following node in the tree based on the results of the comparison.

$$\text{Entropy(S)} = -\Sigma\left(\text{p(i)} * \log 2\left(\text{p(i)}\right)\right)$$

Entropy values vary from 0 to 1, with 0 denoting an entirely pure node (all examples come from the same class), and 1 denoting the largest amount of impurity (examples are evenly dispersed among several classes).

$$\text{Information Gain}\,(\text{S}, \text{A}) = \text{Entropy(S)} - \sum \text{v} \in \text{values(A)}\ |\text{Sv}|\ |\text{S}|\ \text{Entropy(Sv)}$$

Decision tree algorithms heavily rely on the notion of entropy. Entropy measures the chaos or impurity within a collection of training instances. The entropy changes as a decision tree node splits the training instances into more manageable chunks. Information gain is the measurement of this change in entropy. Gaining information enables the algorithm to select the most informative characteristics for data splitting and decision tree construction.

4. Multilayers Perceptron:

Multilayer perceptrons use perceptrons, which are synthetic neurons inspired by biological neurons. These synthetic neurons are in charge of mapping the network's inputs and processing weighted inputs. The multilayer perceptron lowers the network to two layers and streamlines the information flow by using an activation function.

The allocated weights are modified repeatedly as part of the perceptron's learning process. In order to maximize the network's performance in creating accurate classifications, the algorithm adjusts the weights throughout training. The perceptron fine-tunes its decision-making process by incrementally altering the weights, which improves its capacity to distinguish between various classes.

Algorithm for Multi-layered Perceptron

$Initialize\ weights\ and\ biases\ in\ N, where\ N\ is\ the\ Network$

$while\ condition\ is\ true\ \{$

$for\ each\ training\ tuple\ X\ in\ D\ \{$

$for\ each\ input\ layer\ unit\ j\ \{$

$E_j = K_i$

$for\ each\ hidden\ or\ output\ layer\ unit\ j\ \{$

$K_i = \sum_i W_{ij}\ E_j + \theta_j$

$E_j = \frac{1}{1 + e^{-I_j}}$

$for\ each\ unit\ j\ in\ the\ output\ layer$

$Err_j = E_j\ (1 - E_j)(T_j - E_j)$

$for\ each\ unit\ j\ in\ the\ hidden\ layers, from\ the\ last\ to\ the\ first\ hidden\ layer$

$Err_j = E_j\ (1 - E_j) \sum Err_j\ W_{jk}$

$for\ each\ weight\ wij\ in\ N\ \{$

$\Delta W_{ij} = (I)\ Err_j E_j$

$W_{ij} = W_{ij} + \Delta W_{ij}\ \}$

$for\ each\ bias\ \theta_j\ in\ N\ \{$

$\Delta\ \theta_j = (I)\ Err_j$

$\theta_j = \theta_j + \Delta\ \theta_j$

$\}\ \}$

The multilayer perceptron algorithm makes use of perceptrons, which are artificial neurons arranged in several layers, including hidden layers. These neurons process weighted inputs using activation functions, and the perceptron learns by incrementally changing the supplied weights to enhance classification performance.

5. Ensemble Method:

In order to improve classifier accuracy, ensemble approaches are used. To increase the efficacy of the less effective models, they combine weak learners and strong learners. In this study, an ensemble technique is used to improve the precision of different algorithms for heart disease prediction. Improved performance over using a single classifier alone is the main goal of merging numerous classifiers using ensemble techniques. The ensemble approach makes use of the advantages and mitigates the disadvantages of individual classifiers by combining the predictions from various models. This cooperative endeavor aims to improve the overall accuracy and dependability of the heart disease prediction process.

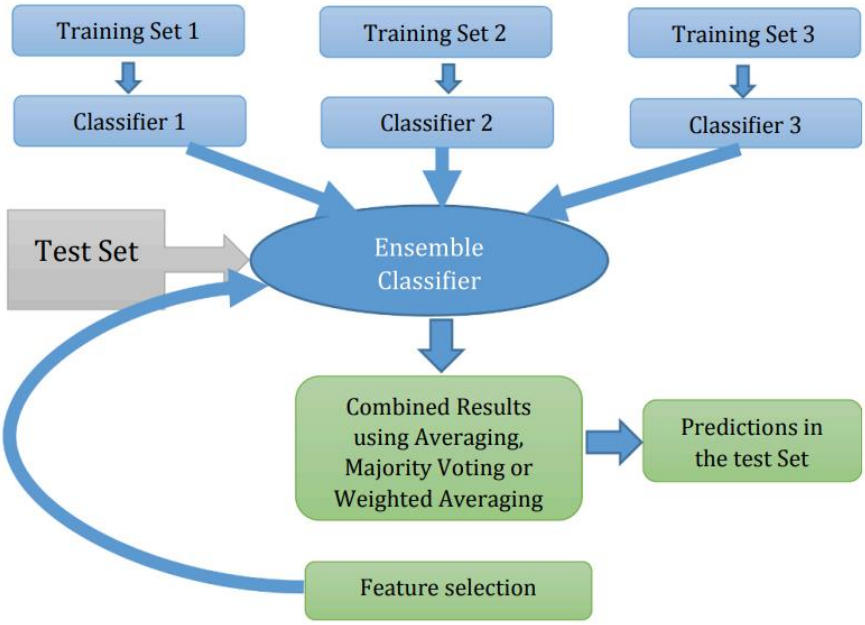

Fig 1: Flow of Ensemble method

The ensemble technique is a meta-classification method that makes use of the power of mixing various algorithms, each of which contributes its own special capacity for generating decisions. The ensemble technique seeks to increase predictive accuracy and produce results that are more reliable in the context of heart disease prediction by utilizing the variety and collective knowledge of numerous classifiers.

5.1 Boosting:

Boosting is an ensemble approach that boosts classifier performance. The first step in the method is to separate the original dataset into several subsets. Each subset is used to train the classifier, which results in a set of models with varying degrees of performance. The misclassified components of the previous model are chosen to construct new subgroups after training the initial model. The following models are then trained using these subsets, with a focus on improving the categorization of the cases that have previously been misclassified.

By integrating the various weak models with the aid of a cost function, the ensembling procedure in boosting tries to improve the performance of the weak models. Each model's performance is assessed by this cost function, and weights are assigned in accordance.

Boosting is an ensemble approach that trains numerous models by segmenting the dataset into smaller sections. New subsets are made using misclassified instances for later model training. The weak models are then combined into the ensemble using a cost function, effectively enhancing their overall performance and raising the classification process's accuracy.

5.2 Bagging:

Bagging is a method that includes picking random subsets of patterns from the training set, also known as bootstrap aggregation. Each pattern has the potential to be selected more than once for a specific subset when bagging because the selection is carried out with replacement.

Bagging divides the training set into numerous subsets by randomly selecting patterns from the original dataset, as opposed to feeding the full training set into a single classifier. The same pattern might show up more than once in a single subset when replacement is allowed, while certain patterns might not be included at all.

Because each subset is made up of a unique combination of patterns from the training set, this sampling procedure with replacement adds diversity to the subsets. A separate classifier is then trained using each subset, which may employ the same method or a different algorithm. Utilizing methods like majority voting or average, the predictions from each of these classifiers are combined to get the final forecast. Utilizing bagging with replacement increases the ensemble's stability and robustness by adding variations to the training subsets. Bagging decreases the influence of individual patterns and enhances the overall accuracy and reliability of the ensemble model by integrating the predictions of numerous classifiers trained on various subsets.

V. Result And Discussion

On the Cleveland dataset, a comparison of various categorization techniques has been done. These algorithms function differently, with some showing good accuracy and others performing poorly. In the present

study, ensemble techniques have been used to improve the performance of weaker classifiers.

In this study, ensemble techniques like bagging, boosting, voting, and stacking are used. The Naive Bayes, Random Forest, Decision Tree, and Multilayer Perceptron algorithms' predictions are combined into an ensemble through the Bagging process. Utilizing the Adaboost.M1 technique, boosting entails building ensembles utilizing Naive Bayes, Random Forest, Decision Tree, and multilayer perceptron classifiers.

Another ensemble technique used has been majority voting, in addition to bagging and boosting. The ultimate choice is determined by majority vote, which includes the predictions of various classifiers. Another ensemble method is stacking, which makes use of the Naive Bayes classifier as the meta classifier. By adding one, two, or three more classifiers on top of the Naive Bayes classifier, the results of stacking are acquired. The study uses these ensemble techniques in an effort to enhance the performance of many classifiers by combining their strengths and collective knowledge. The distinct tactics each ensemble method brings for integrating the predictions could result in increased robustness and accuracy in classification challenges.

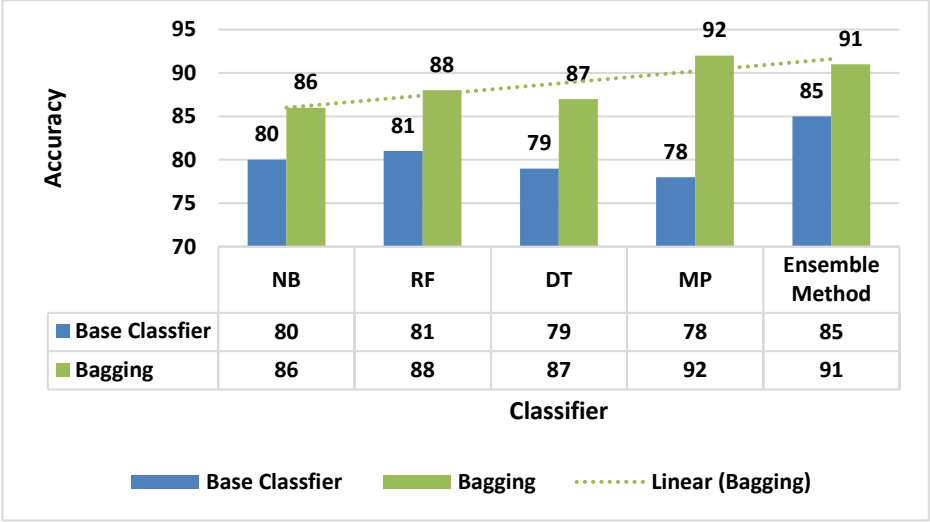

Fig 2: Accuracy of Base classifier with Bagging

The findings of the study demonstrate that when merged, subpar classifiers can perform better in ensembles. The dataset was classified using the Weka tool, which made dealing with inaccurate and missing data easier when cleaning and pre-processing the data. Several classifiers, including Naive Bayes, Random Forest, Decision Tree, and multilayer perceptron classifiers, were utilized for the classification problem. In comparison to Naive Bayes, Random Forest, and Baye Net, Multilayer Perceptron, and Decision Tree were found to be among these classifiers' inferior counterparts. Weak learners were treated to meta-classification algorithms since ensembling has been shown to be a successful tactic for improving classification accuracy.

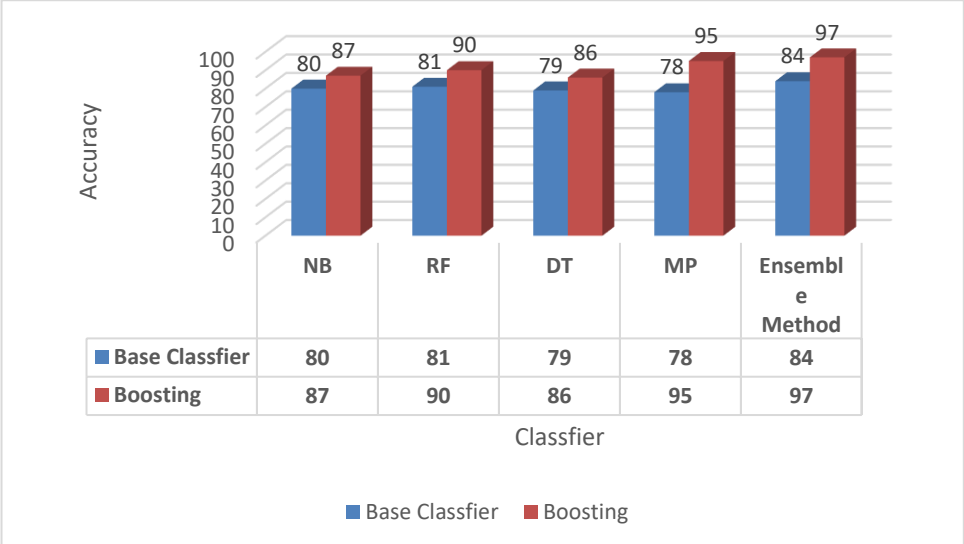

Fig 3: Accuracy of Base classifier with Boosting

The performance of the weak classifiers was improved using ensemble techniques: bagging, and boosting, as shown in figure 2 and figure 3. Each technique's outcomes underwent a thorough analysis. Ten-fold cross-validation was utilized to evaluate how effectively the classification models were performing. The dataset was divided into ten subgroups using this manner, and the classification process was carried out ten times. Each cycle's testing set consisted of nine subsets, and the last subset served as the training set. To arrive at the final results, the outputs from all ten iterations were averaged.

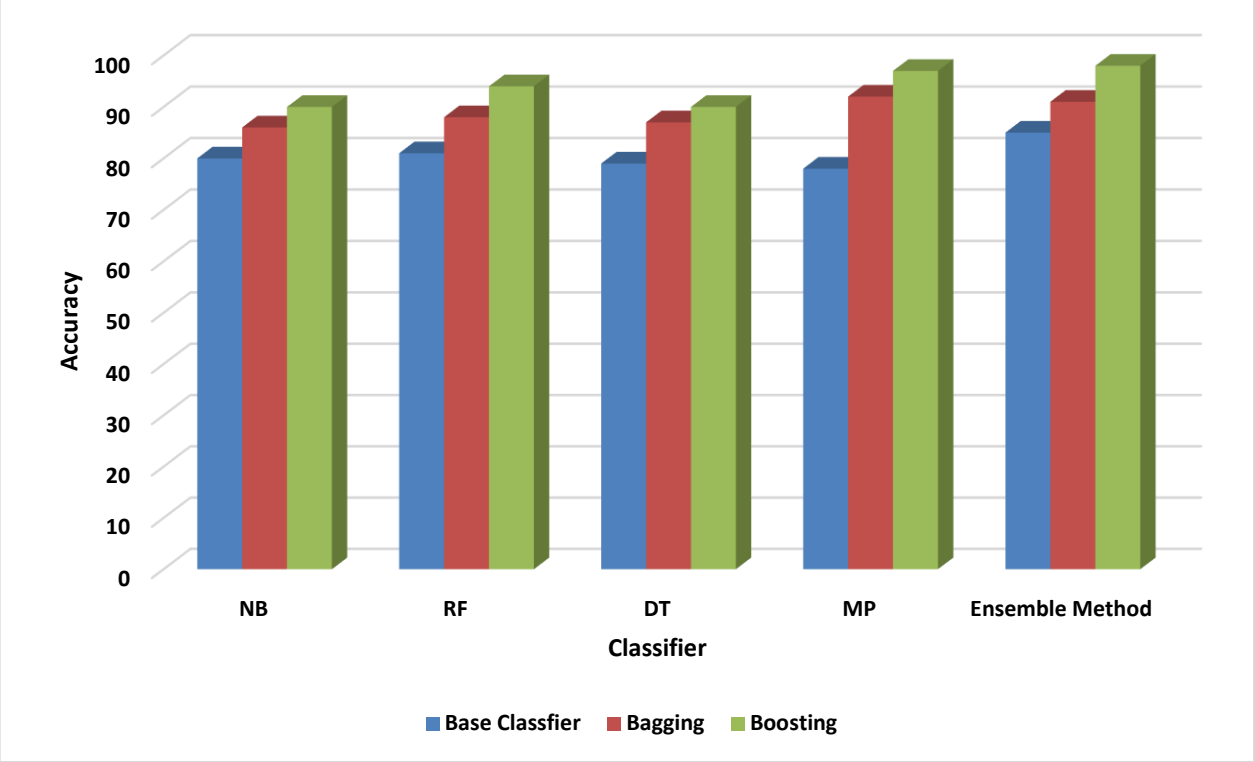

Fig 4: Comparison of Accuracy with Bagging and Boosting of Classifier

Several classification methods were used in the experiment along with feature selection to assess how well they predicted the target variable. The accuracy of the Random Forest method was 81.53%. The accuracy increased to 82.18% after feature selection (FS6) was applied. The accuracy of FS2, a different feature set, was 91.52%. The accuracy of the Multilayer Perceptron algorithm was initially 78.52%. However, accuracy greatly rose to 97.18% when feature selection (FS6) was used. Accuracy values of 97.18% and 98.85% were obtained with FS4 and FS3, respectively, thanks to similar advancements.

Table 3: Improvement in boosting accuracy achieved through feature selection

Method	Accuracy	With Feature Selection and Improvement in Accuracy	Feature Set
RF	81.33	81.81	FS-6
RF	81.35	91.25	FS-2
MP	78.25	97.81	FS-6
MP	78.25	97.81	FS-4
MP	78.525	98.85	FS-3
DT	79.16	94.28	FS-1
NB	80.61	95.94	FS-6
Ensemble Method	84.65	98.78	FS-1

The Decision Tree algorithm achieved an initial accuracy of 79.16%; feature selection (FS1) was then included, increasing accuracy to 94.82%. With FS6, the accuracy of Nave Bayes increased to 95.49% from 80.16%.

Finally, an ensemble technique that merged various classifiers was used. The accuracy of this ensemble approach was 84.56%. When features were chosen (FS1), accuracy increased to 98.87% shown in table 3. Overall, the findings show that feature selection considerably improved the classification systems' performance. The relevance of choosing pertinent features for obtaining improved accuracy in predictive modelling tasks is illustrated by the fact that different feature sets produced variable degrees of improvement.

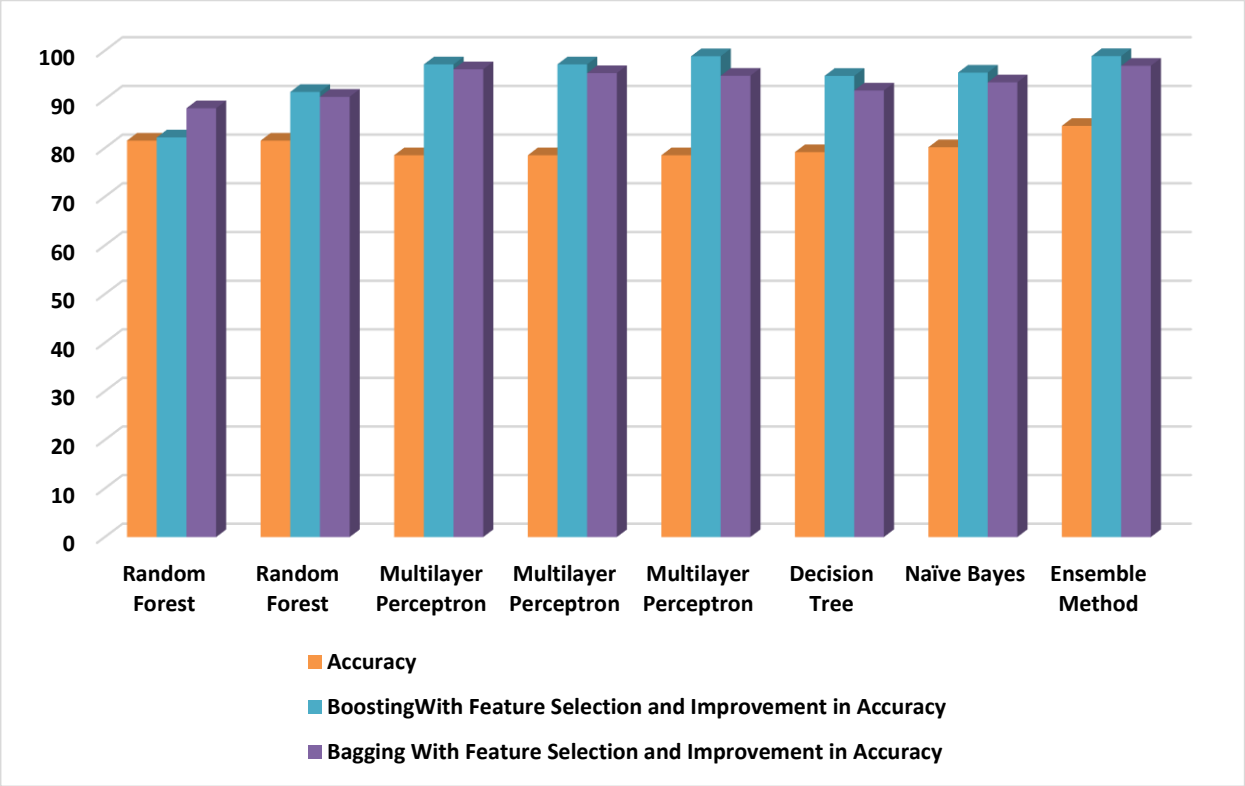

Fig 5: Increase in the Classifier Accuracy with boosting and bagging

Table 4: Improvement in bagging accuracy achieved through feature selection

Method	Accuracy	With Feature Selection and Improvement in Accuracy	Feature Set
RF	81.35	88.81	FS-6
RF	81.36	90.26	FS-2
MP	78.27	95.83	FS-6
MP	78.26	95.84	FS-4
MP	78.25	94.56	FS-3
DT	79.61	91.27	FS-1
NB	80.51	93.36	FS-6
Ensemble Method	84.54	96.88	FS-1

The accuracy of the Random Forest method was 81.53%. Applying feature selection with FS6 increased accuracy to 88.18%. Resulted in an increase in accuracy of 90.52%. The initial accuracy of the Multilayer Perceptron method was 78.52%. However, the accuracy dramatically rose to 96.18% when features were chosen using FS6. The enhanced accuracy rates for FS4 and FS3 were 95.38% and 94.85%, respectively. The initial accuracy of the Decision Tree algorithm was 79.16%.

VI. Conclusion

We aimed to enhance heart disease risk prediction accuracy by employing ensemble classification techniques. We evaluated how well various classification algorithms, including Random Forest, Multilayer Perceptron, Decision Tree, and Nave Bayes, performed by comparing their accuracy. According to our research,

ensemble classification algorithms significantly improved heart disease risk prediction accuracy when compared to individual classifiers. We were able to provide predictions that were more accurate and dependable by using ensemble approaches like Bagging, Boosting, and Stacking, which capitalize on the advantages and combined knowledge of a number of classifiers. The algorithm with the most promise, Random Forest, had an initial accuracy of 81.53%. With the inclusion of feature selection, accuracy improved much more, with FS6 and FS2 attaining maximums of 88.18% and 90.52%, respectively. With FS6, the Multilayer Perceptron method's accuracy increased considerably from an initial accuracy of 78.52% to 96.18%. Similar improvements were also made in FS4 and FS3. Our results highlight the significance of feature selection in raising the accuracy of models that predict the likelihood of acquiring heart disease. By selecting the most relevant qualities, we were able to filter out noise and focus on the primary risk factors for heart disease, producing forecasts that were more accurate. Our research shows that the precision of heart disease risk prediction can be greatly increased by using feature selection and ensemble classification algorithms. With the use of these techniques, several classifiers can be combined in order to provide predictions that are more dependable and accurate. The findings of this study contribute to the improvement of heart disease risk prediction, which aids in the development of early diagnosis and prevention methods for better patient care and better health outcomes.

References

[1] Fida Benish, Nazir Muhammad, Naveed Nawazish, Akram Sheeraz. Heart disease classification ensemble optimization using genetic algorithm. IEEE; 2011. p. 19–25.

[2] Centers for Disease Control and Prevention (CDC). Deaths: leading causes for 2008. Natl Vital Stat Rep June 6, 2012;60(No. 6).

[3] EI-Bialy R, Salamay MA, Karam OH, Khalifa ME. Feature analysis of coronary artery heart disease data sets. Procedia Comput. Sci. 2015;65:459–68.

[4] Lee HeonGyu, Noh Ki Yong, Ryu Keun Ho. Mining biosignal data: coronary artery disease diagnosis using linear and nonlinear features of HRV. LNAI 4819: emerging technologies in knowledge discovery and data mining. May 2007. p. 56–66.

[5] Ajani, S.N., Mulla, R.A., Limkar, S. et al. DLMBHCO: design of an augmented bioinspired deep learning-based multidomain body parameter analysis via heterogeneous correlative body organ analysis. Soft Comput (2023). https://doi.org/10.1007/s00500-023-08613-y

[6] Singh Jagwant, Kaur Rajinder. Cardio vascular disease classification ensemble optimization using genetic algorithm and neural network. Indian J. Sci. Technol. 2016;9(S1).

[7] JyotiSoni Ujma Ansari, Sharma Dipesh. Predictive data mining for medical diagnosis: an overview of heart disease prediction‖. Int. J. Comput. Appl. March 2011;17(8). (0975 – 8887).

[8] Sudhakar K. Study of heart disease prediction using data mining. 2014;4(1):1157–60.

[9] Thenmozhi K, Deepika P. Heart disease prediction using classification with different decision tree techniques. Int J Eng Res Gen Sci 2014;2(6).

[10] KaanUyar Ahmet Ilhan. Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks. 9th international conference on theory and application of soft computing, computing with words and perception. Budapest, Hungary: ICSCCW; 2017. 24-25 Aug 2017.

[11] LathaParthiban, Subramanian R. Intelligent heart disease prediction system using CANFIS and genetic algorithm. Int. J. Biol. Biomed. Med. Sci. 2008;3(No. 3).

[12] Mackay J, Mensah G. Atlas of heart disease and stroke. Nonserial Publication; 2004.

[13] Vasighi Mahdi, Ali Zahraei, Bagheri Saeed, Vafaeimanesh Jamshid. Diagnosis of coronary heart disease based on Hnmr spectra of human blood plasma using genetic algorithm-based feature selection. Wiley Online Library; 2013. p. 318–22.

[14] Amin Mohammed Shafennor, et al. Identification of Significant features and data mining techniques in predicting heart disease. Telematics Inf 2019:82–93.

[15] Nahar J, Imam T, Tickle KS, Chen YPP. Computational intelligence for heart disease diagnosis: a medical knowledge driven approach. Expert Syst Appl 2013;40(1):96–104.

[16] 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.

[17] Drozd˙ z, 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

[18] Murthy, H.S.N.; Meenakshi, M. Dimensionality reduction using neuro-genetic approach for early prediction of coronary heart disease. In Proceedings of the International Conference on Circuits, Communication, Control and Computing, Bangalore, India, 21–22 November 2014; pp. 329–332

[19] Benjamin, E.J.; Muntner, P.; Alonso, A.; Bittencourt, M.S.; Callaway, C.W.; Carson, A.P.; Chamberlain, A.M.; Chang, A.R.; Cheng, S.; Das, S.R.; et al. Heart disease and stroke statistics—2019 update: A report from the American heart association. Circulation 2019, 139, e56–e528.

[20] Shorewala, V. Early detection of coronary heart disease using ensemble techniques. Inform. Med. Unlocked 2021, 26, 100655.

[21] Mozaffarian, D.; Benjamin, E.J.; Go, A.S.; Arnett, D.K.; Blaha, M.J.; Cushman, M.; de Ferranti, S.; Després, J.-P.; Fullerton, H.J.; Howard, V.J.; et al. Heart disease and stroke statistics—2015 update: A report from the American Heart Association. Circulation 2015, 131, e29–e322.

[22] 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 and Information System (ICIMCIS), Jakarta, Indonesia, 24–25 October 2019; pp. 45–48.

[23] Li, J.; Loerbroks, A.; Bosma, H.; Angerer, P. Work stress and cardiovascular disease: A life course perspective. J. Occup. Health 2016, 58, 216–219.

[24] Purushottam; Saxena, K.; Sharma, R. Efficient Heart Disease Prediction System. Procedia Comput. Sci. 2016, 85, 962–969.

[25] Soni, J.; Ansari, U.; Sharma, D.; Soni, S. Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction. Int. J. Comput. Appl. 2011, 17, 43–48.

[26] Mohan, S.; Thirumalai, C.; Srivastava, G. Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. IEEE Access 2019, 7, 81542–81554.

[27] Waigi, R.; Choudhary, S.; Fulzele, P.; Mishra, G. Predicting the risk of heart disease using advanced machine learning approach. Eur. J. Mol. Clin. Med. 2020, 7, 1638–1645.

[28] 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.

[29] Khan, I.H.; Mondal, M.R.H. Data-Driven Diagnosis of Heart Disease. Int. J. Comput. Appl. 2020, 176, 46–54.

[30] Kaggle Cardiovascular Disease Dataset. Available online: https://www.kaggle.com/datasets/sulianova/cardiovascular-diseasedataset (accessed on 1 November 2022).

[31] Khetani, V. ., Gandhi, Y. ., Bhattacharya, S. ., Ajani, S. N. ., & Limkar, S. . (2023). Cross-Domain Analysis of ML and DL: Evaluating their Impact in Diverse Domains. International Journal of Intelligent Systems and Applications in Engineering, 11(7s), 253–262.

[32] Rivero, R.; Garcia, P. A Comparative Study of Discretization Techniques for Naive Bayes Classifiers. IEEE Trans. Knowl. Data Eng. 2009, 21, 674–688.

[33] Khan, S.S.; Ning, H.; Wilkins, J.T.; Allen, N.; Carnethon, M.; Berry, J.D.; Sweis, R.N.; Lloyd-Jones, D.M. Association of body mass index with lifetime risk of cardiovascular disease and compression of morbidity. JAMA Cardiol. 2018, 3, 280–287.

[34] Allauddin Mulla, R. ., M. . Eknath Pawar, S. . S. Banait, S. . N. Ajani, M. . Pravin Borawake, and S. . Hundekari. “Design and Implementation of Deep Learning Method for Disease Identification in Plant Leaf”. International Journal on Recent and Innovation Trends in Computing and Communication, vol. 11, no. 2s, Mar. 2023, pp. 278-85, doi:10.17762/ijritcc.v11i2s.6147.

[35] Kengne, A.-P.; Czernichow, S.; Huxley, R.; Grobbee, D.; Woodward, M.; Neal, B.; Zoungas, S.; Cooper, M.; Glasziou, P.; Hamet, P.; et al. Blood Pressure Variables and Cardiovascular Risk. Hypertension 2009, 54, 399–404.

[36] Yu, D.; Zhao, Z.; Simmons, D. Interaction between Mean Arterial Pressure and HbA1c in Prediction of Cardiovascular Disease Hospitalisation: A Population-Based Case-Control Study. J. Diabetes Res. 2016, 2016, 8714745.

[37] Huang, Z. A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining. DMKD 1997, 3, 34–39. 30. Maas, A.H.; Appelman, Y.E. Gender differences in coronary heart disease. Neth. Heart J. 2010, 18, 598–602.

[38] Bhunia, P.K.; Debnath, A.; Mondal, P.; D E, M.; Ganguly, K.; Rakshit, P. Heart Disease Prediction using Machine Learning. Int. J. Eng. Res. Technol. 2021

[39] Mr. Anish Dhabliya. (2013). Ultra Wide Band Pulse Generation Using Advanced Design System Software . International Journal of New Practices

in Management and Engineering, 2(02), 01 - 07. Retrieved from http://ijnpme.org/index.php/IJNPME/article/view/14

[40] Sharma, S. ., Kumar, N. ., & Kaswan, K. S. . (2023). Hybrid Software Reliability Model for Big Fault Data and Selection of Best Optimizer Using an Estimation Accuracy Function . International Journal on Recent and Innovation Trends in Computing and Communication, 11(1), 26–37. https://doi.org/10.17762/ijritcc.v11i1.5984

[41] Anupong, W., Azhagumurugan, R., Sahay, K. B., Dhabliya, D., Kumar, R., & Vijendra Babu, D. (2022). Towards a high precision in AMI-based smart meters and new technologies in the smart grid. Sustainable Computing: Informatics and Systems, 35 doi:10.1016/j.suscom.2022.100690