**Educational Administration: Theory and Practice**
2024, 30(5) 14762 - 14775
ISSN: 2148-2403
https://kuey.net/

**Research Article**

# Performance Evaluation Of Ensemble Learning Using Light GBM For Enhanced Heart Disease Detection And Prediction

V. Ramesh[1*], M. Swamy Das[2]

[1]Research Scholar, Department Of Computer Science and Engineering, UCEOU (A) Osmania University, Hyderabad, Telangana, India
Email:-rameshvoruganti36@gmail.com
[2]Professor, Department Of Computer Science and Engineering, Chaitanya Bharati Institute of Technology, Hyderabad, Telangana, India
Email:-msdas_cse@cbit.ac.in

**Citation:** V. Ramesh, et.al. (2024), Performance Evaluation Of Ensemble Learning Using Light GBM For Enhanced Heart Disease Detection And Prediction , *Educational Administration: Theory and Practice*, *30(5)* 14762 - 14775
Doi: 10.53555/kuey.v30i5.7411

**ARTICLE INFO**

**ABSTRACT**

Diseases of the heart (CVD) include the primary source of rising death rates as well as major cause of fatality. Improving the predictability as well as accuracy of cardiac disease is the primary goal of constructing the suggested model. Experts who ignore patient complaints put the patient at danger of serious complications that might result in death or disability. Consequently, in order to find patterns and hidden information in the medical data related to heart disease, we require expert systems that act as analytical tools. Finding hidden underlying patterns in vast amounts of data is a cognitive process known as machine learning. This study uses ensemble learning approaches in an attempt to improve the preciseness of the risk of heart disease assessment. Additionally, this research project has included feature selection and hyper parameter tuning approaches, which have increased accuracy even further. Used the information on heart disease to assess its performance using several measures. Six machine learning classifiers, including SVM, LR, RF, DT, and Ensemble techniques, were applied to the final dataset for this purpose, both before and after the hyper parameter tuning of the classifiers. Additionally, by doing specific data pre-processing, dataset standardization, and hyper parameter tweaking, using the common heart disease dataset, we confirm their correctness. The K-fold cross-validation approach was use through the researchers. Lastly, the experimental findings showed that machine learning classifiers' accuracy of prediction increased with hyper parameter tweaking, and they produced noteworthy outcomes with data standardization, hyper parameter tuning, and Light GBM.

**Keywords:** Heart Disease, expert system, Light GBM and ensemble techniques

## 1. Introduction :

Cardiovascular disorders, including heart disease, are among the world's most deadly illnesses. The illness really arises from the heart's failure to pump enough blood to meet the needs of the body. Elevated blood pressure, obesity, smoking, alcohol misuse, sleep apnea, and elevated mental stress are the main risk factors for heart disease. As stated The World Health Organization estimates that heart disease accounts for around 24 percent of the deaths in India caused by non-communicable illnesses. Around the world, cardiovascular illnesses account for a third of all fatalities. Cardiovascular disease (CVD) is a leading cause of death globally, claiming the life of around 17 million people annually; Asia has the highest prevalence of CVD (WHO). And additionally Statistics from the World Health Organization (WHO) indicate that by 2030, heart failure and strokes would account for the majority of CVD-related deaths—more than 23.6 million deaths [13]. Numerous variables, such as stress, alcohol, smoking, poor food, sedentary lifestyle, and other linked health issues including high blood pressure, may contribute to CVD. On the other hand, once identified in their early stages, the majority of CVD-related illnesses are known to be fully treatable [14].It is very difficult to

identify and diagnose cardiac illness [1]. Heart disease may be automatically detected with the use of computer-aided detection, or CAD. The development of machine learning has made the analysis of health facts simple as well as accessible. The performance of machine learning approaches needs to be optimized [2]. By enhancing performance along with yielding improved as well as more perfect findings in the early diagnosis of cardiac disease, the ensemble learning technique offers the answer. Ensemble learning is a machine learning approach with the intention of use several classifiers toward improve system performance. Boosting, stacking, and bagging techniques are part of the ensemble learning framework. These are formed using distinct types of learners for stacking, but the same classifiers are used for bagging and boosting [3]. Relational techniques facilitate the extraction of hidden information and the identification of links between characteristics in a dataset, making them a viable approach for the categorization of CVDs. Providing patients with clinical services of the highest caliber at a reasonable cost is one of the biggest problems confronting health organizations. Accurate patient diagnosis and successful therapy identification are necessary for providing quality care, but incorrect diagnosis must be avoided. The model does far better when several classifiers are coupled than when it is used in isolation for classification. As a result, using ensemble learning increases the prediction accuracy for the identification of heart disease. Numerous researches made use of heart disease dataset. For testing with training the machine learning prediction models, the appropriate data is required. Using a consistent dataset in favor of training and testing machine learning categorization may increase their accuracy.

## 2. Literature Survey:

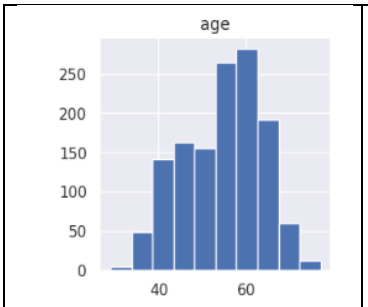
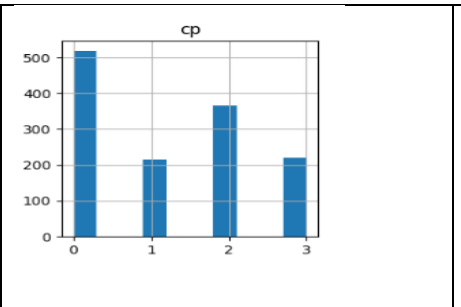
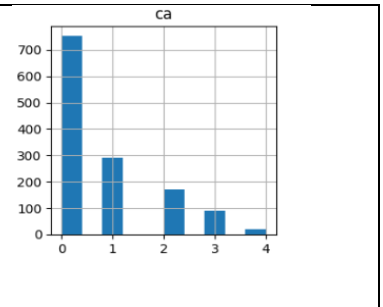
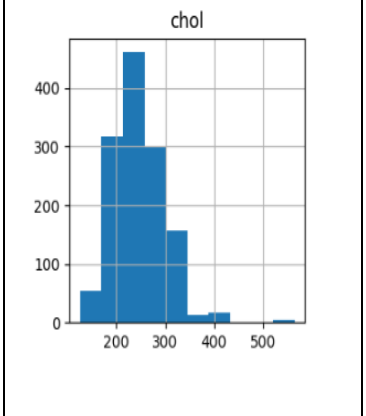
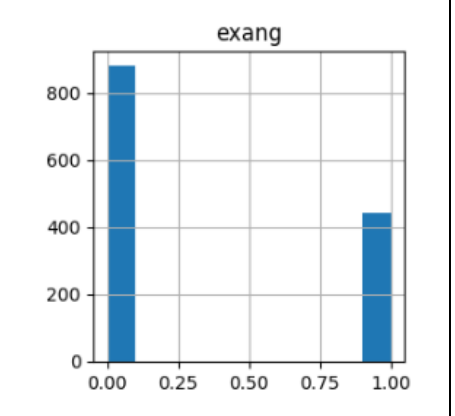
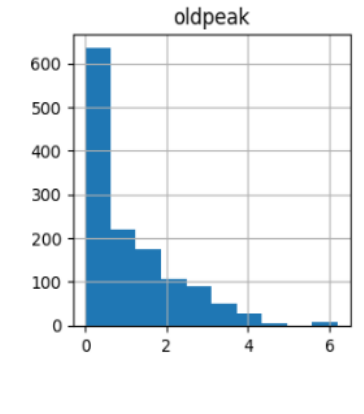
Machine learning may be used to a wide range of issues. It may To estimate the outcome of an experiment by adjusting the values of independent variables to those of the dependent one. Effective diagnostic test costs may be decreased by using computer technology to help in accurate, dependable, and skilled medical diagnosis. Because The healthcare business is a prime example of an application field for data mining due to its massive data resources, which are difficult to handle manually. Heart conditions especially within industrialized nation; have been shown to be individual of the foremost cause of death [7]. Individual of the main cause of heart disease-related deaths be with the aim of the dangers are either not recognized or discovered much afterward in life. The auscultation method was the main technique used by the doctors to differentiate between normal and pathological heart sounds [5]. Using stethoscopes to listen to these heart sounds, doctors were able to diagnose every cardiac condition [4]. There are several disadvantages to the auscultation approach that medical professionals utilize to identify cardiac disease. The ability and experience of physicians, which they acquire during extensive exams, is linked to the categorization of different heart sounds [6]. It has been shown that the model is a helpful resource for helping doctors forecast cardiac disease [9]. An extra feature selection step has been suggested in order to improve accuracy [8]. Several machine learning approaches enclose be presented intended for CVD detection in addition to the human approach.

Amin et al. [25] have the research to categorize the most important characteristics of heart disease prediction. There are seven classification algorithms in use: NB, KNN, LR, DT, NN, SVM, and Vote. The dataset, which include 303 entries and 76 characteristics, were taken from the machine learning library at UCI. Instruction (Training) and evaluation (testing) model is completed using the 10-fold CV (cross-validation) technique. Since there are fewer training instances in the dataset, we employed 10-fold cross-validation instead of data splitting techniques like train-test split, which would have resulted in an underestimate of the model prediction show due to the less quantity of the training set's examples. Nonetheless, Using 10-fold validation will provide the model with 90% of the data for learning. Vote Classifier's accuracy was greater, coming in at 87.4%. A KNN classifier with few parameters was used in a research by Ketut Agung Enriko et al. [26] to predict heart disease with a level of precision that is 81.15%.With the help of KNN, performance decreases the same as more parameter is used, 90% of the input is used for training, which be costly analytical. Subhadra et al. [27] carried out the investigation. A multilayer perceptron neural network (MLP-NN) with back propagation is the training technique used to forecast cardiac disease. F-measure, recall, precision, as well as accuracy are used to assess the system's performance. The Cleveland dataset, which comprises 76 characteristics and 303 records through UC Irvine's machine learning archive, is used for use in testing and training of models. The six records of data had missing values eliminated by pre-processing; the 14 mainly significant heart disease characteristics be then utilized. The experiment's findings showed that, throughout its 3.86-second runtime, MLN-NN achieved a higher accuracy of 93.39%. Another work by Khan et al. [28] included a thorough analysis of a some of the most implemented classifiers in machine learning to predict heart illness. A dataset of 296 records was obtained by doing data pre-processing, of which only 14 characteristics are used for training and testing. Massive datasets are being analyzed using machine learning to uncover hidden, important decision-making information for later examination. The medical industry has vast amounts of patient data. Different machine learning algorithms must mine this data. For the purpose of making an efficient diagnosis, healthcare experts analyze this data. Through analysis, medical data mining using categorization algorithms offers therapeutic assistance. It evaluates the Classification algorithms to forecast patients' risk of heart disease [16].In order to identify different heart disease variables and forecast heart illness, a variety of machine learning methods are

utilized, including regression, clustering, association rules, and classification algorithms including Naïve Bayes, decision trees, random forests, and K-nearest neighbor. We used data from the UCI repository for this study. In order to forecast cardiac disease, classification algorithms are used in the development of the classification model. This study compares the current methods and discusses algorithms used in the prediction of heart disease. The paper training and test dataset's potential for development and more study are also mentioned. While the testing dataset serves as fresh data to gauge the model's correctness, the training dataset sharpens our model.

## 3. Data Source:

We have utilized datasets from the UCI Machine Learning Repository in my research procedure. Including blood pressure, kind of chest pain, ECG result, serum cholesterol, fasting blood sugar, maximum heart rate, and other characteristics, it is an actual dataset consisting of 1322 instances of data with 14 different properties (13 predictors; 1 target). This research used six algorithms to identify the etiology of heart disease and generate the most accurate model possible. The tests have been performed by using the machine learning inventory at UCI Cleveland heart dataset. 1322 instances and 14 characteristics make up the combined dataset. There are six numerical characteristics and eight category attributes. Table 1 displays a description of the dataset. This dataset includes chosen patients with ages ranging between 29 and 79. Gender value 1 is use toward indicate male patients, whereas gender value 0 is used to indicate female patients. Here be four kinds of chest pain with the intention of be thought to be signs of heart disease. as of constricted coronary arteries, type 1 angina be brought on by decrease blood contribute to the heart muscles. Chest discomfort, often known as type 1 angina, is a symptom of emotional or mental stress. Chest discomfort other than angina may arise from a number of causes, not always related to heart disease. It's possible that the fourth category, asymptomatic, has no symptoms of cardiac illness. The resting blood pressure (RBP) measurement is the next assessment. The cholesterol level is represented by Chol , A person's fasting blood sugar level denoted like Fbs, and its value is 1 if less than 120 mg/dl and 0 otherwise. Angina that appears while physical activity is is recorded as Exang, with 1 indicating presence of pain and 0 indicating absence. A resting electrocardiographic result is abbreviated as "RESTECG.", Thalach stands for the maximum pulse. ST depression caused by exercise is denoted by oldpeak, and the slope of the peak exercise ST segment is represented by slope. The number of main vessels colored by fluoroscopy is indicated by ca. Thal represents the exercise test length in minutes. The class attribute, num, has a value of 1 for those without heart disease and 0 for those with heart disease.

| | | |
|---|---|---|
|  |  |  |
| **age** | **cp** | **ca** |
|  |  |  |
| **chol** | **exang** | **Old peak** |

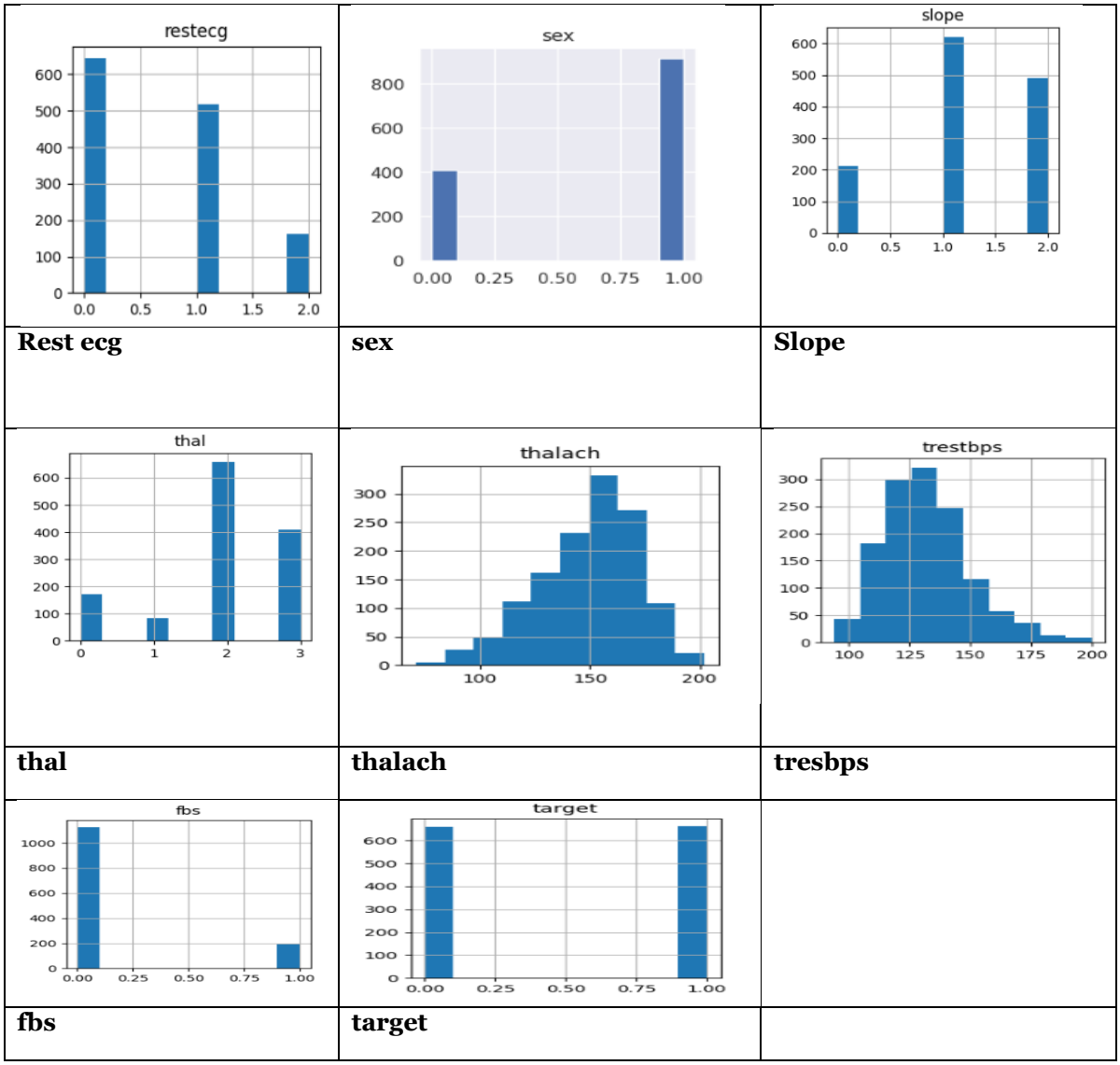



Fig 1: Histogram of each attribute of HDD&P using Ensemble Learning with LIGHT GBM

Table 1: Risk variables from databases on heart disease those are accessible to the general public.

| S.NO | DATASET | NO. OF SAMPLES | FEATURES |
|---|---|---|---|
| 1 | UCI Repository | 303 | 13 |
| 2 | UCI Repository | 1025 | 13 |
| 3 | Merged dataset | 1322 | 13 |

Table 2: Features and information from the heart disease dataset

| sr.no. | attribute | Representativeicon | information |
|---|---|---|---|
| 1 | Age | Age | Patientsage,inyears |
| 2 | Sex | Sex | 0=female;1=male |
| 3 | Chestpain | Cp | 4 typesof chestpain (1—typicalangina; 2— atypical angina;3— pain;4—asymptomatic) |
| 4 | Restbloodpressure | Trestbps | Restingsystolicbloodpressure(in mmHgonadmissiontothehospital) |
| 5 | Serumcholesterol | Chol | Serumcholesterolin mg/dl |
| 6 | Fastingbloodsugar | Fbs | Fastingbloodsugar>120mg/dl(0—false;1—true) |
| 7 | Restelectrocardiograph | Restecg | 0—normal;1—havingST-Twaveabnormality;2—leftventricularhypertrophy |
| 8 | MaxHeart rate | Thalch | Maximumheartrateachieved |
| 9 | Exercise-inducedangina | Exang | Exercise-inducedangina(0—no;1—yes) |
| 10 | STdepression | Oldpeak | STdepressioninducedbyexerciserelativetorest |

| | | | |
|---|---|---|---|
| 11 | Slope | Slope | slopeofthepeakexerciseSTsegment (1-upsloping;2-flat;3-downsloping) |
| 12 | No .of vessels | Ca | No.ofmajorvessels(0–3)coloredbyfluoroscopy |
| 13 | Thalassemia | Thal | Defecttypes;3—normal;6—fixeddefect;7—reversibledefect |
| 14 | Target (class attribute) | | target diagnosis of heartdisease status(0—absent; 1— present; 2— potential risk ; 3— high risk;4—veryhighrisk) |

### 3.1 Data Pre-processing:

Within artificial intelligence, machine learning is a developing field. Designing systems, letting them learn from experience, and enabling them to make predictions is its main goal. It uses a training dataset meant to demonstrate algorithms for machine learning and build a model. The heart disease prediction model makes advantage of the updated input data. So that we may build models, it extracts previously unseen patterns from the given dataset by use of machine learning. For fresh datasets, it generates precise forecasts. After cleaning the dataset, any missing values are filled in. The reliability of the model's cardiac illness prognosis is then evaluated using the updated input data. Standardized the datasets that were gathered. These datasets had inaccurate values and were not collected in a controlled setting. Pre-processing data is thus a crucial stage in the study of data and machine learning. Data normalization is the procedure of ensuring that a dataset's risk factors have distinct values. For instance, the temperature may be measured in multiple ways using Celsius and Fahrenheit. Scaling the risk variables and allocating numbers that illustrate the variation in standard deviations from the mean value are two methods of standardizing data. To enhance ML classifiers' efficiency while using a mean (μ) of 0 as well as a standard deviation (σ) is 1, it rescales the risk factor value (1) provides the standards in mathematical form.

$$\text{Standardization of } Z = \frac{Z - \text{Mean of } Z}{\text{Standard Deviation of } Z} \qquad (1)$$

Applying several machine learning methods to the dataset, such as XGBoost, AdA Boost, Logistic Regression, Naive Bayes, Support Vector Machines, and Random Forest.


Fig.2 .Heat map of machine learning approaches.

## 4. Methodology:

**4.1 Logistic regression (LR)**: Logistic regression (LR) is a powerful classifier within supervised machine learning algorithms [19]. It is a dataset-based extension of generic regression modeling, which indicates the likelihood of a given instance occurring or not [18]. Learning from experience (LR) determines the likelihood that a new observation would fall into a certain class; as a probability, the result falls between 0 and 1. In order to use the LR as binary categorization, a threshold is set to characterizes split into two categories. For example, if the probability assessment is greater than 0.5, it refer to as "class A"; otherwise, it is referred to as "class B." A multinomial logistic regression may be created by generalizing the LR model to create a

categorical variable with more than two values [19]. This research determined the best fit maximum iteration number of 100 and the best fit random state values for the applicable dataset.
Algorithm for Logistic Regression:
INPUT:
Step 1: samples (features)
PROCEDURE: Training data
Step 2: Training data labels (binary target classes)
Step 3: Learning rate
Step 4: Convergence threshold (for stopping criteria)
Step 5: maxi_iterations - Maximum number of iterations (optional)
OUTPUT:
Step 6: Trained weights *w* and bias*b*.

**Sigmoid Function**: Logistic Regression is a extensively worn supervise machine learning algorithm in favor of binary classification tasks, such as predicting Heart disease: whether it exists or not. At its core, logistic regression relies on an equation for the sigmoid, also known as the Logistic function, to model the connection between input characteristics and the likelihood of belonging to the positive class (in this case, having heart disease). The sigmoid function, denoted as σ(p), where The symbol p denotes a linear combination of the weights assigned to the input characteristics. Transforms the output into a number between zero and one. Mathematically, the sigmoid definition of function follows:

$$\sigma(p) = \frac{1}{(1 + e^{-p})} \qquad (2)$$

In this equation, e represents a natural logarithm's foundation, and For each set of characteristics and weights, p is the linear combination: p = w_0 + w_1 * x_1 + w_2 * x_2 + ... + w_n * x_n, where w_i denotes the weight associated with feature x_i.

**4.2 Naive Bayes:** The supervised method used is the Naïve Bayes classifier. It be a straightforward approach to classification base on top of the Bayes theorem. robust (NAIVE) independence within qualities is assumed. A mathematical notion to determine the probability is the Bayes theorem. There is neither a correlation nor a relationship between the predictors. Each characteristic contributes separately to the likelihood in order to optimize it. It does not use Bayesian techniques and may function by applying the Naïve Bayes model. Utilized here are Naive Bayes classifiers in many intricate real-world scenarios [20].

$$P\left(\frac{m}{n}\right) = \frac{P\left(\frac{n}{m}\right) * P(m)}{P(n)} \qquad (3)$$

where:
$p(\frac{m}{n})$ is the posteriorprobability,
p(m) is the class priorprobability,
p(n) is the predictor priorprobability,
$p(\frac{n}{m})$ is the predictor's probability and likelihood.
Algorithm for Naïve Bayes:
INPUT:
Step 1: Training Dataset
PROCEDURE:
Step 2: F = $(f_1, f_2, f_3, f_4, \dots, f_n)$ : Predictor variables in the
Testing dataset
Step 3: Utilizing the Gaussian Density, determine the probability of each predictor variable.
Step 4: Determine the likelihood for Every Class.
Step 5: Get the Class with the Greatest Likelihood
OUT PUT:
Step 6: Predicted class of the Testing dataset

Simple, straightforward, and effective in handling complex, non-linear data is the Naïve Bayes classification method. It be dependent going on assumptions and division restrictive self-determination, which results in a defeat of correctness. With the use of SVM-RFE to choose the ten most significant predictors, Naïve Bayes has produced an accuracy of 84.1584% [21]. All 13 characteristics of the Cleveland dataset have been used to reach an accuracy of 83.49% [21].

**4.3 Decision Tree:** This approach for classification may be used to both numerical and categorical facts. Structures like trees may be created using decision tree. Using decision trees to handle medical datasets is common and straightforward. The data is analyzed on a graph structured like a tree and is easy to implement. Three nodes form the basis of the decision tree model's analysis.

- All other nodes work dependent on the root node, which is the primary node.
- Internal node: manages different properties.
- Leaf node: symbolizes each test's outcome.
- According to significant signs, This scheme partitions the data into many correlated categories. The data are split according on the entropy of each characteristic, with the predictors with the greatest information gain, or minimal entropy, being identified as follows:

Algorithm for Decision Tree:
INPUT:
Step 1: Samples from the Dataset with target classes
PROCEDURE:
Step 2: For all attributes, evaluate their potential to split the data.
Step 3: For each record, apply the Decision Tree Classifier
algorithm to classify the attribute space.
Step 4: Determine the total leaf nodes$n_1, n_2, n_3, n_4, \ldots, n_m$.
OUTPUT:
Step 5: Divide the samples into partitions $m_1, m_2, m_3, m_4, \ldots\ldots\ldots, m_m$
according to the leaf nodes.

$$\text{entropy(S)} = \sum_{i=1}^{c} -\text{pi log2 pi} \qquad (4)$$

$$\text{gain(S, A)} = \text{entropy(S)} - \sum_{\text{v}\in\text{V alues(A)}} \frac{|\text{Sv}|}{|\text{S}|} \text{ entropy(Sv)}$$

The outcomes are simpler headed for read as well as understand [22].For the reason that it examines a network structured like a tree, this method performs extra accurately than other algorithms. To make decisions, just one characteristic is examined whereas the information might be excessively secured Chauhan et al. [23] have attained an accuracy of 71.43% using their decision tree.

**4.4 K-nearest neighbor (K-NN):** One approach as a supervised classification method. Objects are categorized base on top of their closest neighbors. This kind of learning is instance-based. Euclidean distance is used to calculate an attribute's distance from its neighbors [22]. It employs assembled from designated incorporates them into the grading of an additional point. The data are grouped according to how similar they are to one another, and K-NN might be used to complete the data's absent values. The data set undergoes many prediction procedures after the completion of the missing value fill-in. increasing the accuracy may be achieved by combining these algorithms in different ways. The K-NN technique may be easily implemented without requiring the development of a model or additional presumptions. This approach is flexible as well as may be use to search, regression, as well as classification tasks. K-NN is the simplest method, however its accuracy is impacted by characteristics that are irrelevant and noisy. The accuracy was 83.16% attained K = 9 as the value in a Pouriyeh et al. research [24].
Algorithm for K-NN:
INPUT: (m, x , k):
Step 1: m : samples from a dataset with target classes
x : the sample to be classified
k : the number of nearest neighbors to consider
PROCEDURE:
Step 2: Calculate the Euclidean distance between the sample $x$ and all $m$ samples from the dataset.
Step 3: Arrange the calculated distances in ascending order.
Step 4: Take the first $k$ values (the $k$ nearest neighbors).
Step 5: Determine the class of $x$ based on the majority class of the $k$ nearest neighbors.

OUTPUT:
Step 6: Predicted class for the sample $x$.

**4.5 Random Forest (RF) Algorithm:** The Random Forest approach is an algorithmic method for supervised classification. Many trees come together in this process to form a forest. The random forest is structured so that each tree represents a class expectation, and the class with the most votes becomes the model forecast. Higher accuracy in the random forest classifier is correlated with a larger tree count. In forest research, three main methodologies are used:

- Forest RI (randominputchoice);
- Forest RC (randomblend);
- ForestRIandRC (combined).

Algorithm for Random Forest:
INPUT:
Step 1: Training data samples (features)
Step 2: Training data labels (target classes)
PROCEDURE:
Step 3: Number of trees to be built in the forest
Step 4: (maxi_features) Count of characteristics to take into account while choosing the optimum split
Step 5: (maxi_dept) Maximum depth of the trees (can be used to prevent overfitting)
Step 6: (mini_samples_spli) Sample count minimum needed to separate an internal node

OUTPUT:
Step 7: A trained Random Forest model consisting of multiple decision trees

While it is for applications such as classification and regression analyses, it performs best in the former and has the ability to accommodate missing information. Because it need extensive datasets and an increase in tree cover, the results are not only slow to provide predictions, but also mysterious. In [21], the random forest approach obtained 91.6% accuracy with the Cleveland dataset.

**4.6 Ensemble methods**: It was determined how accurate each classifier was, as well as the accuracy of Random Forest, K-NN, Decision Tree (DT), Support Vector Machine (SVM), Logistic Regression (LR), and Gradient Boosting (GB). The classifiers' performance was enhanced by the use of ensemble approaches. A variety of ensemble techniques were used, including bagging and boosting, stacking, and majority voting. By using majority voting and stacking, the separate classifiers were combined to enhance the classifiers' performance. Using a Meta classifier, the stacking method merges the separate classifiers. The decision tree classifier serves the same as the foundational classifier in favor of the bagging method, which enhances performance.

Table 3: percentage of accuracy results of classification techniques

| Accuracy(%) | Logistic Regression | SVM | NaïveBayes | K-nearestneighbor | Decision tree | Randomforest | Ensemble methods |
|---|---|---|---|---|---|---|---|
| Testingset | 69 | 67.5 | 72 | 75 | 90.5 | 93.9 | 92.4 |
| Trainingset | 75.3 | 67.5 | 77 | 87.5 | 100 | 100 | 100 |

**Benchmarking Of The Proposed Model:** LightGBM (Light gradient boosting machine) be a powerful gradient boost outline that plays a significant role in ensemble learning It is designed to be both highly efficient and capable of delivering strong predictive performance. It is optimized for speed and memory usage, making it suitable for handling large datasets and high-dimensional feature spaces. Light GBM adopts a leaf-wise growth strategy for decision trees.
Algorithm for Ensemble method with Light GBM:
INPUT:
Step 1: Training Dataset
PROCEDURE:
Step 2: Number of Light GBM models $n$ .
Step 3: For each Light GBM model, train it on the training dataset using bootstrapped sampling (random sampling with replacement).
Step 4: Hyper parameters for Light GBM (learning rate, maxi depth, number of iterations)
OUTPUT:
Step 5: ensemble model that can predict the class of new samples.
This approach can lead to deeper and potentially more complex trees, which can capture intricate patterns in the data. However, it also requires careful hyper parameter tuning to prevent over fitting. Light GBM can take advantage of parallel and distributed computing, which can further enhance its efficiency, making it well-suited for large-scale applications. In ensemble learning, Light GBM is commonly used as a base model or as a primary model within a gradient boosting ensemble. It is trained sequentially, with each new model in the ensemble focused on correcting the errors made by the existing models. To maximize its performance, Light GBM often requires careful hyper parameter tuning. Parameters such as the learning rate, maximum depth of trees, number of iterations, and regularization strength need to be optimized to attain the most excellent results.

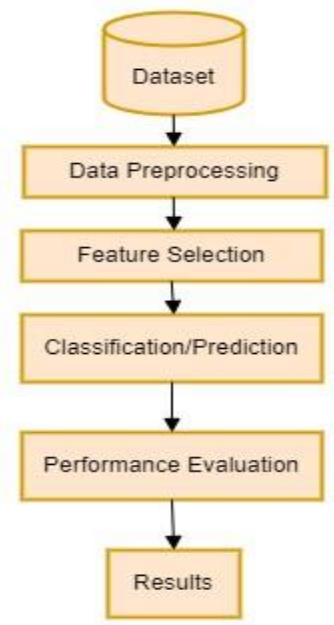



**Fig.3.** Flow chart of proposed model

In ensemble learning, Light GBM is commonly used as a base model or as a primary model within a gradient boosting ensemble. It is trained sequentially, with each new model in the ensemble focused on correcting the errors made by the existing models. To maximize its performance, Light GBM often requires careful hyper parameter tuning. Parameters such as the learning rate, maximum depth of trees, number of iterations, and regularization strength require to be optimized to achieve the most excellent results.

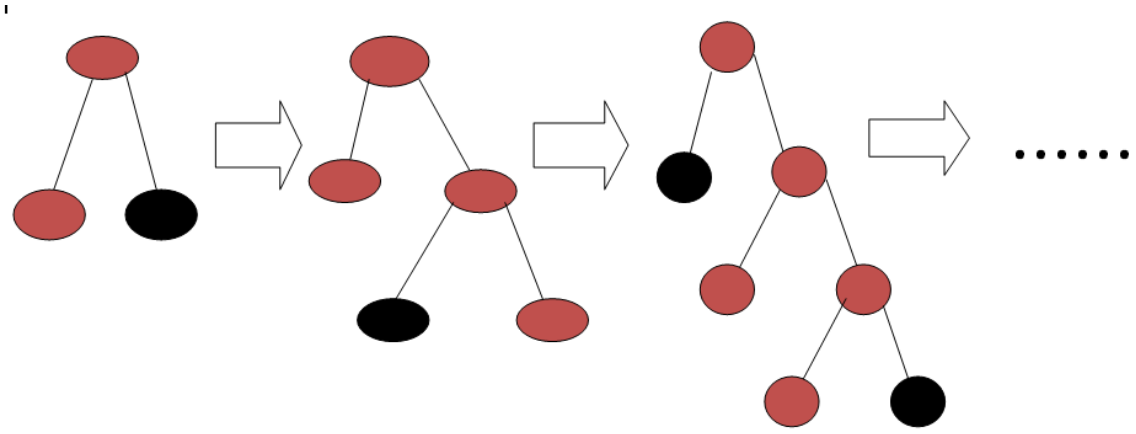


Fig 4: Architecture of Light BGM

## 5. Results and Analysis:

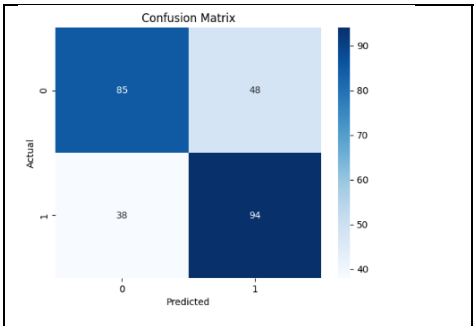
Purpose of the study is determining a patient's likelihood of having heart disease using Naïve Bayes (NB), decision trees(DT), random forests(RF), and K-nearest neighbor(K-NN) on top of the UCI repository, machine learning classification algorithms were studied in this study. After the dataset was classified, its training and test sets were divided. The information is preprocessed and supervised classification techniques such Naïve Bayes, decision trees, K-nearest neighbor, random forests, and ensemble approaches are used in order in the direction of obtain the accuracy score in table 3. We recorded the precision achieve outcome of many classification algorithms using training and test datasets for Python programming. The percentage (%) of accuracy ratings for a number of approaches are shown in Table 4. The models' accuracy scores in predicting heart disease by different authors are compared in Table 4.

| Dataset | Dataset shape | Author | Approaches | Accuracy(%) |
|---|---|---|---|---|
| statlog+heart | (270,14) | KUMAR DWIVEDI [29] | Naïve Bayes | 83 |
| | | | Classification tree | 77 |
| | | | K-NN | 80 |
| | | | SVM | 82 |
| | | | ANN | 84 |

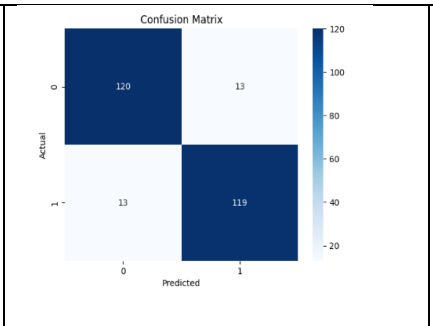
| | | | | |
|---|---|---|---|---|
| | | | Logistic Regression | 85 |
| heart+disease | (303,14) | Devansh Shah[21] | Naïve Bayes | 88.1 |
| | | | K-NN | 90.7 |
| | | | Decision tree | 80.2 |
| | | | Random forest | 86.8 |
| Cleveland UCI | (303,14) | SENTHILKUMAR MOHAN[30] | Naive bayes | 78.8 |
| | | | Generalized linear model | 85.1 |
| | | | Deep learning | 87.4 |
| | | | Logistic regression | 82.9 |
| | | | Decision tree | 85 |
| | | | Random forest | 86.1 |
| | | | Gradient boosting trees | 78.3 |
| | | | SVM | 86.1 |
| | | | VOTE | 87.4 |
| | | | HRFLM | 88.4 |
| Merged dataset | (1322,14) | Proposed model(Light GBM ) | SVM | 67.5 |
| | | | Decision Tree | 90.5 |
| | | | K-Nearest Neighbours | 75 |
| | | | Random Forest | 93.9 |
| | | | ADA Boosting | 76.8 |
| | | | Bagging | 92.5 |
| | | | Gradient Boosting | 87.1 |
| | | | Logistic regression | 69 |
| | | | Naive bayes | 72 |
| | | | Hyper-parameter tuning | 90.9 |
| | | | Light GBM(proposed model) | 94 |

Table 4: Accuracy of heart disease prediction using various methods and a proposed LIGHT GBM model.

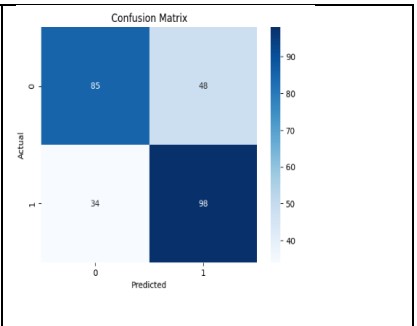
**5.1 Performance Evaluation Measures:** To analyze performance of the proposed model, we computed metrics as expressed in Eqs. 5-9. these metrics include count of correctly identified True Positive (TP), correctly identified True Negative (TN), incorrectly missed False Negative (FN), and incorrectly identified False Positive (FP) in predictions. These basic measures are used to compute sensitivity and specificity which are critical in healthcare diagnosis.




CM of SVM




CM of DT




CM of LR

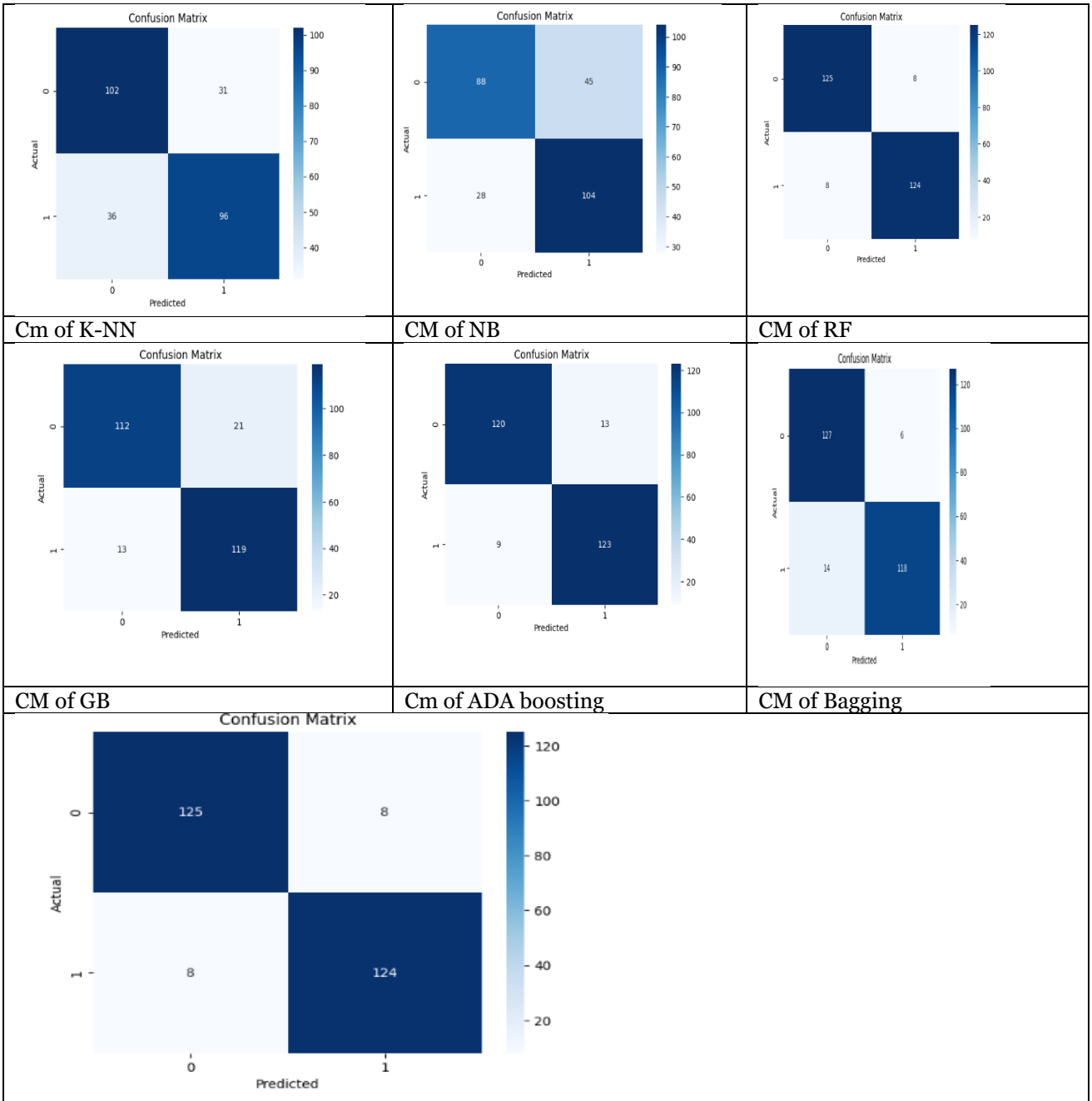



**Fig.5** CM of Proposed model of Light GBM

Fig 5: confusion matrix for classification of machine learning approaches.

| Confusion Matrix | | Predicted | |
|---|---|---|---|
| | | Positive | Negative |
| Actual | Positive | TP | FP |
| | Negative | FN | TN |

Table 6 : Performance analysis for HDD&P Using Ensemble Learning With LIGHT GBM

**5.1.1 Accuracy:** In machine learning, accuracy is a metric for how accurately a model forecasts the future. The ratio of precise forecasts to total predictions is employed to compute it. In terms of math, it is stated as:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \quad (5)$$

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

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

**5.1.3 Recall (sensitivity):** Recall can be defined as the ratio of True Positive cases to total actual Positive cases, and when the recall is equal to 1, it indicates that model is effective in classifying Positive cases. Formula to compute recall is displayed in Eq. 7.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (7)$$

**5.1.4 Specificity:** Specificity (True Negative Rate) calculates the percentage of real negative instances that are accurate negative forecasts. It evaluates a test's effectiveness in correctly classifying individuals without the condition. The formula for specificity is shown in Eq. 8.

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (8)$$

**5.1.5 F1- Score:** The F1 - Score is calculated by considers both Recall and Precision. It ensures the right balance between precision and recall. This is important in situations where both precision and recall are key measures, such as medical diagnosis, spam email detection. The F1-score attains a value of 1 only when both measures, namely recall and precision, also achieve a value of 1.F1 -score is computed using the following Eq.9.

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (9)$$

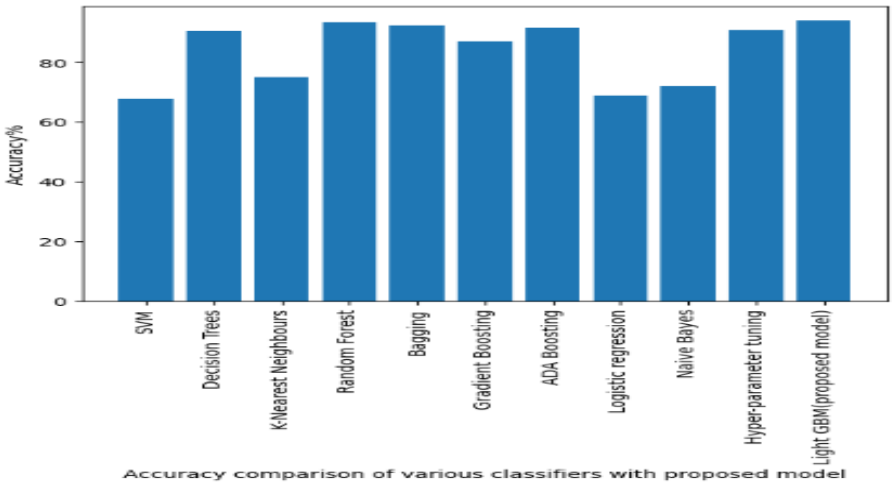



Fig.6. Comparing the accuracy of different classifiers with the suggested model.

## 6. Conclusion:

There is a fundamental disadvantage with the previous suggested algorithms for heart disease prediction: their performance becomes less and less effective with larger datasets. Although efficient attribute extraction might lead to possible improvements, this constraint is ascribed to the inefficiency in dataset classification. A further problem stems from the observation that while the classifiers prediction accuracy improves as the dataset grows, there is a threshold beyond which any increases have a deleterious effect on accuracy. The suggested approach applies ML techniques to the prediction of heart problems with more accuracy and at a lower cost. There have been many classifiers used, and Light GBM has shown an amazing 94% accuracy. According to the research, effectively managing features may result in significant improvements in the classification of heart disease predictions. The research highlights the significance of effectively handling features to improve the categorization of heart disease prognoses. Upcoming research on the performance of heart illness forecast studies is anticipated to use the results of these suggested methodologies as standards.

## References:


[1] Kim JK & KangS (2017) Neural Network-Based Coronary Heart Disease Risk Prediction Using Feature Correlation Analysis, Journal of Healthcare Engineering,Article ID 2780501] https://doi.org/10.1155/2017/2780501

[2] Miao K H, Miao J H and G J Miao (2016) Diagnosing Coronary HeartDisease UsingEnsemble MachineLearning,InternationalJournalofAdvancedComputerScienceand Applications,(IJACSA),30-39.

[3] Li H et al. (2018) Ensemble learning for overall power conversion efficiencyoftheall-organicdye-sensitizedsolar cells,IEEE Access34118–26.

[4] G E Guraksin, U Ergun and O Deperlioglu,(2010) Classification of the heart sounds via artificial neural network, International Journal of Reasoning-Based Intelligent Systems, 272–278,

[5] R K Sinha, Y Aggarwal, and B N Das, (2007) Backpropagation artificial neural network classifier to detect changes in heart sound due to mitral valve regurgitation, Journal of MedicalSystems, 205–209.

[6] A Kandaswamy, C S Kumar, R P Ramanathan, S Jayaraman, and N Malmurugan, (2004) Neural classification of lung sounds using wavelet coefficients, Computers in Biology and Medicine, 523–537.

[7] Vanisree K, JyothiSingaraju,(2011) Decision support system for congenital heart diseasediagnosis based on signs and symptoms using neural networks, Int J Comput Appl, 0975 8887.

[8] Chauhan Shraddha, Aeri Bani T, (2015) The rising incidence of cardiovascular diseases inIndia, assessing its economic impact, J Prev. Cardiol,735–40.

[9] LathaParthiban, Subramanian R, (2008) Intelligent heart disease prediction system usingCANFIS and genetic algorithm, Int. J. Biol. Biomed. Med. Sci.

[10] Cleveland Clinic Foundation, Heart disease data set, Available at http://archive.ics.uci.edu/ml/datasets/ Heart+Disease.

[11] Statlog, Heart disease data set, Available at :https://archive. ics.uci.edu/ml/datasets/ Statlog+%28Heart%29.

[12] C G D Dua, Oct 21 (2021)(UCI) Machine Learning Repository. Accessed:, [Online]. Available: http://archive.ics.uci.edu/ml

[13] G Bazoukis, S Stavrakis, J Zhou, S C Bollepalli, G Tse, Q Zhang,J P Singh, and A A Armoundas, (2021) Machine learning versus conventional clinical methods in guiding management of heart failure patients_A systematic review, Heart Failure Rev.

[14] A Makhlouf, I Boudouane, N Saadia, and A R Cherif, (2018) Ambient assistance service for fall and heart problem detection, J. Ambient Intell. Humanized Comput, 1527_1546,.

[15] http://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/heart-disease.names

[16] amalingam VV, Dandapath A, Raja MK (2018) Heart disease prediction using machine learning techniques: a survey. Int J Eng Technol684–7.

[17] Patel J, TejalUpadhyay D, Patel S (2015) Heart disease prediction using machine learning and data mining technique. Heart Dis129–37.

[18] S Uddin, A Khan, ME Hossain, MA Moni, (2019) Comparing different supervisedmachine learning algorithms for disease prediction, BMC Med. Inf. Decis. Making 1–16.

[19] Md Mamun Ali a, Bikash Kumar Paul a,b,c, Kawsar Ahmed b,c,*, Francis M Bui d, Julian M W Quinn e, Mohammad Ali Moni (2021)Heart disease prediction using supervised machine learning algorithms, Performance analysis and comparison Computers in Biology and Medicine136,104672,https://doi.org/10.1016/j.compbiomed.2021.104672.

[20] Fatima M, Pasha M (2017|) Survey of machine learning algorithms fordiseasediagnostic.JIntellLearnSystAppl,1–16.https://doi.org/10.4236/jilsa.2017.91001.

[21] Devansh Shah Samir Patel Santosh Kumar Bharti(2020) Heart Disease Prediction using Machine Learning Techniques SN Computer Science 345 https://doi.org/10.1007/s42979-020-00365-y

[22] Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N (2017) Can machine-learning improve cardiovascular risk prediction using routineclinicaldata?PLoSONE,e0174944.

[23] Chauhan R, Bajaj P, Choudhary K, Gigras Y (2015)Framework to pre-dicthealthdiseasesusing attributeselectionmechanism.In:2ndinternationalconferenceoncomputingforsustainableglobaldevelopment(INDIACom).IEEE,1880–84.

[24] PouriyehS,VahidS,SanninoG,DePietroG,ArabniaH,Gutierrez-JA(2017)comprehensiveinvestigationandcomparisonofmachinelearningtechniquesinthedomainofheartdisease.IEEEsymposium on computers and communications (ISCC) IEEE 204–207

[25] M S Amin, Y K Chiam, and K D Varathan, (2019) Identification of significant features and data mining techniques in predicting heart disease, Telematics and Informatics, 82–93.

[26] I Ketut Agung Enriko, M Suryanegara, and D AgnesGunawan, (2016)Heart Disease Prediction System Using K-Nearest Neighbor Algorithm with Simplified Patient's Health Parameters, Springer, Berlin, Germnay.

[27] K Subhadra and B Vikas, (2019) Neural network based intelligent system for predicting heart disease, International Journal of Innovative Technology and Exploring Engineering, 484–487.

[28] S N Khan, N M Nawi, A Shahzad, A Ullah, and M F Mushtaq, (2019) Comparative analysis for heart disease prediction, International Journal on Informatics Visualization, 227–231.

[29] DwivediAK, (2018)Performanceevaluationofdifferentmachinelearn-ing techniques for prediction of heart disease, Neural ComputAppl,685–693.
[30] Senthilkumar Mohan , Chandrasegar Thirumalai, And Gautam Srivastava, (2019)Effective Heart Disease Prediction UsingHybrid Machine Learning Techniques Digital Object Identifier 10.1109/ACCESS,2923707.