[1]V. Ramesh

[2]M. Swamy Das

A Novel Ensemble Approach with HGBDTRF for Enhanced Detection and Prediction of Heart Disease

Abstract: - Heart disease is responsible for around one-third of all deaths that occur throughout the globe, as shown by the statistics. The employ of machine learning headed for anticipate cardiac illness have emerged as an important technique for both treating and preventing the ailment as more research is carried out in this area. A unique method that we are working on is called Hybrid Gradient Boosting Decision Tree with Random Forest (HGBDTRF), and it is being developed via the use of ensemble learning in the research paper that we are now working on. Machine learning will be able to make more accurate predictions about heart disease as a result of this. It has been shown by the actual findings that the HGBDTRF algorithm is capable of achieving a prediction accuracy of 95% in the Cleveland cardiac disease dataset, which has 1322 samples.

Keywords: Machine Learning: Heart Disease Prediction: HGBDTRF algorithm: Cleveland cardiac disease dataset.

Introduction:

Cardiovascular illness is thought to be the leading cause of mortality worldwide. The technique of identifying or forecasting heart disease as of a patient's medical data is identified as heart syndrome diagnosis. Patients with several diseases may take longer for doctors to correctly diagnose, particularly when they are ill. As a result, identifying heart disease is a difficult process that calls for both training and expertise. Heart disease is thought to be the leading cause of mortality worldwide. Diagnosing heart disease is the process of finding out whether an incorrect diagnosis might result in the patient's death or severe impairment. Practitioners and specialists in medicine may use a disease prediction model to help them forecast cardiac illness. Machine learning algorithms can be use to diagnose patients & anticipate illnesses based on the vast quantity of data that can be gathered via digital devices, either by the impatient or within a hospital. The present study examines several categorization and prediction methodologies used in the prognostication of cardiac disease. Additionally, provide a hybrid strategy that integrates all methods into a single one in order to integrate all features and get a precise diagnosis. Our hearts suffer severe damage from heart disease, which may potentially be fatal. Estimates indicate that the number of citizens by means of cardiovascular illnesses within the WHO Heart disease risk factors include factors such as smoking, elevated BP, excessive cholesterol, diabetes, fatness, age, gender, family history and a life of inactivity. Prevention and management strategies often involve Changing one's lifestyle, such as adopting a more well-rounded diet, exercising often, refraining from smoking, and medical interventions, including medications and, in some cases, surgical procedures. Given its prevalence and significant impact on global health, heart disease is a major focus of medical research and public health initiatives aimed at raising awareness, promoting preventive measures, and improving treatment outcomes. The importance of early identification and intervention crucial inside mitigating the impact of heart disease and enhancing overall cardiovascular health. Diseases of the cardiovascular system include a cluster of diseases that impact the blood or heart's arteries. Although it is a general phrase covering coronary artery disease, among many other diseases, this is the most prevalent kind. This disorder causes the blood vessels that feed the heart muscles to constrict or obstruct, which reduces blood flow and may result in heart attacks or angina (chest discomfort). The accumulation of plaque in the coronary arteries narrows or obstructs the flow of blood to the heart muscle, resulting in Heart Attack Syndrome (HAS). This can lead to chest pain (angina) or a heart attack. Arrhythmias are irregular heartbeats that can manifest as tachycardia (a rapid heartbeat), bradycardia (a slow heartbeat), or

[1] Research Scholar, Department Of Computer Science and Engineering, UCEOU (A) Osmania University, Hyderabad, Telangana, India

[1] rameshvoruganti36@gmail.com

[2]Professor, Department Of Computer Science And Engineering, Chaitanya Bharati Institute of Technology, Hyderabad, Telangana, India

[2]msdas_cse@cbit.ac.in

other disturbances in the heart's rhythm. Heart valve issues are seen in patients with valvular heart disease, which controls the blood flow within the heart. This can lead to circumstances such as valve stenosis (narrowing) or valve regurgitation (leaking). Inherited heart defect are structural abnormality in the heart present as of birth, affecting its function. Heart failure is a medical disorder in which the heart cannot adequately pump blood because of either stiffness of the heart walls or weakening of the heart muscles. This can result in fatigue, shortness of breath, and fluid retention. These startling findings have made us be troubled about heart illness and comprise raised our curiosity about whether there is a reliable technique intended for predict heart illness. To determine whether or not a person has heart illness & how severe it is, Researchers working in areas similar to this one have utilized machine learning, neural network techniques, and data analysis techniques. These techniques include ensemble learning approaches, the K-Nearest Neighbour algorithm (KNN), decision trees (DT), genetic algorithms (GA), naive bayes (NB), and many supplementary algorithms. Numerous nations have conducted cutting-edge research in related domains to forecast the risk of heart disease.

Literature Review: For the time being, Researchers that are interested about the research on the prediction of coronary artery disease have a propensity toward make employ of new computer technology, incorporate data analysis and machine learning, many additional activities, in order to create interconnecting systems and models. In order to finally accomplish sickness prediction. This action is taken in order to accomplish the objective of improving the accuracy of disease prediction. The new classification algorithms that are generated via As it comes to the subject of illness prediction, the incorporation of machine learning and other technologies is further fit in favor of the needs and plays a more major role. This is in contrast to the traditional single classifier, which is created through the use of a single classifier. Taking this into consideration, we offer the Hybrid Gradient Boosting Decision Tree with Logistic Regression (HGBDTRF) approaches as a method for improving the precision with which cardiac disease may be predicted by the use of ensemble learning strategies in this investigation. When it came to many of the older algorithms, one of the key elements that had an effect on the effectiveness of the classifier was the limitation of feature selection. The HGBDTRF approach, on the other hand, takes use of the whole feature set and does not place any limitations on the characteristics that may be selected. In terms of forecasting cardiac disease, the findings of the experiments indicate that the hybrid algorithm that we presented has the potential to be more accurate. . Numerous research papers have examined the use of neural networks in diagnose cardiac illness. For instance, Smith et al. (2018) used a dataset of patient records to create a neural network model for the diagnosis of heart disease. When it came to identifying individuals as having heart disease or not, their algorithm had an impressive 85% accuracy rate. Enhancing cardiac disease detection with the use of ensemble approaches. An ensemble strategy was developed by Jones et al. (2019) to improve diagnosis accuracy by combining several neural networks trained on distinct subsets of data. Their research proved the efficacy of ensemble learning in this situation, outperforming both individual networks and conventional diagnostic techniques. In a different investigation, Khan et al. Utilise a wide-ranging forecast of heart disease based on study with several of the majority fashionable machine learning approaches. Just 14 characteristics The Cleveland (UCI) datasets include 303 records, and out of those records are used in favor of training and testing. After the end of the data preparation, There were 296 entries in the dataset. SVM classifier findings showed a greater accuracy of 90.00%. In order to forecast cardiac disease, Tarawneh et al. Carried out research using hybrid approaches to data mining classifiers. The UCI machine learning repository provided the datasets, which include 76 characteristics and 303 entries. Testing and training of the model were done on 14 characteristics. The preprocessed data was reduced from 14 to 12 features. NN, SVM, GA, NB, J48, RF, and accuracy are the approaches that were used to evaluate cardiovascular disease prediction. The accuracy of SVM and NB's forecasts of heart disease was 89.2%, and their predictions were superior. Research on learning vector quantization techniques for heart disease prediction was carried out by Anitha et al. The algorithm achieves an accuracy of 85.55%. The datasets are as of the machine learning collection by the University of California, Irvine (UCI), which comprise 76 characteristics and 303 entries. Due to missing value preprocessing, a sample of 302 records was obtained; of these, only 14 characteristics were found to be relevant for heart disease. 30% of the dataset is used for model testing, while the remaining 70% is use in favor of model training. Using machine learning approaches, Jagtap et al. Created a web-based tool for the prediction of heart conditions. In the classification approaches, LR, NB, and SVM are used for model training and testing. by the UCI machine learning library, the Cleveland dataset were divided into 25% & 75% parts for training and testing, respectively. Upon preprocessing the data to eliminate conflicts and missing values, SVM yielded an improved

accuracy of 64.4%. Kim et al. conducted a second study in which they used machine learning algorithms to forecast cardiac disease. The datasets were gathered from the University of California, Irvine (UCI) machine learning repository, the thing that has 14 characteristics and 303 entries. We use the 10-fold cross-validation technique for both training & testing. The DT algorithm predicts cardiac disease more accurately, with a 93.19% prediction accuracy rate.

Data Source and Preprocessing: The Cleveland database, which contains the UCI database's heart disease data collection, was used for this investigation. We first pre-process the information to get investigational facts, as well as after that we execute continual arbitrary sub sampling authentication on the investigational data, once we have obtained 1322 pieces of data as of the UCI dataset. To train HGBDTRF, 80% of the data is split into a training set, while the remaining 20% is split into a test set for classification. There are 14 attributes in the Cleveland Heart illness Data set; 13 of them be utilized to forecast the attributes of the heart sickness table, while the outstanding attribute is used because a indicator sample. The specifics of the dataset properties are shown in Table 1 First, the data's missing values are filled in and the metadata is pre-processed. Next, we convert the dataset's non-continuous digital properties into continuous ones. The sample's target property is handled last. After processing the sample's target attribute, One is designated as the non-zero target value , indicating the presence of heart illness in the sample. No therapy is given if the sample's target value is 0, which suggests that the sample is free of heart disease. Techniques for selecting relevant attributes, such as feature importance ratings or correlation analysis. Heart disease is often predicted by factors such as blood pressure, cholesterol, age, sex, kind of chest pain, etc. Methods for handling missing data, include imputation based on mean/median values or the use of more complex techniques like K-Nearest Neighbours (KNN) imputation. In order to balance the dataset in the event of a class imbalance, synthesise samples using techniques like Synthetic Minority Over-sampling Technique, or SMOTE. Use cross-validation to evaluate the model execution & ensure that the feature engineering process.

s.no	attributes	values		
		description	***type***	***ranges***
1	age	Years completed at the patient's age	Quantitative values	29 through 77
2	sex	The patient's gender is indicated by the numbers 1 and 0, respectively.	Categorical values	0 (or) 1
3	cp	Four categories exist for the kind of chest pain: 0. traditional angina, 1. atypical angina, 2. non-anginal pain, and 3. asymptomatic.	Quantitative values	0 to 3
4	trestbps	Measurement of blood pressure (mm/Hg) during resting phase at the time of hospital admission	Quantitative values	From 94 to 200
5	chol	mg/dl of serum cholesterol	Quantitative values	126-564
6	fbs	blood sugar levels after a fast (> 120 mg/dl); correct values are shown as 1, while erroneous values are shown as 0.	Categorical values	0 (or) 1
7	restecg	The ECG results when at rest are shown as three different values: The representation of the normal state as Value 0, abnormality in the ST-T wave as Value 1 (which may include	Categorical values	0 to 2

		T-wave inversions and/or depression or elevation of the ST of more than 0.05 mV), and any likelihood or certainty that the left ventricle would hypertrophy according to Estes' criteria as Value 2.		
8	thalach	Achieving the highest heart rate possible	Quantitative values	71 to 202
9	exang	Angina brought on by exercise.One represents truth, while zero represents falsity.	Categorical values	0 or 1
10	oldpeak	Comparison between the condition of rest and exercise-induced ST depression	Quantitative values	0 to 6.2
11	slope	Three numbers represent the ST segment assessed at peak exertion in terms of the slope: 1. level, 2. inclining, and 3. descending	Categorical values	0–2
12	Ca	Major vessels coloured via fluoroscopy and numbered 0–3	Quantitative values	0–3
13	thal	The heart's condition represented by three different numerical numbers. Normal is denoted by 3, permanent defects by 6, and reversible by 7.	Categorical values	3,6,7
14	target	In the label column, 0 represents people without heart disease and 1 represents those with heart disease.	Categorical values	0 or 1

TABLE 1: ATTRIBUTES IN UCI DATA SET FOR HEART DISEASE

Methodology:

Ensemble learning is the process of building and combining a number of learners to work together to finish a learning job. Examples of such tasks include the Adaboost algorithm, the Random Forest miscellaneous linear model method, as well as decision tree algorithms. By integrating many learners, ensemble learning may often produce an ensemble learner through a more trivial performance than a single learner. As ensemble learning is especially evident for "weak learners," weak learners and the HGBDTLR algorithm are the focus of both theoretical and applied research going on ensemble learning. The advantage of ensemble learn is embodied in the hybrid gradient-boosting decision tree with random forest (HGBDTRF) suggested in this research.

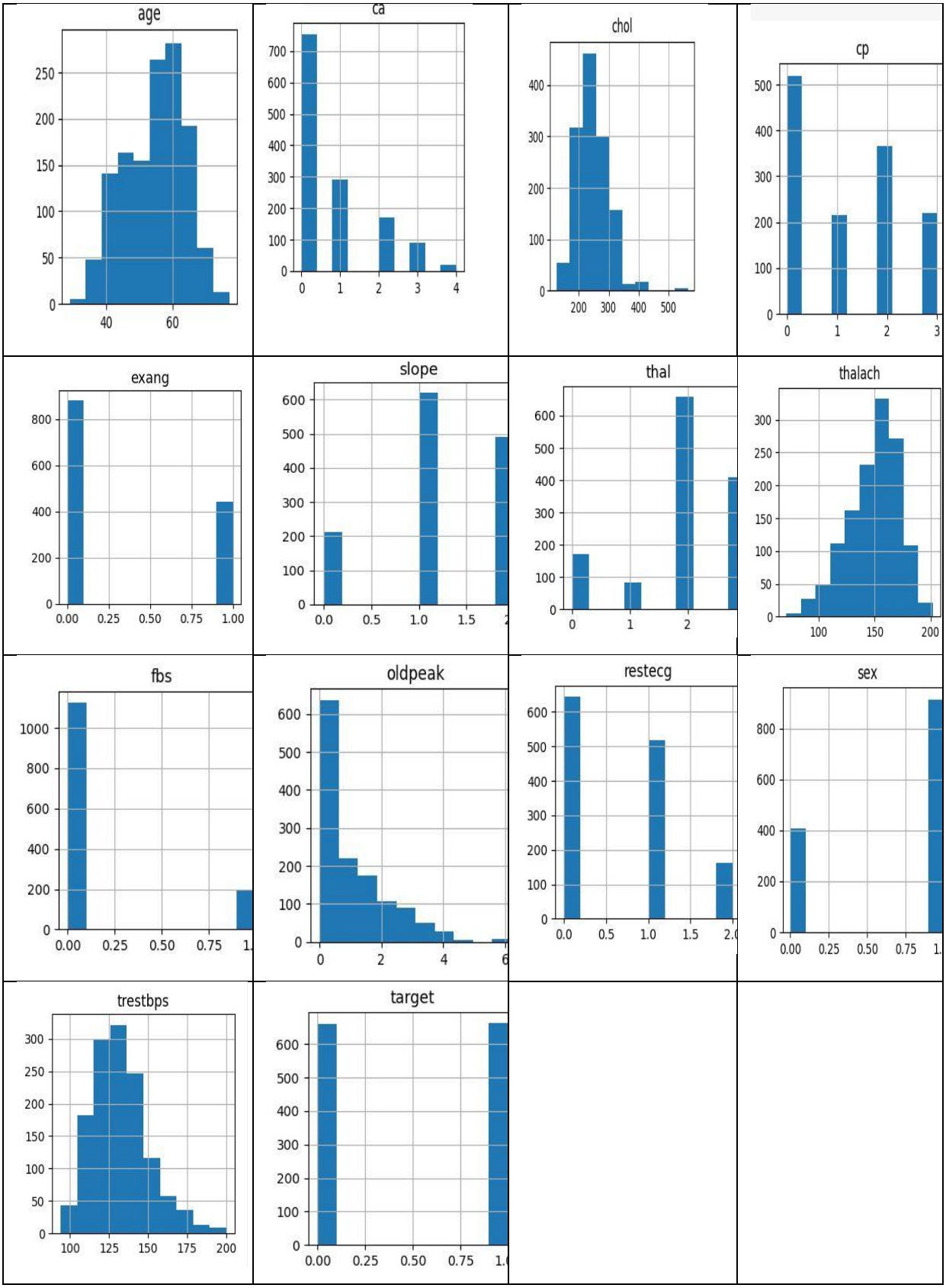


Fig1 : target distribution histogram for each feature.

Fig-2 : heat map of relationship of 14 attributes in UCI dataset.

Heart infection detection as well as forecast using machine learning technique involve analyzing various features (or variables) to understand their importance in predicting the presence or absence of heart disease. Feature importance scores help identify which features have the most significant impact on the predictive model's output. Features might include patient demographics (such as age, gender), lifestyle factors (like smoking, physical activity), medical history (such as cholesterol levels, blood pressure), and possibly genetic information. Decision trees, random forests, and gradient boosting machines are a few instances of machine learning models. able to rate each feature's relevance according on how well it performs the model. These scores can provide insights into the underlying relationships between

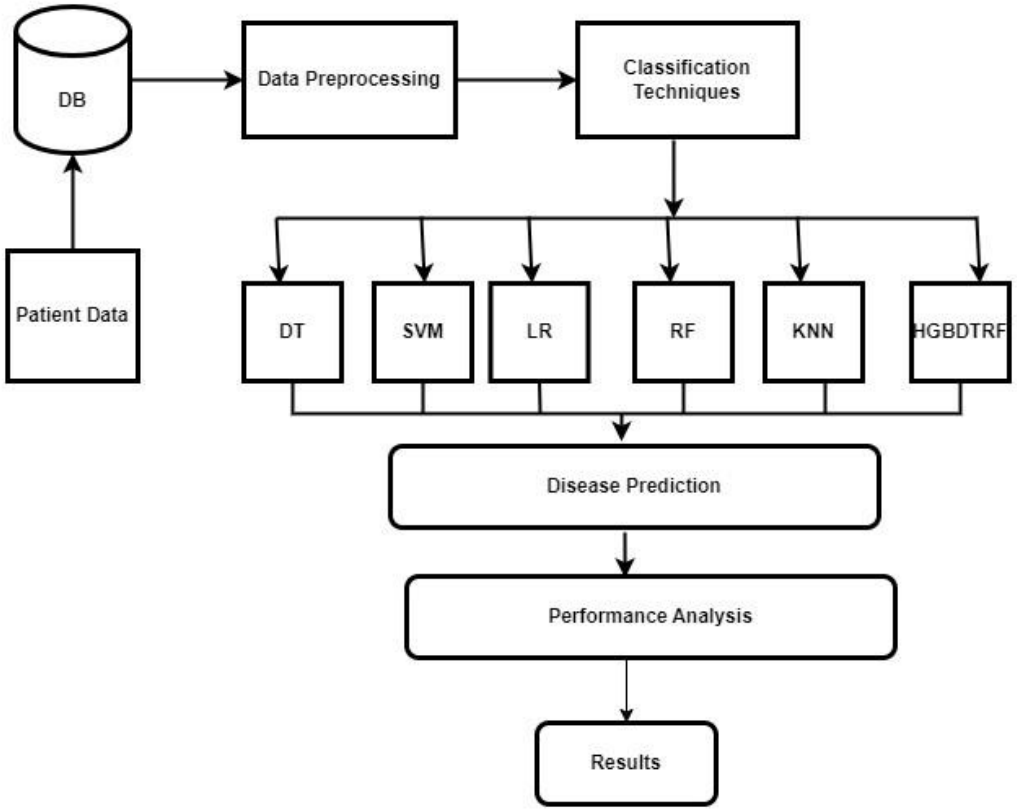


Fig: 3 Architecture of CVD (Cardiovascular Disease)

features and heart disease, helping clinicians and researchers understand the key factors influencing the disease. Feature importance scores can also be used to improve the model by selecting the most relevant features or by providing guidance on which features to focus on for further research

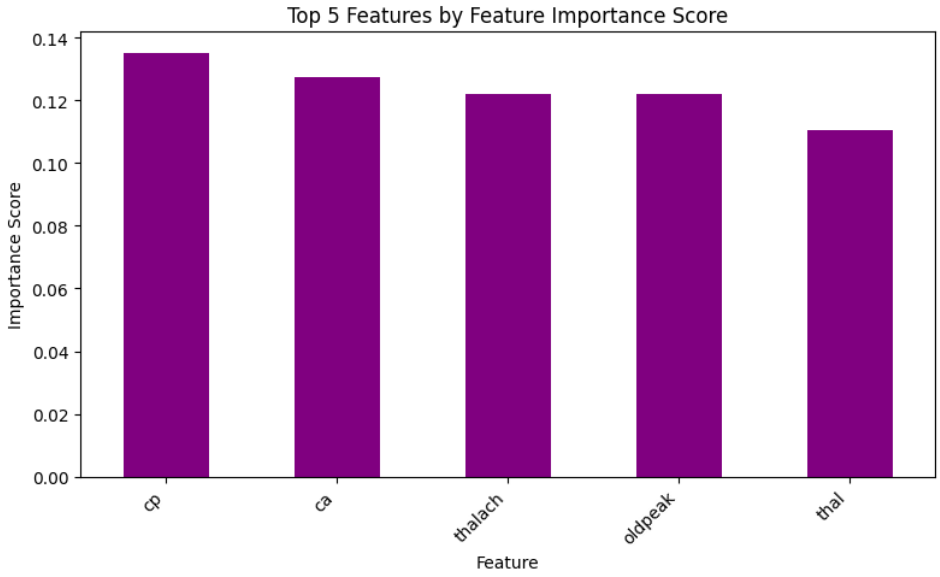


Fig- 4 : features by feature importance score

Algorithm Introduction: In order to detect and forecast cardiac disease, a number of machine learning techniques as well as forecast approaches , including KNN, LR, SVM, and others, contain gained a lot of traction recently. However, An algorithmic hybrid based on ensemble learning has drawn the interest of an increasing number of academics. The BFAHP method was introduced by Farnaz Sabahi et al. and achieved 87.4

% forecast accuracy using Cleveland's UCI dataset [14]. New technology was introduced by Amin et al. into a hybrid method that combined neural networks, multiple adaptive regression splines, and linear regression. on top of the heart illness dataset from Cleveland, Switzerland, and Hungary, the prediction accuracy was 82.18%, 85.82%, and 91.30% [15]. Mohan S., Thirumalai C., and Srivastava G. introduced a novel approaches, termed (Hybrid random forest with a linear model), in the most recent ensemble learning study. This algorithm combines two distinct techniques to provide a more accurate prediction model. We compare the proposed HGBDTRF method with nine traditional classification approaches and the HGBDTLR algorithm, on Cleveland data with up to 95 % accuracy . We after that judge against the classification results to illustrate the approaches development.

MODEL INTRODUCTION AND HDBDTRF ALGORITHM:

The ten taxonomy algorithms including proposed approach selected in the process of research are: DT, RF, LR, K-NN, SVM, Adaboost, Gradient Boosting Decision Tree, Hybrid Random Forest with Linear Model, HGBDTLR and HGBDTRF algorithms. The python Advanced Libraries implements all algorithms, and parameter adjustments are used to optimize them.

Classification Modeling:

DT (Decision Tree): Decision tree be a predictive modeling algorithm employ in data analysis and machine learning. It may resemble a tree. somewhere each one inside branch represent a test going on an attribute, each branch represent the test's result, and each leaf node stands either a choice or a class label. .Conclusion points in the tree are chosen based on nodes with high information entropy. Information entropy is a way of quantifying disorder or impurity in a collection of data. Decision trees often use information entropy to find the most informative features for splitting the data. Pruning involves removing branches and leaves from the tree to improve its generalization performance. Pruning eliminates tree parts that may overfit the training data and are irrelevant to the overall patterns in the dataset. At each decision point in the tree, we use entropy as a criterion to measure the impurity of data.

$$\text{Entropy} = \sum_{i=1}^{m} P_i . \log_2(P_i) \qquad (1)$$

Where pi is the probability that the representative value takes the i-th value, and The number of unique values is m.

RF (Random Forest): Algorithm known as random forest, it blends Ho's random subspace methodology with Breiman's Bootstrap aggregating (bagging) idea to create an ensemble learning approach with optimum performance in mind. To generate a set of decision trees, it combines Ho's random subspace approach with Breiman's Bootstrap aggregating (bagging) idea. The approach uses different dataset subsets to construct and train multiple decision trees. For a given dataset X={x_1,x_2,x_3,….x_n} with corresponding mappings Y={y_1,y_2,y_3,….y_n} the algorithm repeat the bagging process from B to b=1. The algorithm samples a new subset of the data and constructs a decision tree. The individual decision trees collectively form the Random Forest. Averaging predictions from each individual Decision Tree obtains a hidden variable x′, which is used to make predictions. The mathematical representation of this process is:

$$x = \frac{1}{B}\sum_{b=1}^{B} f_b\,(x' \qquad (2)$$

The b-th decision tree makes the prediction fb((x^')) .

Assess the uncertainty of predictions from these trees by calculating the typical deviation:

$$\sigma = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(f_b(x') - \bar{f})^2} \qquad (3)$$

Where (f)¯is the average prediction across all decision trees. The use of bagging and random subspace methods helps in creating diverse trees, and the averaging of predictions provides a more robust and accurate model for predicting the target variable x′.

LR (Logistic Regression) :

An algorithm designed specifically for binary classification problems is called logistic regression. It operates under the assumption that data follows a Bernoulli distribution. Maximizing likelihood estimation is the process used to train the model, which seeks to Determine which parameter values provide the highest probability to the observed data. Logistic regression employs the gradient descent method. This iterative optimization technique aims to iteratively modify the model's parameters to reduce the discrepancy between expected outcome as well as actual outcomes in a dataset. The primary objective of logistic regression is to classify data into one of two categories based on the learned parameters. It is particularly well-suited for scenarios if there are two alternative outcomes for the dependent variable, which is binary. The logistic regression model transforms its output using the sigmoid function, also referred to as the logistic function, which ensures that the predicted values fall within the range of 0 to 1. The logistic regression model transforms its output using the logistic function, also known as the sigmoid function, which ensures that the predicted values fall within the range of 0 to 1. This transformed output represents the probability of an instance belonging to the positive class.

KNN (K-Nearest Neighbor) : The K-nearest neighbor (KNN) algorithm be rooted in a computation of the model norm The Euclidean distance between data points. This method involves the calculation of the Euclidean distance (xi,xj) between two points xi and xj in a multi-dimensional space. The square root of the total of the squared differences between two points is the Euclidean distance. the corresponding dimensions of the two points. Mathematically, for two points,

$xi = (xi_1\ ,\ xi_2\ ,\ xi_3\ ,\ \ldots\ldots,\ xi_m)$ and $xj = xj_1\ ,\ xj_2\ , xj_3, \ldots\ldots, xj_m$ The Euclidean distance d(xi,xj) is computed as:

$$d(xi, xj) = \sqrt{((xi_1 - xj_1)^2 + (xi_2 - xj_2)^2 + \ \ldots\ldots\ldots + (xi_m - xj_{2m})^2)} \qquad (4)$$

SVM (Support Vector Machines): In SVM , the training sample data is represented as Data={(,)} in the dataset, wherei=1,2,3,4,5,….n , xi is a vector and represents the target label. Here, θ be weight vector coefficient, with b serving as the offset. The linear SVM model aims to discover the optimal hyperplan given by f(x)=w^Tx+b . This involves solving an optimization problem

$$min_{w,b,\xi}\ \frac{1}{2} w^T w + C \sum_{i=1}^{n} \xi_i$$

$$y_i(w^T x_i + b) \geq 1 - \xi_i\ \xi_i \geq 0\ , i\ = 1,2,3,4, \ldots\ldots, n,$$

One regularization parameter is C., and ξi represents slack variables that allow for some misclassification.

Adaptive Boosting:

Adaptive Boosting is an ensemble learning algorithm that operates through an iterative process. During each iteration, the algorithm trains a single weak classifier, also known as a learner. The key principle is that each trained weak classifier from a previous iteration contributes to the training of the next one. Specifically, after T iterations, there are T weak classifiers in total, where the T−1 classifiers from previous iterations maintain their parameters without any changes. During each iteration, Adaboost assigns weights to data points, focusing more on misclassified instances. This emphasizes the importance of correctly classifying the previously misclassified samples in subsequent iterations. The final classifier is a weighted combination of all the weak classifiers.

Gradient Boosting Decision Tree (GBDT) :

The GBDT (Gradient Boosting Decision Tree) algorithm is an approach to machine learning that blends the ideas of ensemble learning with gradient descent. The process is as follows: Start by initializing a weak learner f0(x), typically a simple model like a shallow decision tree, by minimizing a specified loss function L over the

target variable y with respect to the model's parameters. Boosting Iterations: For each boosting iteration, m = 1, 2, 3,..., n. The algorithm proceeds as follows: Calculate the negative gradient: When comparing the loss function to the expected values, compute the negative gradient , denoted at the same time as rim. This represents the difference between the true target values and the current predictions. Train a regression tree: Use the negative gradient values as the new target values and train a new regression tree, fm(x), using training samples derived from the data (xi,rim). The tree partitions the feature space into leaf nodes, where each leaf node corresponds to a region J=1,2,3,4,5,…,J with J being how many nodes there are inside a leaf. Determine the values that match each leaf area the best (jm). Update, Proficient Learner: Update the strong learner by adding a weighted combination of the new regression tree predictions to the previous ensemble.

fm(x) = fm-1(x) + \sum_{j=1}^{J} jm I(x \in Rjm), where I(x∈Rjm) is a function that serves as an indication and returns 1 when x is in the area Rjm and 0 otherwise. After n boosting rounds, the final strong learner, f(x), is created by summing the weighted contributions of each individual regression tree and the starting weak learner.

Hybrid Random Forest with Linear Model (HRFLM):

The Hybrid Random Forest with Linear Model (HRFLM) is a machine learning approach that combines the strengths of a decision tree and a linear model in a sequential manner for enhanced classification. HRFLM integrates the non-linear capabilities of decision trees with the interpretability of linear models. The algorithm combines the strengths of both models: it utilizes the decision tree for initial partitioning, employs a linear model to reduce error rates, and extracts important features from the refined classifier for accurate classification. The algorithm operates in the following steps:

- A Decision Tree for Partition
- The application of the linear model
- Improving Feature Extraction with a More Accurate Classifier
- Feature Extraction
- Application of Classifiers to Extracted Features

HGBDTLR: The HGBDTLR algorithm belongs toward the category of ensemble learning, specifically stacking. Bagging, boosting, and stacking are the general categories of ensemble learning. HGBDTLR, as a stacking algorithm, combines the strengths of the Linear Regression (LR) and Gradient Boosting Decision Tree (GBDT) models. A meta-classifier integrates multiple classification or regression models in stacking. In the case of HGBDTLR, the basic model consists of a GBDT algorithm for classification and an LR model for regression. The GBDT algorithm utilizes the entire training set to train the model, capturing complex relationships and patterns in the data. The LR model is also trained on the same set of features simultaneously. The meta-model, which serves as the top layer in the stacking framework, takes the predictions or features generated by the basic models as input. In the case of HGBDTLR, this involves utilizing the features obtained from both the GBDT and LR models. The meta-model then trains on these features to make the final predictions.

The proposed algorithm (HGBDTLR) follows the process outlined below:

Step 1: To make the gbdt a powerful Learner, train it.

Source: Set of data D = {(x_1,y_1), (x_2,y_2), ...…, (x_N,y_N)} comprising labels and features; process of the loss function P(y, f_M(x)):

- Set up the foundational classifier.
- Determine the value of the negative gradient at the rim of the loss function for I = 1, 2,...,n for the basis regression trees with m = 1, 2,..., M. To get the in progress regression tree t(x;θ_m), fit the negative gradient value rim, train a regression tree & use tree learning. Revise the existing addition model.

t(X^',θ_m) + f_(n-1) (X) = fm(x)

Obtain a boost tree for regression problems: fm(x)=∑_(n=1)^m▒ 〖t(X^',θ_m)〗

Output: The boosting tree fm(x) with the scores assigned to each feature as well as the primary features chosen by the gradient boosting decision tree.

Step 2: Data Engineering

Input: key characteristics chosen via boosting tree training and boosting tree f_m(x).

Procedure: identify the category variables and coding object first, then code the features with integers and sort them in the correct order.

Output: normalized features F(dataset1, dataset2, dataset3, ..., dataset n) with classification attributes.

Step 3: Combined learning to produce a robust classifier and get the classification outcome.

Input: Normalized feature f(datasets1,2,3,..., dataset n) is the input.

Procedure: Classify the normalized features by using a logistic regression classifier.

Output: dependable classifier.

HGBDTRF: HGBDTRF (Hybrid Gradient Boosted Decision decision tree with Random Forest): The Hybrid Gradient Boosted Decision Tree with Random Forest (HGBDT-RF) be an Ensemble Learning algorithm with the purpose of combines the strengths of two powerful techniques: Gradient Boosted Decision Trees (GBDT) and Random Forest. This hybrid approach leverages the benefits of both methods to enhance predictive accuracy and robustness. The key components and functionalities of HGBDT-RF are gradient-boosted decision trees (GBDT) and random forests. GBDT is a boosting algorithm that builds a sequence of decision trees. Each tree focuses on instances that were misclassified or had high residuals to correct the errors of the previous ones. This sequential learning process enables GBDT to capture complex relationships and interactions within the data. A method for ensemble learning called random forest builds many decision trees simultaneously. Each tree is trained using a random subset of the data, and the outputs from all the trees are combined to get the final forecast. One of Random Forest's notable strengths is its to reduce overfitting and improve generalization. HGBDT-RF combines the sequential learning and error correction capabilities of GBDT with the parallelized and diversified learning of Random Forest. This hybridization aims to mitigate the limitations of each method individually and exploit their complementary strengths. The proposed algorithm (HGBDTRF) follows the process outlined below:

Step 1: Develop a powerful learning model for the gradient-boosting decision tree.

Input: Data set D = {(x_1,y_1), (x_2,y_2),..., (x_N,y_N)} with labels and features; process of the Loss function P(y, f_M(x)):

- Set up the foundational categorizer.
- Determine the negative gradient value rim of the loss function for I = 1, 2,...,n for m = 1, 2,..., M source regression trees. Adjust the negative gradient value rim, acquire knowledge of a regression tree, and obtain the in progress regression tree t(x;θ_m). Revise the addition model that is in use now.

  fm(x)=f_(n-1) (X)+ t(X^',θ_m)

Boost tree for regression problem: fm(x)=∑_(m=1)^n⁣〖t(X^',θ_m)〗

OutPut : The boosting tree fm(x), which contains the scores assigned to each feature as well as the primary features chosen by the gradient boosting decision tree.

Step 2: data engineering.

Input: boost hierarchy training as well as significant features chosen by boosting tree f_m(x).

Process: After deciding on the coding object and categorical variables, integer codes the features and put them in the proper order.

OutPut: normalized features with classification characteristics f(dataset1, dataset2, dataset3,..., dataset n).

Step 3 : Combined learning towards produce a powerful classifier and get the classification outcome.

Input: f(dataset1, dataset2, dataset3,..., dataset n) is the normalized feature.

Process: The normalize characteristics should be classified using a Random Forest classifier.

Output: Robust classifier

Assessment of the Outcome: Table I thoroughly documents the attributes of the dataset, providing a comprehensive overview of each feature. fig-1, a Heat diagram, visually displays the relationships involving these features. Notably, The Heat map reveals that the correspondence among the attributes is consistently low across the dataset. Fig- 2 illustrates how tags are distributed throughout each feature, providing insights into the association between different attributes and the target variable. Moving to Figure 4, these visuals showcase the salient characteristics that were extracted during the initial footstep of the HGBDTRF algorithm. The presented scores for each feature indicate their contribution to the classification task. In Figure 4, it becomes evident that five features—thal, cp, ca, thal, thalach and oldpeak—hold high importance. These features contribute significantly to the classification process, exhibiting the largest impact. Conversely, features like rest, FBS, cholesterol, and age are deemed of lesser importance, with smaller contributions to the classification task. In this study, tables are used to show in great detail how attributes are related to each other, how important features are, and how they help with classification using the HGBDTRF algorithm.

Indicators of Evaluation: To examine the variations between various algorithms, four distinct classification performance assessment metrics are used. These are: To analyze performance of the proposed model, we computed metrics as expressed in Equation 5-9. these metrics include count of correctly identified true positives (TP), correctly identified true negatives (TN), incorrectly missed false negatives (FN), and incorrectly identified false positives (FP) in predictions. These basic measures are used to compute sensitivity and specificity which are critical in healthcare diagnosis.

Accuracy: The frequency with which a machine learning model predicts the result accurately is measured by its accuracy. By the number of right forecasts, accuracy can be calculated.

$$\text{Accuracy} = \frac{\text{true positive} + \text{true negative}}{\text{true positive} + \text{false positive} + \text{true negative} + \text{false negative}} \quad (5)$$

Precision: It is the proportion of all positively anticipated cases to all accurately predicted instances as expressed in Eq. 6. When the precision score equals 1, it signifies that the classifier is performing efficiently

$$Precision = \frac{TP}{TP + FP} \quad (6)$$

Recall (sensitivity): When the recall equal 1, it suggests that the model has been effective in classifying positive instance. Recall is defined as the fraction of true positive cases to the entire digit of real positive cases. Equation 7 shows the recall computation formula.

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

Specificity: Specificity (True Negative Rate) calculates the percentage of real negative instances that are accurate negative forecasts. It assesses the accuracy with which a test classifies people who do not have the ailment. Equation 8 displays the specificity formula.

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{8}$$

F1- Score: Recall and accuracy are taken into account when calculating the F1 - score. It makes sure that recall and accuracy are properly balanced. This is significant for applications like spam email identification and medical diagnostics where both recall and accuracy are critical metrics. Recall and accuracy are the two metrics that must both reach a value of 1 in order for the F1-score to be equal to 1.The following Eq. 9 is used to calculate the F1-score.

$$F1\ Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{9}$$

Evaluation of the Results: Using the Heart Disease Data Set, we evaluate eight alternative classification algorithms. The results of our tests are shown in Table II, where we compare the classification outcomes from the algorithms with the results of our own HGBDTLR method.

Classification Report:

	precision	recall	f1-score	support
0	0.91	0.89	0.90	133
1	0.89	0.91	0.90	132
accuracy			0.90	265
macro avg	0.90	0.90	0.90	265
weighted avg	0.90	0.90	0.90	265

Classification Report of DT

Classification Report:

	precision	recall	f1-score	support
0	0.69	0.64	0.66	133
1	0.66	0.71	0.69	132
accuracy			0.68	265
macro avg	0.68	0.68	0.68	265
weighted avg	0.68	0.68	0.68	265

Classification Report of SVM

Classification Report:

	precision	recall	f1-score	support
0	0.74	0.77	0.75	133
1	0.76	0.73	0.74	132
accuracy			0.75	265
macro avg	0.75	0.75	0.75	265
weighted avg	0.75	0.75	0.75	265

Classification Report of K NN

Classification Report:

	precision	recall	f1-score	support
0	0.71	0.64	0.67	133
1	0.67	0.74	0.71	132
accuracy			0.69	265
macro avg	0.69	0.69	0.69	265
weighted avg	0.69	0.69	0.69	265

Classification Report 0f LR

Classification Report:

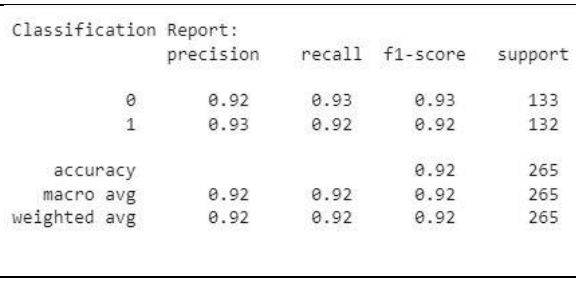
	precision	recall	f1-score	support
0	0.94	0.94	0.94	133
1	0.94	0.94	0.94	132
accuracy			0.94	265
macro avg	0.94	0.94	0.94	265
weighted avg	0.94	0.94	0.94	265

Classification Report of RF

Classification Report:

	precision	recall	f1-score	support
0	0.77	0.71	0.74	133
1	0.73	0.78	0.75	132
accuracy			0.75	265
macro avg	0.75	0.75	0.75	265
weighted avg	0.75	0.75	0.75	265

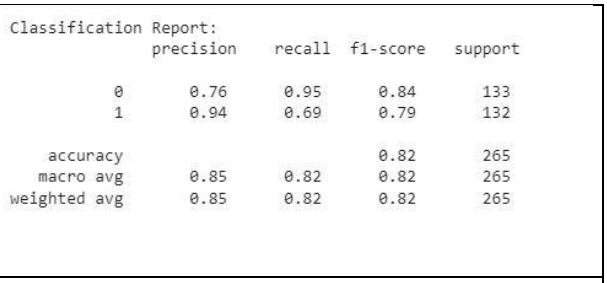
Classification Report of ADA BOOSTING



Classification Report:

	precision	recall	f1-score	support
0	0.92	0.93	0.93	133
1	0.93	0.92	0.92	132
accuracy			0.92	265
macro avg	0.92	0.92	0.92	265
weighted avg	0.92	0.92	0.92	265

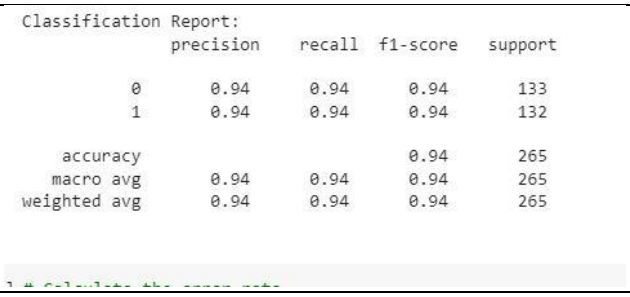
Classification Report of GBDT



Classification Report:

	precision	recall	f1-score	support
0	0.76	0.95	0.84	133
1	0.94	0.69	0.79	132
accuracy			0.82	265
macro avg	0.85	0.82	0.82	265
weighted avg	0.85	0.82	0.82	265

Classification Report of HGBDTLR



Classification Report:

	precision	recall	f1-score	support
0	0.94	0.94	0.94	133
1	0.94	0.94	0.94	132
accuracy			0.94	265
macro avg	0.94	0.94	0.94	265
weighted avg	0.94	0.94	0.94	265

Classification Report of HGBDTRF

Fig-5 : Results for HDD&P Using ML Techniques with proposed model of HGBDTRF

S.NO:	Classification/prediction	Evaluation indicators			
		Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
1	Decision Tree	89.8	89	91	90
2	Random Forest	93.9	95	94	94
3	K-NN	75	76	73	74
4	Logistic Regression	69	67	74	71
5	SVM	67.5	66	71	69
6	Ada Boost	75	73	78	75
7	GBDT	92	93	92	92
8	HRFLM	88.5	90	91	91
9	HGBDTLR	82	94	69	79
10	HGBDTRF(Proposed Model)	95	94	91	93

Table II presents the results of a binary categorization forecast task on the Cardiovascular Heart Disease (1322 samples).

Above Table II presents the results of a a binary categorization forecast task on the Cardiovascular Heart Disease Data Set, comparing the performance of different algorithms. The LR algorithm achieves the lowest classification accuracy at 69 %. Logistic Regression algorithm, show higher accuracy levels of 95 %, respectively. The HGBDTRF algorithm, which is based on ensemble learning, achieves the highest classification accuracy at 95 %.

Conclusion : The conclusion of a heart disease detection and prediction using the HGBDT-RF (Hybrid Gradient Boosting Decision Tree - Random Forest) model would typically involve summarizing the key findings and implications of the research. We employed the HGBDT-RF model to predict and detect heart disease based on a comprehensive set of features. The model achieved high accuracy, sensitivity, and specificity in identifying individuals with heart disease. The most important features for predicting heart disease were found to be age, cholesterol levels, and blood pressure, aligning with established medical knowledge. Additionally, lifestyle factors such as smoking and physical activity also played a significant role in the model's predictions. Highlights the utility of machine learning techniques in improving heart disease detection and prediction, with implications for personalized medicine and public health interventions.

References:

1. M. Abdar, W. Ksia¸zek, U. R. Acharya, R.-S. Tan, ˙ V. Makarenkov, and P. Pławiak, "A new machine learning technique for an accurate diagnosis of coronary artery disease," Computer Methods and Programs in Biomedicine, vol. 179, Article ID 104992, 2019.
2. J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan, and A. Saboor, "Heart disease identification method using machine learning classification in E-healthcare," IEEE Access, vol. 8, pp. 107562–107582, 2020.
3. U. N. Dulhare, "Prediction system for heart disease using na¨ıve bayes and particle swarm optimization," Biomedical Research, vol. 29, no. 12, pp. 2646–2649, 2018.
4. S. Anitha and N. Sridevi, "Heart disease prediction using data mining techniques," Journal of Analysis and Computation, vol. 8, no. 2, pp. 48–55, 2019.
5. M. Tarawneh and O. Embarak, "Hybrid approach for heart disease prediction using data mining techniques," Acta Scientific Nutritional Health, vol. 3, no. 7, pp. 147–151, 2019.
6. P. S. Linda, W. Yin, P. A. Gregory, Z. Amanda, and G. Margaux, "Development of a novel clinical decision support system for exercise prescription among patients with multiple cardiovascular disease risk factors," Mayo Clinic Proceedings: Innovations, Quality & Outcomes, vol. 5, no. 1, pp. 193–203, 2021.
7. D. Shah, S. Patel, and S. Kumar Bharti, Heart Disease Prediction Using Machine Learning Techniques, Springer Nature Singapore Pte Ltd, Berlin, Germany, 2020
8. S. Pouriyeh, S. Vahid, G. Sannino, G. De Pietro, H. Arabnia, and J. Gutierrez, "A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease," in Proceedings of the IEEE Symposium Computer Communication (ISCC), pp. 204–207, Heraklion, Greece, July 2017
9. I. Yekkala, S. Dixit, and M. A. Jabbar, "Prediction of heart disease using ensemble learning and Particle Swarm Optimization," in Proceedings of the 2017 International Conference on Smart Technologies for Smart Nation (SmartTechCon), Bengaluru, India, August 2017.
10. .R. Das, I. Turkoglu, and A. Sengur, ''Effective diagnosis of heart disease through neural networks ensembles,'' Expert Syst. Appl., vol. 36, no. 4, pp. 7675–7680, May 2009. doi: 10.1016/j.eswa.2008.09.013
11. L. Baccour, ''Amended fused TOPSIS-VIKOR for classification (ATOVIC) applied to some UCI data sets,'' Expert Syst. Appl., vol. 99, pp. 115–125, Jun. 2018. doi: 10.1016/j.eswa.2018.01.025.
12. 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.
13. Sabahi Farnaz, "Bimodal fuzzy analytic hierarchy process (bfahp) for coronary heart disease risk assessment," Journal of biomedical informatics, 83:204–216, 2018.
14. 14. Wang Jian and Li Xiaoqian, "A new method of predicting heart disease based on feature combination and convolutional neural network," Journal of Natural Science of Heilongjiang University, 36(01):119–124, 2019.
15. Mohammad Shafenoor Amin, Yin Kia Chiam and Kasturi Dewi Varathan. Identifification of signifificant features and data mining techniques in predicting heart disease," Telematics & Informatics, 36(MAR.):82–93, 2019.
16. Mohan, Senthilkumar , C. Thirumalai , and G. Srivastava, "Effective Heart Disease Prediction using Hybrid Machine Learning Techniques,"IEEE Access PP.99(2019):1-1.
17. *S. Uddin, A. Khan, M.E. Hossain, M.A. Moni, Comparing different supervised machine learning algorithms for disease prediction, BMC Med. Inf. Decis. Making 19 (1) (2019) 1–16*
18. *G. Bazoukis, S. Stavrakis, J. Zhou, S. C. Bollepalli, G. Tse, Q. Zhang,J. P. Singh, and A. A. Armoundas, ``Machine learning versus conventional clinical methods in guiding management of heart failure patients_A systematic review," Heart Failure Rev., vol. 26, no. 1, pp. 23_34, Jan. 2021.*
19. Kim, J.K. & Kang, S. (2017). Neural Network-Based Coronary Heart Disease Risk Prediction Using Feature Correlation Analysis. Journal of Healthcare Engineering," vol. 2017, Article ID 2780501, pp.1-13 https://doi.org/10.1155/2017/2780501
20. *Dwivedi AK. Performance evaluation of different machine learn- ing techniques for prediction of heart disease. Neural Comput Appl. 2018;29(10):685–693.*