Research Article

An Ensemble Basedheart Disease Predictionusing Gradient Boosting Decision Tree

1st S.Irin Sherly, 2nd G. Mathivanan

Research Scholar
Sathyabama Institute ofScience andTechnology
Chennai, India
irinsherly15@gmail.com

Professor
Dept of IT
Sathyabama Institute of Science andTechnology
Chennai, India
mathivanan.it@sathyabama.ac.in

Article History: Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 28 April 2021

ABSTRACT: Heart disease is now one of the deadliest diseases among the world. This crisis is affecting the bulk of people across the globe. Considering the massive death rate and huge amount of people suffering from the disease, the importance of early diagnosis of heart disease has been proven. Known cause for forecasting such a disease exist, but they really do not seem to be adequate. It is critical to develop a standard medical device that can foresee early heart diagnosis and have a more precise diagnosis than existing technologies such as Logistic Regularization, Lasso, Elastic Net, and Lasso Community Regularization.Ensemble classifiers are used in a variety of machine learning models that can increase forecasting ability in healthcare. Four databases are assembled in this paper, and 14 clinical features are fed into Ensemble. Traditional methods like SVM, AdaBoost, Logistic Regularization, and the existing Ensemble Prediction Model are compared to the proposed Ensemble Prediction Model in this paper. Across each experiment, the accuracy rate of the four datasets was 99 percent, outperforming other machine learning techniques and related academic studies. The performance metrics clearly show that the developed ensemble learning approach is superior. The findings show that the suggested ensemble can accurately predict the risk of cardiovascular disease.

Keywords: Cardiovascular disease, Machine Learning, Ensemble learning techniques, Gradient Boosting

1. INTRODUCTION

According to a recent WHO survey, 17.9 million people die each year. It's unsurprising that by 2030, it'll be 75 million. India had some of the highest rates of cardiovascular death in the country (CVD). In India, the mortality rate due to cardiovascular disease is set to reach from 2.26 million in 1990 to 4.77 million in 2020 [1].Over the last few generations, the worldwide proportion of coronary heart disease in India has increased from 1.6 to 7.4 percent within remote communities and from 1% to 13.2 percent in urban communities [2].Most cardiovascular diseases (CVDs) destroy individuals in various of ways, including nicotine use, unhealthy diet, physical dormancy, and deadly liquor utilize. People with CVD are more likely to suffer a heart attack. As previously stated, the disease necessitates early detection and instruction on the use of short-term drugs.Ultimately, the production of fatty stores inside the ducts and blood groups concludes CVD. Tissue damage in locations like the head, eyes, heart, and kidneys can also trigger it.Strokes and cardiac accidents are frequently caused by powerful incidents, and a blood clot that inhibits blood flow to the brain or heart is a major condition. Many of the most difficult responsibilities here is to foresee the disease that exists in the human body, which has piqued the attention of researchers. In addition, health professionals are inept at forecasting disease [3]. In order to predict the illness, they often need a support network. Some enhancements are encouraged, but the achievement of the concept must be demonstrated even outside the existing system.As a result, there is a significant amount of research being conducted to assist physicians in predicting human CVD disease. Meanwhile, researchers have developed a variety of machine learning models depending on the known risk of heart disease [4].

Many areas in medical science have incorporated machine-learning methods. But at the other hand, researchers looked for opportunities to grow and strengthen these structures. Ensemble learning, on the other hand, has been shown to improve computational tasks [5]. The Ensemble Classifier is mainly composed of a framework for combining element expectations, such as majority voting, as well as several specific classifiers.Assembly classifiers

outperform traditional classifiers in certain cases, according to research [6]. To accurately determine the suggested procedure's efficacy, data from Switzerland, Hungary,Clevelandand the VA Long Beach Data sources are being used. After that, a comparative study with certain recent scholarly work including well machine learning algorithms such as the Bagged Decision Tree, SVM, AdaBoost, Gradient Boosting DT, and random forest is performed. The suggested framework had a precision of 98 percent and an accuracy of 99 percent. In comparison to other existing models, the proposed model's current results suggest that it is effective at predicting cardiovascular disease.

2. RELATED WORK

Using data mining methods, [7] predicts the emergence of cardiovascular disease. The probability of developing heart disease is expressed as a percentage. The medical parameters classified in the databases are evaluated using a data mining classification methodology. Python programming is used to analyze the datasets using two machine learning algorithms, the Nave Bayes Algorithm and the Decision Tree Algorithm. The best algorithm is chosen among these two algorithms based on the overall accuracy of heart disease.

Using the WEKA tool, alternative Decision Trees classification models are applied in the search for better results in the diagnosis of heart disease [8,]. Even amongst the algorithms being studied are the J48, the Random Forest, and the Logistic Model Tree. The focus of this study was to use data mining methodologies to uncover hidden patterns associated with heart diseases and to estimate the prevalence of heart disease in patients, which ranged from non-existent to potentially present.

The relevance and accuracy of clinical machine learning methods such asNeural Network, Logistic Regression,Support Vector Machine, Random Forest and Decision Tree were studied by Beunza et al.[9]. Zhao et al. [10] used time analysis, machine learning, and CNN models to look at the pulse shift cardiovascular rate. The support vector machine outpaced all classification algorithms in the evaluation of cardiovascular recognition. Over the course of a year, Chen et al. [11] planned to use machine learning models to track cardiac problems in patients with chronic DCM.The ML algorithm gave 32 clinical expertise highlights, and Information Gain (IG) chose key functions that were extremely relevant to cardiovascular risk. Drug-induced cardiovascular pathogens in humans were investigated in this study [12]. The Base of Fisiologiab Clinica used two ML methods: one was the Stomach Associated Association and the Vault of Renal Diseases and other was American database promoting regional diabetes.

Certain ensemble learning methods are discussed in this section. Classification algorithms based on machine learning are commonly used in a number of fields. As a result, developers are actively developing new methods to increase classification accuracy. Ensemble learning, which may be simple or complex, is one approach. Bootstrap aggregating (bagging) [13], boosting [14], and random decision forest [15] are instances of ensemble classification algorithms. When these ensembles have been used, the classification efficiency typically improves. Others, on the other hand, have used a range of methodologies to accomplish ensemble learning, such as methods for merging different classifiers or separating data using majority vote, among other approaches.

A research by Leon et al. [16] looked further into impacts of different voting processes on classifiers. On databases of great variety, the thesis looked at just how different voting methods influenced the effectiveness of various algorithms. Despite the widespread use of plurality voting in the literature, the test results indicate that the single transferable form could be a viable choice.Banfield et al. [17] compared seven different randomization-based techniques for building decision tree ensembles, including bootstrap aggregation. When using publicly available databases, Random Forests, Boosting, and Randomized Trees outpaced bagging, according to the results.

While some ensemble learning techniques concentrate on combining two base classifiers, ensembles could also be built by segmenting the database into subgroups and combining the findings. Additionally, without the use of bagging, boosting, or random forest, a variety of datasets can be gathered. To create separate datasets, Ruta et al. [18] used random permutation and splitting up the real data.Result of individual models' greatly improved consistency, the corresponding ensemble attained impressive generalization. Ensemble learning has since been applied to a variety of medical diagnostic challenges [19, 20].

Yadav and Pal [21] performed their studies at the University of California, Irvine's repository. This data consists 14 attributes. Random TreeFour tree-based classification procedures were used in the classification: M5P, Reduced Error Pruning Method, and the Ensemble Random Forest Technique. In this analysis, three feature-based mechanisms were used: Recursive Feature Elimination, Pearson Correlation, and Lasso Regularization. Precision and accuracy of the procedures have also been linked. The final approach was the most successful. In a recent article [22], Gupta et al. developed a machine intelligence platform using factor analysis of mixed data (FAMD) and RF-based MLA.

Rashmi et al. [23] tested dataset 303, which they got from the Cleveland datasets. Decision Tree, the proposed algorithm, had a 75.55 percent accuracy rate. Dinesh et al. [24] examined 920 databases from the University of California, Irvine's machine learning repository (Long Beach, Cleveland,Switzerland, VA, and Hungarian). Random

Forest reached 80.89 percent precision mostly on AFIC dataset, and Saqlain attained 68.6 percent accuracy [25]. On a same data, Sharma et al. [26] as well as Dwivedi et al. [27] used the K-Nearest Neighbors methodology. The probabilities were 90.16 percent and 80 percent, respectively.Enriko [28] used the Kita Hospital Jakarta (450) dataset to attain a 46 percent accuracy. Kaur et al. [29] achieved a strong outcome using AdaBoost on the Cleveland dataset, for instance 56.13 percent.Shetty et al. [30] achieve 89 percent accuracy with the 270 samples from the Statlog dataset, and Chaurasia et al. [31] attain 75.9% accuracy with a Boosting hybrid approach. The effectiveness of the Boosting ensemble technique was also tested using the UCI laboratory dataset. Cheng et al. and Chaurasia et al. used an ANN model to achieve 82.5 percent accuracy [32] and a hybrid model to achieve 78.88 percent accuracy [31].

Dinesh et al. [24] used a combination of four datasets to achieve 84.27 percent accuracy, while Bhuvaneeswari et al. [33] used 583 records from the Cleveland and Statlog datasets to achieve 95.19 overall accuracy. On the Rajaie cardio vascular medical dataset [34], a survey result was developed with 79.54 percent accuracy by using hybrid method. The Decision Tree [35] Bagging technique, on the other hand, yielded an average of more than 85.03 percent. Three separate datasets were merged into one to achieve a more coherent result. Mohan et al. [36] used a hybrid system to achieve an accuracy of 88.4 percent. Latha et al. [31] used the Bagging technique to achieve 80.53 percent accuracy with 303 datasets of Cleveland heart disease. Tan et al. [37] assembled 303 samples from the Cleveland Heart Disease Database using a hybrid technique, also attained an accuracy of 84.07 percent, whereas Latha et al. [31] reached 85.48 percent. We hope to build on previous research and create a solid ensemble learning forecasting model for cardiovascular disease risk in this work[40,41,42].

3. METHODOLOGYUSED

The heart disease data sets are incorporated in this section. The reports are saved in a database. To generate feedback for the trained model, attributes are chosen from the uploaded data. In the trained model, certain characteristics are processed. The output is expressed in terms of 0, 1, 2, 3, and 4. The Gradient Boosting algorithm is used to develop heart disease classification and predictions based on ensemble learning. Finally, the ensemble learning model's assessment methods are explained in depth. The overview of the proposed system is shown in Figure 1.

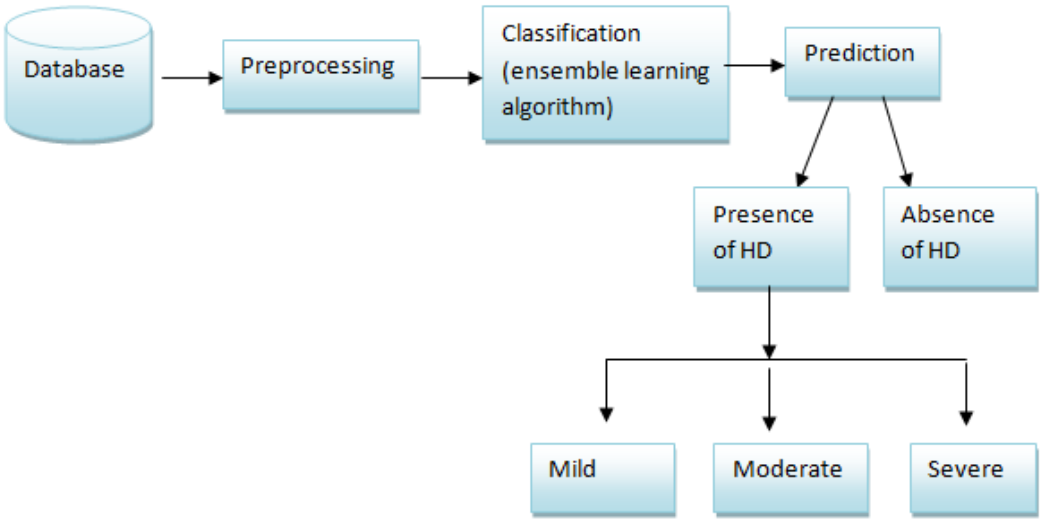

Figure 1. Proposed methodology for Heart disease prediction

A. Dataset forCardiovascular Disease

The data for the cardiovascular disease sets in this study came from the Heart Disease Databases of the UCI Machine Learning Repository [13]. The Cleveland Clinic Foundation (CCF), the Hungarian Institute of Cardiology (HIC), the Long Beach Medical Center (LBMC), and the University Hospital in Switzerland all contributed reports on heart disease clinical cases. Each set of data in the databases contains 303, 294, 200 and 123 clinical incidents accordingly.As an outcome, there are a total of 920 clinical incidents.Each patient's clinical instances appear to be the same in each heart disease database. There are 300 different examples in total, each with 76 different traits. The only 14 characteristics taken into consideration by this technique are age, gender, chest pain, trestbps, chol, fbs,

restecg, thalch, exang, oldpeak, slope, ca, thal, and target. Datasets are used to derive disease-specific patterns. Preparedness and study records are separated into two categories.

B) Preprocessing

The merged dataset is searched for missing values during data preprocessing. The best features from the dataset are then extracted. The efficiency of classifiers created with these techniques is studied, and also the original features. The dataset is divided into two sections after feature selection: training and testing. 80 percent of the data is allocated to the training phase based on model learning rates, while the remaining 20% is allocated to the testing phase. To draw a contrast over the consolidated dataset, all ensemble models with classifiers are presented. The dataset included a variety of features with differing magnitudes, ranges, and units. This is a major stumbling block since our proposed embedded algorithms are extremely sensitive to these characteristics. It's crucial to avoid the Machine Learning algorithms overfitting to the wrong features. The data is structured to compensate for variations in the measurements used for different features using the standardization scaling technique. With a unit standard deviation, the values are based around the mean.As a result, the attribute's mean is set to zero, as well as the resulting distribution has a standard deviation of one[43].

The standardization formula is as follows:

$$X' = \frac{X-\mu}{\sigma} \quad (1)$$

μ is the mean of the feature values, σ and is the standard deviation of the attribute values. It's essential to mention that the ratings in this case aren't restricted to a small range.

The degree to which the traits are associated to one another or to the target variable is indicated by correlation. A positive correlation (a rise in one feature's value increases the significance of the target variable) or a negative correlation (a fall in one feature's value lowers the value of the target variable) exists (increase in one value of the feature decreases the value of the target variable). The ability to classify the characteristics which are most relevant to the target variable is provided by using a heatmap. A Correlation Coefficient indicates a high correlation value near 1. As a result, values from the Heatmap that correspond most closely with the first column are crucial for training Machine Learning algorithms.Since the knowledge for the Machine Learning to train is already present, values near 1 between other attributes are not ideal. We created a data visualization of the number of patients with and without a heart disorder. Figure 2 illustrates the connection among all features using a heat map.

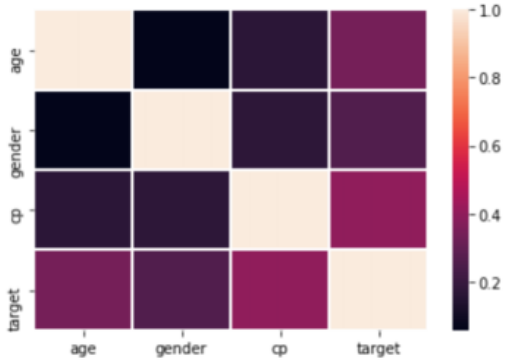

Figure 2: Heatmap with Correlation Coefficient for Features

Males are more likely than females to develop heart disease, according to the data. Men are more likely than women to have a heart attack. Figure 3 depicts a countplot on gender vs data, with the person's sex depicted as Male = 1 and Female = 0.

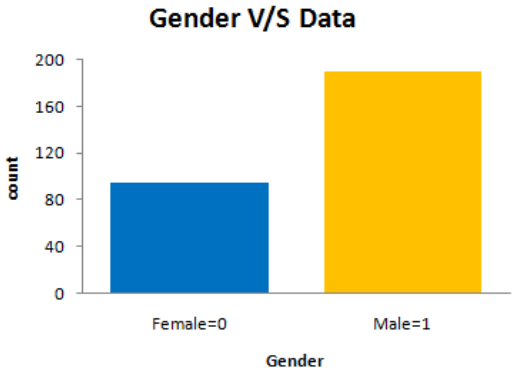

Figure 3: Countplot on Gender v/s Dataset where the person's sex (1=male, 0=female)

The aim feature refers to whether or not the patients have heart disease. It's measured on a 0 to 4 scale, with 0 denoting no heart disease and 1, 2, 3, and 4 denoting the prevalence and the intensity of cardiovascular disease, accordingly.The countplot on target v/s data can be seen in Figure 3. The thalach function, on the other hand, refers to a person's maximum rate. Heart disease is more common in individuals who have a high blood pressure of over 140. Figure 4 depicts the scatterplot of age vs. thalach.The effects of the blood flow as seen by the radioactive dye are expressed by Thalach.

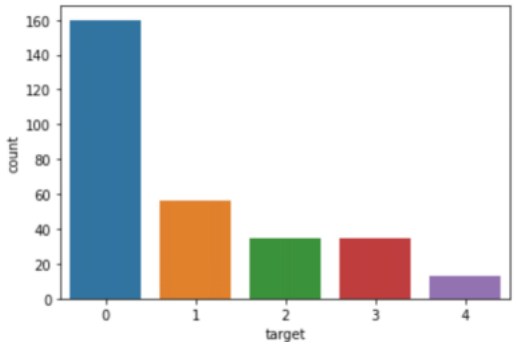

Figure 3: Countplot on Target v/s Data

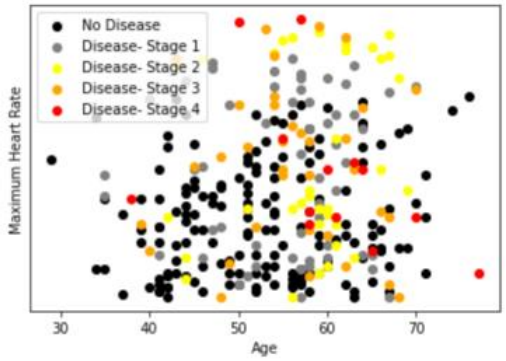

Figure 4. Scatter plot between Age and Thalach

C) Gradient Boosting Classifier (GBC):

Gradient boosting is a regression and classification machine learning method that generates a prediction process consists of a set of weak estimation methods, the most prevalent of which is a decision tree.It combines several weaker models into a strong, large model with highly predictive performance. Models of this kind are common because of their ability to efficiently classify datasets. In most cases, decision trees are used to construct models for gradient boosting classifiers. Optimizing a loss function, teaching a weak learner to make accurate predictions, and

adapting weak learners to an optimization strategy to lessen the loss function are all key features of Gradient Boosting.The loss function must be one-of-a-kind and specific to the problem under consideration.Gradient Boosting considers Decision Trees to be the feeble learner.Because regression trees generate real values as dispersed outcomes and their outputs can be added up, they can be used to insert consecutive model outputs and correct residuals in predictions.Trees are built to minimize loss or to choose the best split score based on Gini's purity ratings [10]. In an additive model, trees are presented one at a time, with earlier trees in the framework remaining intact. When adding trees, the gradient descent approach can reduce the odds.

1. Loss Function

The solution domain determines the type of loss function used. It must be distinct, but there are many commonly used loss functions and also the potential to construct your own. A squared error, for example, can be used in regression, and a logarithmic loss could be used in classification. The gradient boosting mechanism has the advantage of not requiring each differentiable loss function to define a new boosting algorithm; rather, because it is such a basic approach, any variational loss function is being used.

2. Weak Learner

For gradient boosting, decision tree algorithm were often used as a weak learning process. Regression trees were included because they deliver superior split values and it can be linked together, enabling subsequent model outputs to be introduced and residuals in predictions to be "corrected." Trees are designed greedily, with the best split tips chosen based on purity scores inorder to reduce loss. Decision stumps, or very short decision trees with only one break, were used in the beginning, as in the case of AdaBoost. Weak learners are often restrained in some way, such as the number of layers, nodes, breaks, or root node they can access.

3. Additive Model

Existing trees in the model aren't substituted; instead, new trees are added one by one. The effect of adding trees is reduced using a gradient descent strategy. In neural networks, gradient descent is widely used to explain a set of parameters, such as regression coefficients or weights. To get the precision, the weights are changed after the error or loss has been computed. Rather than restrictions, poor learner sub models, or more accurately decision trees, are used. After computing the loss before running the gradient descent procedure, we can create a tree to the model to mitigate the loss (i.e. follow the gradient).

To implement a gradient boosting classifier, the number of steps carried out are as follows:

a) Fit the model;

b) Fine-tune the parameters and hyperparameters of the model;

c) Make predictions

d) Evaluate the outcomes.

The algorithm of Gradient boosting:

1. Set $f_o(x) = \arg min_\gamma \sum_{i=1}^{N} L(y_i, \gamma)$ as initial value.
2. For m=1 to M:
   a) For i= 1,2,……,N calculate
   $$r_{im} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f=f_{m-1}}$$
   b) Create terminal regions by fitting a regression tree to the targets r_{im}.
   $R_{jm}, j = 1,2, \ldots . J_m$
   c) In the case of j= 1,2,…., J_m calculate
   $$\gamma_{jm} = \arg\ \min_\gamma \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma).$$
   d) Update the information
   $$f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm}\, I(x \in R_{jm}).$$
3. $\hat{f}(x) = f_M(x)$, is the output.

By using gradient boosting technique, a parameter importance of the attribute related to predicting heart disease in this datasets is given.

D. The Methods of Gradient Boosting Model Evaluations

Examining the model's sensitivity (also known as recall), F1score, and comparing them using the training and testing data sets is another of the effective methods to quantify the accuracy and misclassification error of Gradient Boosting Ensemble classification and forecasts. These figures are determined by the number of false positive and

false negative cardiovascular disease instances discovered in this study. Table 1 summarizes the confusion matrix for positive and negative outcomes from the ensemble learning classification and prediction methodology for confirming heart disease risk or absence.

Table I. A confusion matrix depicts the diagnostic results of developed ensemble learning classification and prediction models for distinguishing between the existence and absence of heart disease.

	Presence of Heart Disease	Absence of Heart Disease	Total Number
Predicted	True Positive (TP)	False Positive (FP)	TP+FP
Unpredicted	False Negative (FN)	True Negative (TN)	FN+TN
Total Number	TP+FN	TN + FP	FP+TP+TN+FN

The probability of properly acknowledging the presence of heart disease patients is referred to as sensitivity/recall. The percentage of properly expected positive observations to all events in the actual class is measured by recall, which is a quantity metric.

$$\text{Recall} = \frac{TP}{TP+FN} \tag{2}$$

The most straightforward success metric is accuracy, which is essentially the number of correctly expected observations to all observations. Accuracy is a valuable measure, and even when you have symmetric datasets with virtually equal values for false positives and false negatives. As a result, other parameters must be considered when evaluating the efficiency of your model. The accuracy of the model is calculated by

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{3}$$

The ratio of True Positives to all Positives is known as precision. That will be the number of patients who we correctly classify as having heart disease out of all the patients who currently have it for our problem statement. In terms of mathematics:

$$\text{Precision} = \frac{TP}{TP+FP} \tag{4}$$

The weighted average of Precision and Recall is the F1-Score. As a consequence, F1-score takes both false positives and false negatives into account. Despite the fact that it is less clear, F1 is usually more useful than accuracy. This is particularly true if the distribution of classes is unequal. Accuracy increases when the costs of false positives and false negatives were equivalent. It's important to view at F1-Score when the cost of false positives and false negatives are somewhat distinct.We may simply aim for a high F1-score, which would also reflect a high Accuracy and Recall value, rather than weighing precision and recall.

$$F1score = \frac{2TP}{2TP+FP+FN} \tag{5}$$

(or)

$$F1score = 2 * \frac{Precision*Recall}{Precision+Recall}$$

where an F1-score of 1 depict the classification and forecast system's highest level of accuracy, and an F1-score of 0 depict the system's lowest level of accuracy. As a consequence, the F1-score in Eq. (4) is used to evaluate model efficacy as well as the precision of a model throughout study.

4. Experimental Resultsand Discussion

The Gradient Boosting Decision Tree Algorithm is used in this study to develop an ensemble machine learning technology for precision cardiovascular diagnosis and effect forecasting. The proposed Gradient Boosting model is one of the ensemble machine learning algorithm that converges many computational theories into a good and enhanced classification and prediction framework by combining a group of results from several other learning

algorithms into a weighted average.As various machine learning algorithms are used to test the models, each algorithm offers a different accuracy rate for the characteristics that are known to be the cause of cardiovascular disease. The Uncertainty Matrix plots [38] can be used to measure sensitivity or recall. It is necessary to avoid getting too many false-positives or false-negatives when diagnosing cardiovascular disease. As a result, the model with the highest overall accuracy is specified (accuracy is the sum of the diagonals. on the confusion matrix divided by the total). Table II indicates the number of cases with or without heart failure in each of the four data sets, as well as the percentage. As shown in Table II, the number of patients with heart disease varies significantly across the data sets, with the lowest being 36.18 percent and the highest being 93.44 percent. Table III compares the results of various algorithms for various parameters. The model's classification report reveals that 99 percent of the time, the absence of heart disease was correctly predicted, and 97 percent of the time, the occurrence of heart disease was correctly predicted.

Table II. Information on the number of patients with or without heart disease.

	Individual with presence of Heart disease	Individual with absence of Heart disease	% of Presence of Heart disease in Database
CCF	125	157	44.33%
HIF	106	187	36.18%
SUH	114	8	93.44%
LBMC	113	32	77.44%

Table III. Comparison result of different algorithm for different parameters.

Parameters	Bagging DT	Gradient Boosting DT	KNN	Random forest	SVM	Logistic regression
Accuracy	96.6%	99%	71.6%	98.3%	81.6%	86.6%
Recall	80%	97%	42%	90%	58%	75%
Precision	75%	98%	54%	94%	58%	69%
F1 score	77%	97%	41%	92%	57%	70%
Model Score	97%	99%	72%	98%	82%	87%

The Confusion Matrix for the different Classifier algorithms included in this scheme, such as Bagging Decision Tree, Gradient Boosting DT, KNN, Support Vector Machine, Random forest, and Logistic Regression, can be seen in the Figure 5 (a), (b), (c), (d), (e), (f), where “0,1,2,3,4” corresponds to Healthy, Stage 1, Stage 2, Stage 3, and Stage 4, respectively.

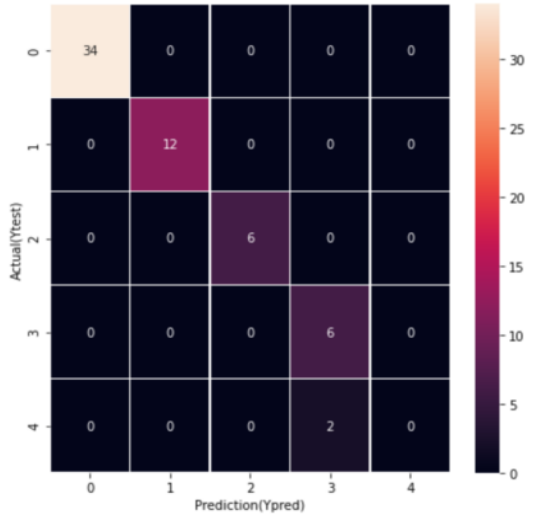

Figure 5 a) BaggingDT Classifier Confusion matrix

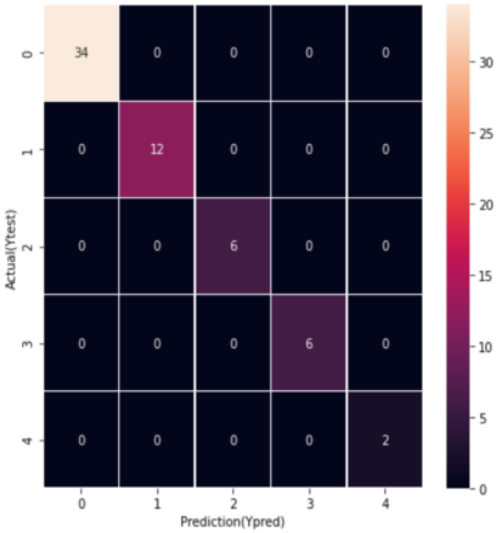

Figure 5 b) Gradient Boosting DT classifier Confusion matrix

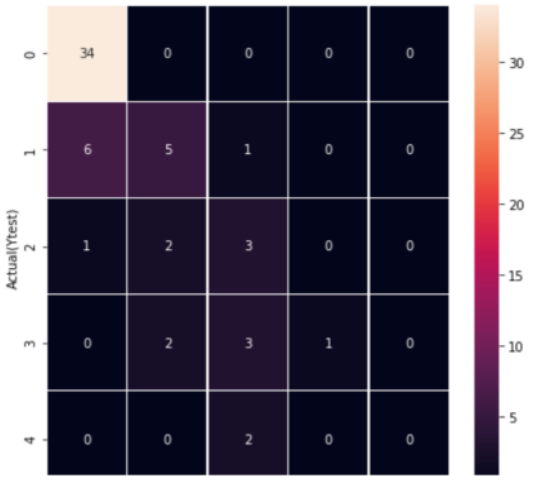

Figure 5 c) Confusion Matrix for KNN Classifier

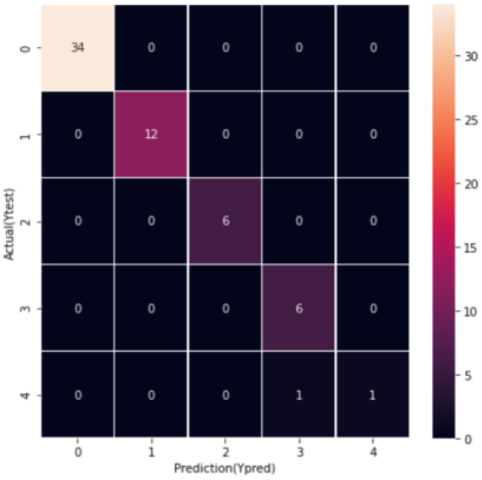

Figure 5d) ConfusionMatrix for Random forest Classifier

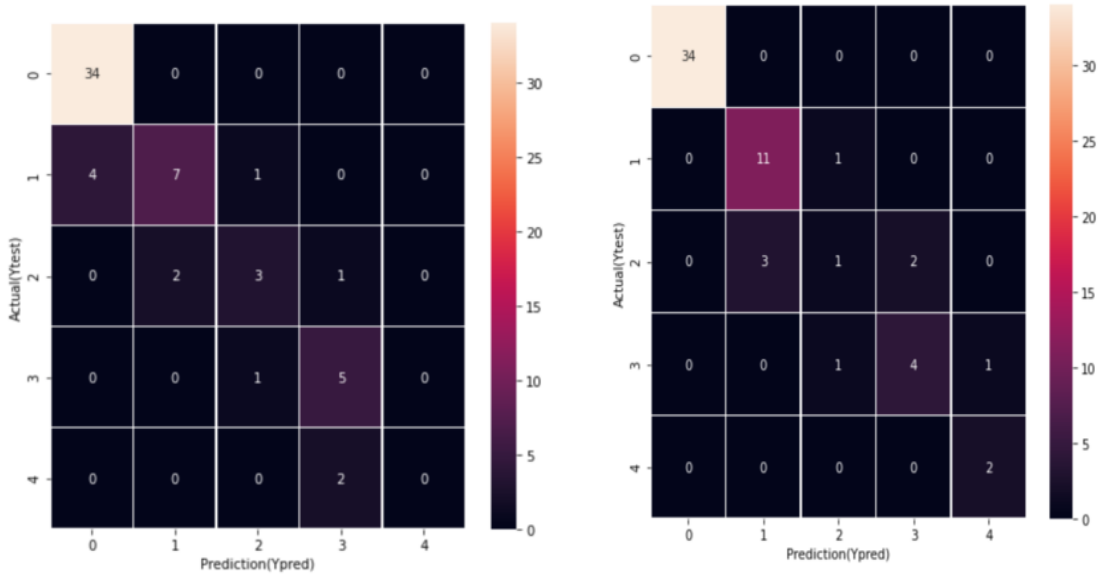

Figure 5 e) Confusion Matrix for SVM Classifier

Figure5 f) Confusion Matrix for Logistic Regression Classifiers

Figure 6 a) represents different output parameters for various Classifier algorithms used in this procedure, such as Bagging DT, Gradient Boosting DT, KNN, Random forest, Support Vector machine, and Logistic Regression, such as Accuracy, Recall, Precision, F1 Score, and Model Score, for various Classifier algorithms used in this method, such as Bagging DT, Gradient Boosting, KNN, Random forest, Support Vector machine, and Logistic Regression. Gradient Boosting DT has the highest accuracy (99%), recall (97%), precision (98%), F1 score (97%), and score (97%) of any method (99 percent).

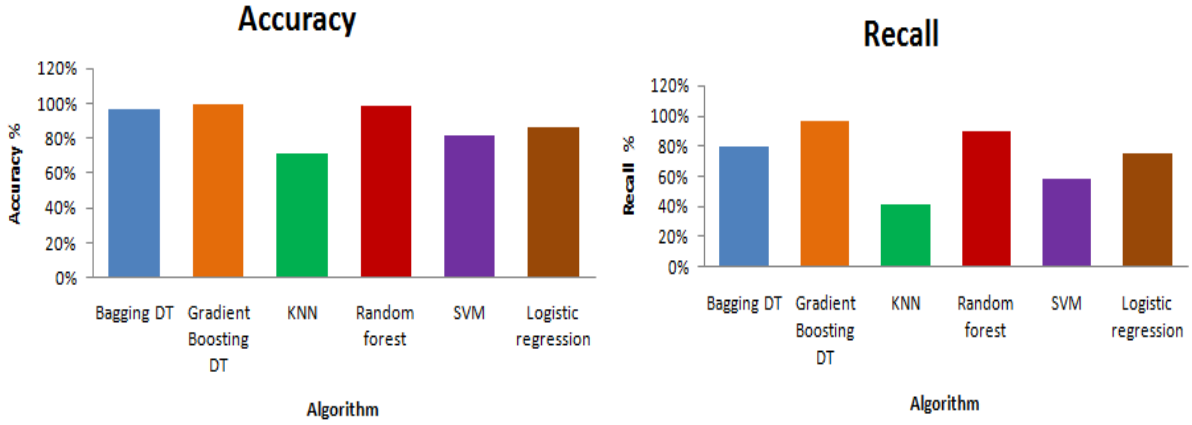

Figure 6 a) Percentage of Accuracy for different Algorithm

Figure 6 b) Percentage of Recall for different Algorithm

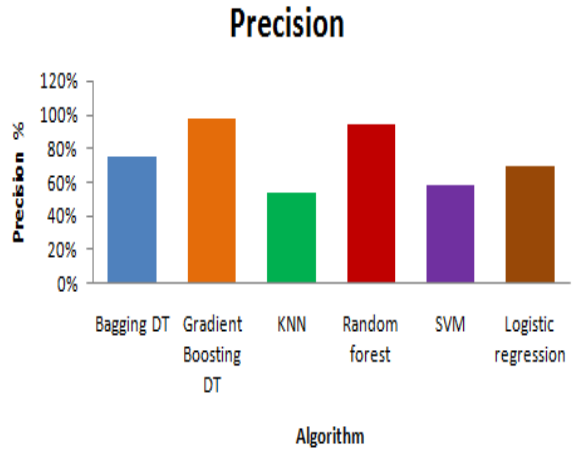

Figure 6 c) Percentage of Recall for different Algorithm

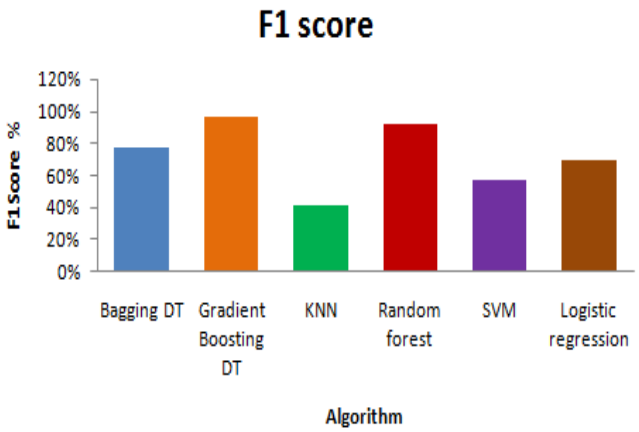

Figure 6 d) Percentage of F1 score for different Algorithm

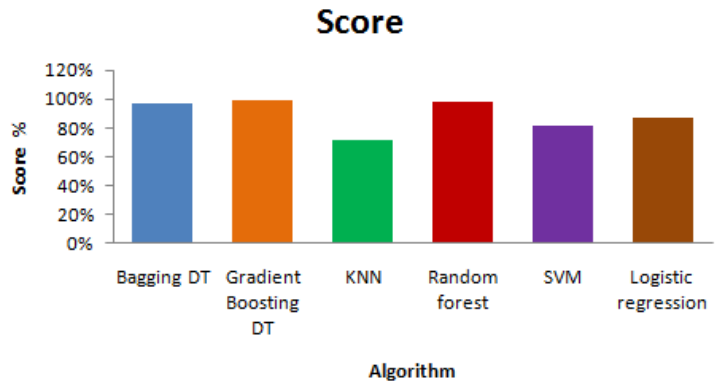

Figure 6 e) Percentage of Scoredifferent Algorithm

V. CONCLUSION AND FUTURE WORK

This study used ensemble learning classification and modeling techniques to identify and classify the existence and absence of cardiovascular disease in patient outcome forecasts, as well as model precision and accuracy, sensitivities (or recalls), F1 ratings, and confusion matrix. While using an appropriately weighted majority vote of a number of weak classifiers using the Gradient Boosting algorithm, the designed classification and prediction models had fantastic features such as manipulating a weighting vector to generate a strong, single aggregate ensemble classification and predictive models. The current ensemble learning classification and modeling techniques, which were made up of four different data sets from four different hospitals, were trained and checked using the stronghold method. The developed ensemble learning classification and forecast models had a model summary accuracy of 99 percent, an average sensitivity (or recall) of 97 percent, an average precision of 98 percent, an average model F1-score of 97 percent, and an average model Score of 99 percent in assessing both the occurrence and exclusion of cardiovascular disease.As a result, the developed ensemble learning modeling and classification techniques, which use 14 input parameters, provide extremely accurate and timely evaluations for cardiovascular disease patient outcomes projections, allowing patients to prevent unnecessary, erroneous assessments and clinical therapy.We want to generalize the model even further in the future since it can be combined with many other feature selection tools and applied to datasets with a lot of missing data. Convolutional Neural Networks are another potential future solution.The study's main goal was to improve on previous work by devising a fresh and novel method of building the model, and to make the model practical and simple to implement in real-world scenarios.

REFERENCES

1. Murray CJ, Lopez AD. Alternative projections of mortality and disability by cause 1990–2020: Global Burden of Disease Study. *Lancet.* 1997;349:1498–504. [PubMed] [Google Scholar]
2. Gupta R, Joshi P, Mohan V, Reddy KS, Yusuf S. Epidemiology and causation of coronary heart disease and stroke in India. *Heart.* 2008;94:16–26. [PubMed] [Google Scholar]
3. Jabbar, M.A., Chandra, P., Deekshatulu, B.L. (2012). Prediction of risk score for heart disease using associative classification and hybrid feature subset selection. 2012 12th International Conference on Intelligent Systems Design and Applications (ISDA), Kochi, pp. 628-634.
4. Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques Informatics in Medicine Unlocked, 16 (Jan. 2019), p. 100203.
5. R.K. Sevakula, N.K. Verma "Assessing generalization ability of majority vote point classifiers" IEEE Transactions on Neural Networks and Learning Systems, 28 (12) (Dec. 2017), pp. 2985-2997, 10.1109/TNNLS.2016.2609466
6. H. Li, *et al.* Ensemble learning for overall power conversion efficiency of the all-organic dye-sensitized solar cells IEEE Access, 6 (2018), pp. 34118-34126.
7. Shadman Nashif, Md. Rakib Raihan, Md. Rasedul Islam, Mohammad Hasan Imam, "Heart Disease Detection by Using Machine Learning Algorithms and a Real-Time Cardiovascular Health Monitoring System", Scientific Research Publishing, World Journal of Engineering and Technology, ISSN Online: 2331- 4249 ISSN Print: 2331-4222 November 2018, Vol. 6, pp. 854-873.
8. R. Alizadehsani, et al. "Machine learning-based coronary artery disease diagnosis: A comprehensive review", Computers in Biology and Medicine, Elsevier Ltd. June 2019, pp. 01-14
9. Beunza, J.J., Puertas, E., García-Ovejero, E., Villalba, G., Condes, E., Koleva, G., Hurtado, C., Landecho, M.F. (2019). Comparison of machine learning algorithms for clinical event prediction (risk of coronary heart disease). Journal of Biomedical Informatics, 97: 103257.
10. Zhao, L., Liu, C., Wei, S., Liu, C., Li, J. (2019). Enhancing detection accuracy for clinical heart failureutilizing pulse transit time variability and machine learning. IEEE Access, 7: 17716-17724.
11. Chen, R., Lu, A., Wang, J., Ma, X., Zhao, L., Wu, W., Du, Z., Fei, H., Lin, Q., Yu, Z., Liu, H. (2019). Usingmachine learning to predict one-year cardiovascular events in patients with severe dilated cardiomyopathy. European Journal of Radiology, 117: 178-183.
12. Mezzatesta, S., Torino, C., De Meo, P., Fiumara, G., Vilasi, A. (2019). A machine learning-based approach for predicting the outbreak of cardiovascular diseases in patients on dialysis. Computer Methods and Programs in Biomedicine, 177: 9-15.
13. L. Breiman, Bagging predictors, Mach Learn, 24 (2) (Aug. 1996), pp. 123-140.
14. R.E. Schapire, Y. Singer, Improved boosting algorithms using confidence-rated predictions, Mach Learn, 37 (3) (Dec. 1999), pp. 297-336.
15. Tin Kam Ho ,The random subspace method for constructing decision forests, IEEE Trans Pattern Anal Mach Intell, 20 (8) (Aug. 1998), pp. 832-844.
16. F. Leon, S.-A. Floria, C. Badica, Evaluating the effect of voting methods on ensemble-based classification, 2017 IEEE international conference on Innovations in intelligent Systems and applications (INISTA) (Jul. 2017), pp. 1-6,
17. R.E. Banfield, L.O. Hall, K.W. Bowyer, W.P. Kegelmeyer, A comparison of decision tree ensemble creation techniques, IEEE Trans Pattern Anal Mach Intell, 29 (1) (Jan. 2007), pp. 173-180,
18. D. Ruta, B. Gabrys, C. Lemke, "A generic multilevel architecture for time series prediction",IEEE Trans Knowl Data Eng, 23 (3) (Mar. 2011), pp. 350-359.
19. B. Zhang, *et al.* "Ensemble learners of multiple deep CNNs for pulmonary nodules classification using CT images",IEEE Access, 7 (2019), pp. 110358-110371.
20. L. Han, S. Luo, J. Yu, L. Pan, S. Chen, "Rule extraction from support vector machines using ensemble learning approach: an application for diagnosis of diabetes", IEEE Journal of Biomedical and Health Informatics, 19 (2) (Mar. 2015), pp. 728-734.
21. D. C. Yadav and S. Pal, ''Prediction of heart disease using feature selection and random forest ensemble method,'' Int. J. Pharmaceutical Res., vol. 12, no. 4, 2020.
22. A. Gupta, R. Kumar, H. S. Arora, and B. Raman, ''MIFH: A machine intelligence framework for heart disease diagnosis,'' IEEE Access, vol. 8, pp. 14659–14674, 2020.
23. G. O. Rashmi and U. M. A. Kumar, ''Machine learning methods for heart disease prediction,'' Int. J. Eng. Adv. Technol., vol. 8, no. 5S, pp. 220–223, May 2019.

24. M. Saqlain, W. Hussain, N. A. Saqib, and M. A. Khan, ‘‘Identification of heart failure by using unstructured data of cardiac patients,’’ in Proc. 45th Int. Conf. Parallel Process. Workshops (ICPPW), Aug. 2016, pp. 426–431.
25. K. G. Dinesh, K. Arumugaraj, K. D. Santhosh, and V. Mareeswari, ‘‘Prediction of cardiovascular disease using machine learning algorithms,’’ in Proc. Int. Conf. Current Trends Towards Converging Technol. (ICCTCT), Coimbatore, India, Mar. 2018, pp. 1–7.
26. S. Sharma and M. Parmar, ‘‘Heart diseases prediction using deep learning neural network model,’’ Int. J. Innov. Technol. Exploring Eng., vol. 9, no. 3, pp. 1–5, Jan. 2020.
27. A. K. Dwivedi, ‘‘Evaluate the performance of different machine learning techniques for prediction of heart disease using ten-fold crossvalidation,’’ Neural Comput. Appl., vol. 29, pp. 685–693, Sep. 2016.
28. I. K. A. Enriko, ‘‘Comparative study of heart disease diagnosis using top ten data mining classification algorithms,’’ in Proc. 5th Int. Conf. Frontiers Educ. Technol., 2019, pp. 159-164.
29. A. Kaur, ‘‘A comprehensive approach to predict heart diseases using data mining,’’ Int. J. Innov. Eng. Technol., vol. 8, no. 2, pp. 1–5, Apr. 2017.
30. A. A. Shetty and C. Naik, ‘‘Different data mining approaches for predicting heart disease,’’ Int. J. Innov. Sci. Eng. Technol., vol. 5, pp. 277–281, May 2016.
31. C. B. C. Latha and S. C. Jeeva, ‘‘Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques,’’ Informat. Med. Unlocked, vol. 16, no. 2, 2019, Art. no. 100203.
32. C. A. Cheng and H. W. Chiu, ‘‘An artificial neural network model for the evaluation of carotid artery stenting prognosis using a national-wide database,’’ in Proc. 39th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC), Jul. 2017, pp. 2566–2569.
33. R. Bhuvaneeswari, P. Sudhakar, and G. Prabakaran, ‘‘Heart disease prediction model based on gradient boosting tree (GBT) classification algorithm,’’ Int. J. Recent Technol. Eng., vol. 8, no. 2, pp. 41–51, Sep. 2019.
34. R. Alizadehsani, J. Habibi, Z. A. Sani, H. Mashayekhi, R. Boghrati, A. Ghandeharioun, F. Khozeimeh, and F. Alizadeh-Sani, ‘‘Diagnosing coronary artery disease via data mining algorithms by considering laboratory and echocardiography features,’’ Res. Cardiovascular Med., vol. 2, no. 3, pp. 133–139, Aug. 2013.
35. V. Chaurasia and S. Pal, ‘‘Data mining approach to detect heart diseases,’’ Int. J. Adv. Comput. Sci. Inf. Technol., vol. 2, no. 4, pp. 56–66, 2014.
36. S. Mohan, C. Thirumalai, and G. Srivastava, ‘‘Effective heart disease prediction using hybrid machine learning techniques,’’ IEEE Access, vol. 7, pp. 81542–81554, 2019.
37. K. C. Tan, E. J. Teoh, Q. Yu, and K. C. Goh, ‘‘A hybrid evolutionary algorithm for attribute selection in data mining,’’ Expert Syst. Appl., vol. 36, no. 4, pp. 8616–8630, May 2009.
38. Dangare, C.S. and Apte, S.S. (2012), “Improved Study of Heart Disease Prediction System Using Data Mining Classification Techniques”, International Journal of Computer Applications, 47, pp. 44-48.
39. Nagarajan, G., and R. I. Minu. "Fuzzy ontology based multi-modal semantic information retrieval." Procedia Computer Science 48 (2015): 101-106.
40. Dhanalakshmi, A., and G. Nagarajan. "Convolutional Neural Network-based deblocking filter for SHVC in H. 265." Signal, Image and Video Processing 14 (2020): 1635-1645..
41. Nagarajan, G., and K. K. Thyagharajan. "A machine learning technique for semantic search engine." Procedia engineering 38 (2012): 2164-2171.
42. Indra, Minu Rajasekaran, Nagarajan Govindan, Ravi Kumar Divakarla Naga Satya, and Sundarsingh Jebaseelan Somasundram David Thanasingh. "Fuzzy rule based ontology reasoning." Journal of Ambient Intelligence and Humanized Computing (2020): 1-7.
43. Nagarajan, G., R. I. Minu, and A. Jayanthila Devi. "Optimal nonparametric bayesian model-based multimodal BoVW creation using multilayer pLSA." Circuits, Systems, and Signal Processing 39, no. 2 (2020): 1123-1132.