IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 26, 2022, accepted February 12, 2022, date of publication February 21, 2022, date of current version March 7, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3153047

Comparative Study of Optimum Medical Diagnosis of Human Heart Disease Using Machine Learning Technique With and Without Sequential Feature Selection

GHULAB NABI AHMAD[1], SHAFIULLAH[2], ABDULLAH ALGETHAMI[3], HIRA FATIMA[1], AND SYED MD. HUMAYUN AKHTER[4]

[1]Institute of Applied Sciences, Mangalayatan University, Aligarh, Uttar Pradesh 202146, India
[2]Department of Mathematics, K. C. T. C. College, Raxual, a Constituent Unit of BRA, Bihar University, Muzaffarpur, Bihar 842001, India
[3]Department of Mechanical Engineering, Taif University, Taif 26571, Saudi Arabia
[4]Department of Applied Sciences and Humanities, Institute of Technology and Management, Aligarh, 202001 Uttar Pradesh, India

Corresponding author: Shafiullah (sha.stats@gmail.com)

ABSTRACT Predicting heart disease is regarded as one of the most difficult challenges in the health-care profession. To predict cardiac disease, researchers employed a variety of algorithms including LDA, RF, GBC, DT, SVM, and KNN, as well as the feature selection algorithm sequential feature selection. For verification, the system employs the K-fold cross-validation approach. These six strategies were used to conduct the comparative study. The Dataset for Cleveland, Hungray, Switzerland, and Long Beach V, as well as the Dataset Heart Statlog Cleveland Hungary, were used to assess the models performance. For both Hungary, Switzerland & Long Beach V and Heart Statlog Cleveland Hungary Dataset, Random Forest Classifier sfs and Decision Tree Classifier sfs produced the highest and almost identical accuracy values (100%, 99.40% and 100%, 99.76% respectively). The findings were compared to previous research that focused on cardiac prediction. In the future, we hope to extend the model even further so that it may be used with various feature selection techniques; another possibility is to use a random forest classifier. The major goal of this study is to improve on previous work by developing a new and unique technique for creating the model, as well as to make the model relevant and easy to use in real-world situations.

INDEX TERMS Heart disease, sequential feature selection, DT, RF, SVM, GBC, LDA, confusion matrix · ROC curve.

I. INTRODUCTION

The study of disease diagnosis is crucial in the realm of healthcare [1]. A disease is defined as any cause or set of conditions that lead to suffering, sickness, malfunction, or finally death of a human person. Individuals have the fundamental right to good health, according to WHO principles [2]. It is thought that proper health care services should be offered for frequent health checks. Heart disease is the leading cause of death in the world, accounting for nearly 31% of all deaths. Early detection and treatment of many cardiac disorders are highly challenging, especially in poor countries, due to a lack of diagnostic centers, skilled doctors, and other resources that affect the proper prognosis of heart disease [3]. Some prevalent risk factors, such as diabetes, high blood pressure, and excessive cholesterol, make it difficult to detect heart disease. Underlying disorders induce irregular cardiac rhythms and breathing difficulties, such as pulmonary cracks, improved jugular vein weight, and borderline edema [4]. Because the symptoms of cardiac disease are so varied, they must be treated with extreme caution. Failure to do so may have a negative impact on the heart [5]. According to the American College of Cardiology, there are 26 million individuals globally who have heart disease, and 3.6 million people are tested each year. Within a year 15–35 % of individuals with heart disease will die, and the rest will die in 4–5 years. Diseases can affect a person physically and mentally, and they can have a significant impact on how they live. The pathological process is defined as the study of the causes of disease. Signs or symptoms that are evaluated by

The associate editor coordinating the review of this manuscript and approving it for publication was Mauro Gaggero.

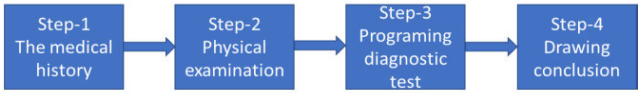


FIGURE 1. Block diagram of the medical diagnosis process.

clinical professionals form the basis of disease. Diagnosis is described as the process of determining the path physiology of a disease based on its indications and symptoms. As illustrated in **Figure-1**, diagnosis may also be described as the process of determining which disease an individual has, based on their symptoms and indicators. The information was gleaned from medical records. The knowledge necessary for diagnosis is based on a physical examination of a person with a medical pathology. During this treatment, at least one diagnostic procedure, such as medical testing, is frequently performed. A medical practitioner will go through a procedure that includes multiple steps in order to make an accurate diagnosis enabling them to get the most quantity of data feasible [6].

Disease diagnosis is the most difficult procedure and, at the same time, a crucial phenomenon for a medical care expert to understand before reaching a judgment. The diagnostic procedure might be lengthy and difficult. Care specialists collect empirical facts to determine a patient's disease to reduce the uncertainty in medical diagnosis health. Due to a mistake in the diagnostic process, the patient's necessary therapy may be postponed or ignored, resulting in major health complications. Unfortunately, not every doctor is a specialist in every field of medicine. As a result, an independent verdict system was needed that combined human understanding with Machine Learning (ML) precision [7]. To get accurate outcomes from the diagnosis procedure at a lower cost, we need a good decision support system. For human specialists, classifying disease based on multiple factors is a difficult task, but machine learning techniques might assist to detect and treat such situations. Various machine learning approaches are currently being applied in medicine to accurately diagnose cardiovascular disease. ML is a component of computer science that enables computers to become more intelligent. Learning is a must for every intelligent system. Learning-based strategies in machine learning include sequential feature technique. One of the most significant is Artificial Intelligence (AI) technologies in the medical field is a rule-based intelligent system, which provides a collection of if-then rules in medical healthcare and works as a decision support system. AI-based autonomous techniques need very little human interaction are progressively adding intelligent systems in the medical business [8]. In recent years, medical aid software has been developed using computer technology and machine learning techniques as a support system for the early identification of cardiovascular disease [9]. Early detection of any heart-related disease can lower the chance of mortality [10]. In medical data, many ML algorithms are utilized to comprehend the pattern of data and make predictions from it [11]. Healthcare data is typically large in volume and structured in a complicated way [12]. ML algorithms can manage large amounts of data and mine it for the useful information [13]. ML algorithms use historical data to learn and predict real-time data. This type of machine learning framework for cardiac sickness prediction can urge cardiologists to act more quickly, allowing more patients to get medications in a shorter amount of time, perhaps saving a substantial number of lives [14]. Machine Learning is a branch of Artificial Intelligence study that has gained a lot of attraction in recent years. Machine Learning algorithms can perform a variety of tasks, including prediction, classification, and decision-making. To learn ML algorithms, we need training data [15]. The majority of investigations at the Literary Research Centre illustrate the use of feature determination techniques and other machine learning algorithms, which is a strategy to organize individuals in an expected fashion or as cardiovascular disease patients. Previously K. J. Shanthi, D. K. Ravish *et al.* (2015), used a vast amount of data created by the medical business that was not efficiently utilized. The innovative procedures proposed here are simple and successful in lowering the cost and improving the calculation of temperament ailment. The numerous research strategies used in this study for the prediction and classification of heart disease utilizing ML and deep learning (DL) techniques are extremely accurate in determining the usefulness of these methods [16], [17].

K. Polaraju *et al.* [18] suggested a Multiple Regression Model for Heart Disease Prediction, which shows that Multiple Linear Regression is suitable for predicting heart disease risk. The task is done using a training data set of 3000 examples with the 13 different characteristics mentioned before. The data set is divided into two parts, with 70% of the data being used for training and 30% being used for testing. Based on the findings, it is obvious that the Regression method has higher classification accuracy than other techniques [19]. Different data mining algorithms were developed by Brahmi and Shirvani [20] to test heart disease prediction and diagnosis. The major purpose is to compare and contrast J48, Decision Tree, KNN, SMO, and Nave Bayes, as well as other classification algorithms. Following that, the accuracy, precision, sensitivity, and specificity of certain performances are evaluated and compared. J48 and decision tree [21] are the strongest strategies for predicting cardiac disease. ''An Intelligent Learning System based on Random Search Algorithm and Optimized Random Forest Model for Improved Heart Disease Detection'' was developed by Ashir Javeed, Shijie Zhou, and colleagues [22]. This study employs the random search algorithm (RSA) for factor selection and the random forest model for cardiovascular disease diagnosis. This model is primarily designed to be used with the grid search algorithmic software. For cardiovascular disease prediction, two types of experiments are employed. Only a random forest model is built in the first experiment, while a suggested Random Search Algorithm-based random forest model is developed in the second. This technique is more efficient and straightforward than the traditional random forest model.

It has a 3.3 % greater accuracy than a traditional random forest. Physicians can benefit from the suggested learning system by improving the quality of heart failure detection [23]. A support vector machine recursive feature elimination feature selection method based on artificial contrast variables and mutual information [24]. The first features to be removed are noise, redundancy, or irrelevant features; the most significant feature is the last to be removed. The accuracy of SVM-RFE and Principal Component Analysis-Support Vector Machine (PCA-SVM) in predicting cardiac disease was 88.24% [25]. By removing duplicate data, this strategy can enhance heart disease prediction accuracy. It is, however, a greedy strategy that attempts to discover the best possible categorization combination by deleting the poorest attributes one at a time. When paired with other features, the features that were eliminated before might give a large performance boost. Mohan *et al.* (2019) [23] developed an effective hybrid machine learning strategy. Random forest and linear methods are combined in the hybrid technique. For prediction, the dataset and subsets of characteristics were gathered. The pre-processed knowledge (data) collection of cardiovascular disease was used to choose a subset of specific properties. Hybrid approaches were used to diagnose cardiovascular disease after prep-processing [26]. ''Prediction and Diagnosis of Heart Disease Patients Using Data Mining Technique'' was created by Mamatha A. P. and Shaicy P Shaji [27]. This article employs the Artificial Neural Network, KNN, Random Forest, and Support Vector Machine methodologies. The Artificial Neural Network [28] is compared to the above-mentioned classification algorithms in data mining to predict improved accuracy in diagnosing heart disease. A disease prediction system based on Feature Selection (FS) was introduced by Sandhiya and Palani *et al.* [29]. As the FS method, the incremental FS algorithm (IFSA) was used. This showed that IFSA was a mixture of intelligent conditional random fields and linear correlation coefficient-based FS. The distance between characteristics was used to categorize the features. Finally, diseases including cancer, heart disease, and diabetes were predicted with a lower false alarm rate (FAR) utilizing temporal-convolutional NN (T-CNN) [30]. According to the authors, results may be attained with excellent precision by following this approach. ML can forecast different diseases by utilizing electronic information. Prediction and Diagnosis of Heart Disease Patients Using MLT were recently done by different researchers [31], [32].

The Sequential Feature Selection (SFS) method starts with an empty set and adds an element that provides the most persuasive incentive for the target effort in the first phase. Initial with the second phase, the remaining characteristics are added to the current subset precisely. In a good approach, the algorithm picks several features from a collection of features and assesses them for model iteration, lowering and improving the number of features for the model to achieve optimal performance and outcomes. Bharti *et. al* [58] have used the ML for predicting heart disease and got 100% accuracy. In any event, determining the best feature subset to employ in the diseases prediction and analysis framework is still a work in progress. Present relevant writings [33], [34] always focus on picking a subclass of pieces to magnify the correctness of a single/large-scale arrangement. Predicting cardiac disease is considered one of the most challenging tasks in the medical field. Researchers used a range of algorithms, including LDA, RF, GBC, DT, SVM, and KNN, as well as the feature selection method SFS, to predict cardiac illness. The system uses a K-fold cross-validation technique for verification. The comparison research was conducted using these six methodologies. The models performance is evaluated using the Datasets for Cleveland, Hungray, Switzerland, and Long Beach V, as well as the Dataset Heart Statlog Cleveland Hungary. Random Forest Classifier sfs and Decision Tree Classifier sfs achieved the highest and almost comparable accuracy scores for both the Hungary, Switzerland & Long Beach V and Heart Statlog Cleveland Hungary Datasets (100 %, 99.40 % and 100%, 99.76% respectively). The findings were compared to previous research that focused on cardiac prediction.

By training the dataset, the algorithm determines the best answer. Various performance indicators, including Accuracy, Sensitivity, Specificity, Precision, and F1-Score, may be used to evaluate the models' performance is tested on a subset of features selected by Sequential Forward Selection (SFS) method with 5-fold cross-validation for Heart Disease Clinical Record Data Set 2020. Our work suggests the combination of the ML modes and optimization technique that predicts heart disease with the highest accuracy.

The rest of the paper is organized as follows: **Section-A** includes similar work on prior research that used different machine learning algorithms to predict cardiac disease. **Section-B** more clearly explains the technique of the proposed study. **Section-C** explains the experimental result findings, as well as a comparison of past investigations and the methodology utilized and different models. **Section-D** explains our results, conclusion and study directions in the future.

II. MACHINE LEARNING CLASSIFIERS

An AI classification method is used to distinguish between heart disease patients and healthy people. The assumptions behind various well-known categorization algorithms are briefly discussed in this section.

A. LINEAR DISCRIMINANTS ANALYSIS (LDA)

LDA is employed when all populations variation covariance grids are homogenous. Our selection strategy in LDA is based on the linear score function, which is the population element represented by each θ_i in our group, and the set difference covariance frame the linear score function's features are [43].

$$\begin{aligned} S_I^{\alpha}(x) &= -\frac{1}{2}\theta_i'\Sigma^{-1}\theta_i + \theta_i'\Sigma^{-1}x + logQ(\Pi_i) \\ &= D_{i0} + \sum_{j=1}^{\theta} D_{ij}X_i + logQ(\Pi_i) \\ &= D_i^L(x) + logQ(\Pi_i) \end{aligned} \tag{1}$$

TABLE 1. Comparative study table of highly sited papers of machine learning techniques for the prediction of heart disease.

S.No.	Authors	Disease prediction	Techniques	Features	Evaluation measures
1.	Kumar et al. [35]	Heart murmur classification	Nonlinear classifier (SVM), foating sequential forward method (SFFS)	Out of 17 features 10 were selected *Feature Set 1:* loudness 1, zcr 1, transition ratio1, spectral power 2, fundamental frequency 1,spectral shape 1, spectral power 3, fux 1, stat skewness 1, Lyapunovexponent 1, *Feature Set 2:* PoS 1–32 no selection, kept as in original work *Feature Set 3:* WT detail 7, VFD 8, Shannon energy 5, Shannon energy 6, GMM cycle 5,Shannon energy 4, GMM murmur 5, Eigenfrequency1 2, WT entropy 10, GMM cycle 4, Eigenfrequency1 1, Shannon energy 8, VFD 2, HOS 1 *Feature Set 4:* loudness 1, transition ratio1, spectral power 2, stat skewness 1, Lyapunovexponent 1, PoS 13, PoS 16, PoS 17, PoS 18, PoS 22, Shannon energy 8, WT details 5, WT entropy 6, ST map 3, Eigenfrequencies1 4, Eigenfrequencies2 10, Eigentimes1 5, HOS 6, HOS 13, GMMx 7, GMMx murmur 4, VFD 3	Set 1 Set 2 Set 3 Set 4 *Sensitivity (Se in %)* 95.74 93.02 90.51 96.15 *Specifcity (Sp in %)* 95.01 96.79 91.26 96.16
2.	Khemphila et al. [36]	Heart disease classification	Multi-layer Perceptron (MLP) with Back Propagation learning algorithm, feature selection algorithm, Artificial neural networks (ANN)	Out of 13, 8 features selected by feature selection algorithm Selected features are: Thal, Chest Pain Type, Number Colored Vessels, Old Peak, Maximum Heart Rate, Induced Angina, Slope, Age	*With 13 features* Training Accuraacy 88.46% Validation Accuracy 80.17% *With 8 features* Training Accuraacy- 89.56% Validation Accuracy 80.99%
3.	Mokeddem et al. [37]	Coronary artery disease	Genetic Algorithm (GA), wrapped Bayes Naïve (BN), Best First Search (BFS), Sequential Floating Forward Search (SFFS)	Features selected by FS approach GA wrapped BN: cp, Sex, restescg, oldpeak, slope, ca, thal; GA wrapped SVM: cp, Age, exang, oldpeak, slope, ca, thal; GA wrapped MLP: cp, Age, Sex, restbps, slope, ca, thal; GA wrapped C4.5: cp, fbs, ca, thal; BFS wrapped BN: chol, fbs, thalach, exang, ca, thal; SFFS wrapped BN: cp, restescg, thalach, oldpeak, ca, thal	*Classifcation accuracy* SVM: 83.5% MLP: 83.16% C4.5: 80.85% *Wrapper based feature selection algorithms accuracy* GA wrapper: 85.50
4.	Usman et al. [38]	Heart disease prediction	cuckoo search algorithm (CSA) and cuckoo optimization algorithm (COA), SVM, MLP, NB and, RFC	*Features selected by COA and CSA approach in 5-heart disease data set* ***Data set*** / ***Features*** / ***COA*** / ***CSA*** **Eric** / **7** / **4** / **4** Echocardiogram / 12 / 5 / 4 Hungarian / 13 / 6 / 4 Stat log / 13 / 6 / 4 Z-Alizadeh Sani / 55 / 14 / 7	*In all data sets* CSA better performed than COA after feature selection. SVM has highest accuracy before and after feature selection among all the 5-datasets
5.	Haq et al. [39]	Heart disease prediction	Sequential Backward Selection (SBS), K-Nearest Neighbor (k-NN)	**Features eliminated by SBS approach** Number of selected features: 13 (selected one feature at time)	Out of k=8 Kernal of K-NN, highest average accuracy is 90% after eliminating 6 features
6.	Javeed et al. [40]	Heart risk failure prediction	Floating window with adaptive size for feature elimination (FWAFE-ANN) artifcial neural network and (FWAFE-DNN) deep neural network	**Experiment No. 1** Feature selected by FWAFE method: n= 6, n= 7 and n= 11, where n=size of features subset. The optimal accuracy found at n= 11(subset of features) by applying ANN **Experiment No. 2** Feature selected by FWAFE method: Optimal accuracy found at n = 11 (subset of features) by applying DNN	**Experiment No. 1** FWAFE-ANN accuracy: 91.11% **Experiment No. 2** FWAFE-DNN accuracy: 93.23%
7.	Yadav et al. [41]	Heart disease	Pearson correlation, recursive features elimination and lasso regularization and M5P, random tree, Reduced Error Pruning and Random forest ensemble method	*Features selected by 3 method are* Pearson correlation: cp, exang, oldpeak and target Recursive Features selection: 12 features selected Lasso Regularization: 10 features selected	*Accuracy* Pearson correlation- 99.9% Recursive Features selection- 94.12% Lasso Regularization- 99.9%
8.	Ritu Aggarwal [42]	Heart disease	Sequential fearture selection(SFS), LDA, RF, GBC, DT, SVM, and KNN	Out of 13, 9 features selected by feature selection algorithm **Selected features are: age,** creatinine, phoshokinase , ejection_fraction , serum_creatinine, diabetes platelets , serum_creatinine , smoking , time **LDA:** Age, creatinine_phoshokinase ,	Accuracy RFC_FS with fivefold cross validation by FS algorithm SFS, its best accuracy is 86.67%, the RFC and GBC followed closely, and performed better

TABLE 1. *(Continued.)* **Comparative study table of highly sited papers of machine learning techniques for the prediction of heart disease.**

				ejection_fraction ,serum_creatinine ,time RF: Age, diabetes , ejection_fraction ,serum_creatinine DT: Diabetes, ejection_fraction , smoking GBC: Diabetes, ejection_fraction , smoking KNN:Anaemia , ejection_fraction ,platelets , serum_creatinine , time SVM: serum_creatinine , ejection_fraction ,smoking	in terms of accuracy, both of which were 85.56%, the average ROC_AUC, GBC results have a higher accuracy of 74%.
9.	Proposed Method	Heart disease	Sequential fearture selection(SFS), LDA, RF, GBC, DT and SVM	**Data Analysis for cleveland,hungray,switserland & Long Beach V, Data:1025** Out of 13, 11 features selected by feature selection algorithm **Selected features are:** Sex , age ,cp, trestbps ,oldpeak , ca ,chol ,thalach , slope , restecg , thal **LDA:** sex , cp , trestbps , oldpeak , ca **RFC:** age , sex , cp , chol , thalach **DTC:** age ,cp , chol , thalach , slope **GBC:** age , cp , chol, oldpeak , thal **SVM:** cp , restecg, thalach , ca , thal	Accuracy DT Classifiers SFS with fivefold cross validation by FS algorithm SFS, its best accuracy is 100%. The RFC, DTC and GBC on the other hand, came in second and third, respectively, in terms of accuracy 100%., GBC findings have a greater accuracy of 98 % in terms of average ROC_AUC
10	Proposed Method	Heart disease	Sequential fearture selection(SFS), LDA, RF, GBC, DT and SVM	**Data Analysis for statlog_cleveland_hungary, Data:1190** Out of 13, 11 features selected by feature selection algorithm **Selected features are:** Sex , age ,cp, trestbps ,oldpeak , ca ,chol ,thalach , slope , restecg , thal **LDA:** sex , cp , trestbps , oldpeak , ca **RFC:** age , sex , cp , chol , thalach **DTC:** age ,cp , chol , thalach , slope **GBC:** age , cp , chol, oldpeak , thal **SVM:** cp , restecg, thalach , ca , thal	Accuracy DTC Classifiers SFS with fivefold cross validation by FS algorithm SFS, its best accuracy is 99.76% The Random Forest Classifiers, on the other hand, came in second and third, respectively, in terms of accuracy 99.40%.

Were, $D_{i0} = -\frac{1}{2}\theta_i'\sum^{-1}\theta_i$,

$$D_{ij} = \theta_i'\sum^{-1} j^{th} element,$$
$$D_i^L(x) \text{ is a linear discriminent function}$$

The unknown parameters θ_i and $\sum^{-1}$ are used to calculate the linear scoring function. As a result, we must rely on the training data to predict their values.

B. RANDOM FOREST CLASSIFIER

The supervised learning approach is used by Random Forest, a well-known machine learning algorithm. It may be used for both classification and regression problems in machine learning. It is based on ensemble learning, which is a technique for combining a large number of classifiers to solve a difficult problem and improve the model's performance. "A random forest is a classifier that contains a number of decision trees on various subsets of a given dataset and takes the average to improve the dataset's prediction accuracy," as suggested by the name Rather of relying on a single decision tree [44]. The random forest considers the predictions from each tree and predicts the final output based on the majority votes of projections. The flowchart of the Random Forest Classifier is given in **Figure-2.**

C. GRADIENT BOOSTING CLASSIFIER

Gradient boosting is a type of Artificial intelligence (AI) that may be used to solve regression and classification problems. As a prediction model, it provides a set of overall prediction models and decision trees. It builds models in a stage-savvy approach, similar to previous improvement strategies, and summaries them by permitting arbitrarily distinguishable work [45]. In order to minimize the target work, in each cycle, we adjust the basic learners to the negative angle of the negative gradient, progressively increasing the expected value and adding it to the previously emphasized incentives:

$$f_p(X) = f_{p-1}(X) - \lambda_p \sum_{i=1}^{n} f_{p-1} L(\gamma_i - f_{p-1}(X_i)) \quad (2)$$

$$\lambda_p = \frac{Argmin}{\lambda} \sum_{i=1}^{n} L(\gamma_i - f_{p-1}(X_i) - \lambda f_{p-1} L(\gamma_i - f_{p-1}(X_i)) \quad (3)$$

where $L(\gamma_i f(X))$ is differentiable loss function.

D. DECISION TREE (DT)

Trees are a type of supervised machine learning in which data is split on a regular basis based on a parameter. In the training data, we define what the input is and what the related output is. The leaves symbolize the decisions or final results. This technique has a tree or structure like a flowchart, with the branches, leaves, nodes, and root node.

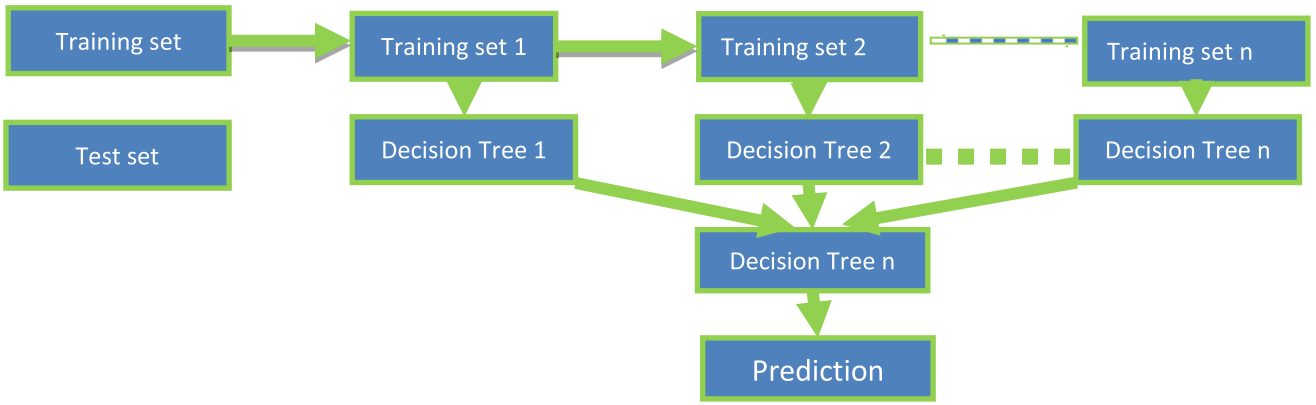


FIGURE 2. Random forest classifier flowchart.

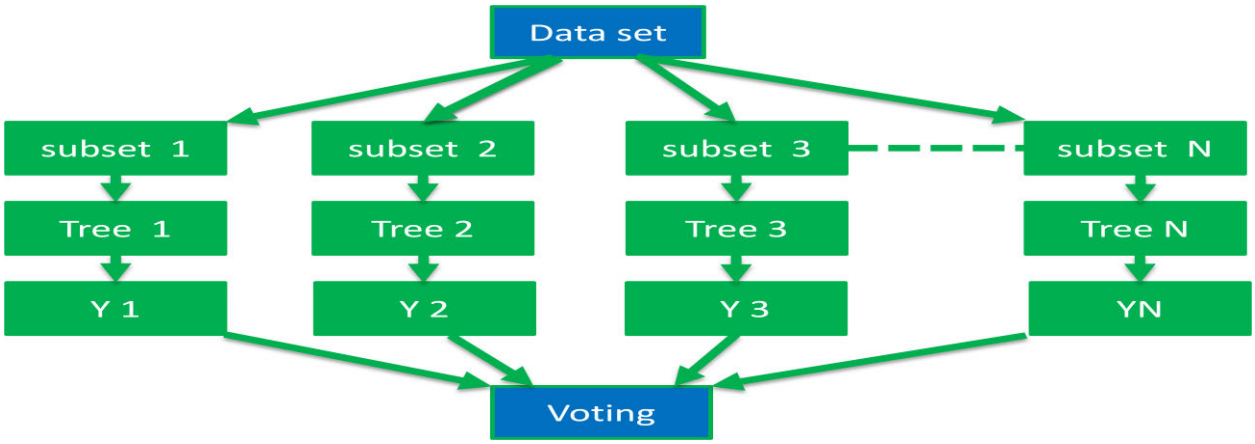


FIGURE 3. Decision tree classifier flowchart.

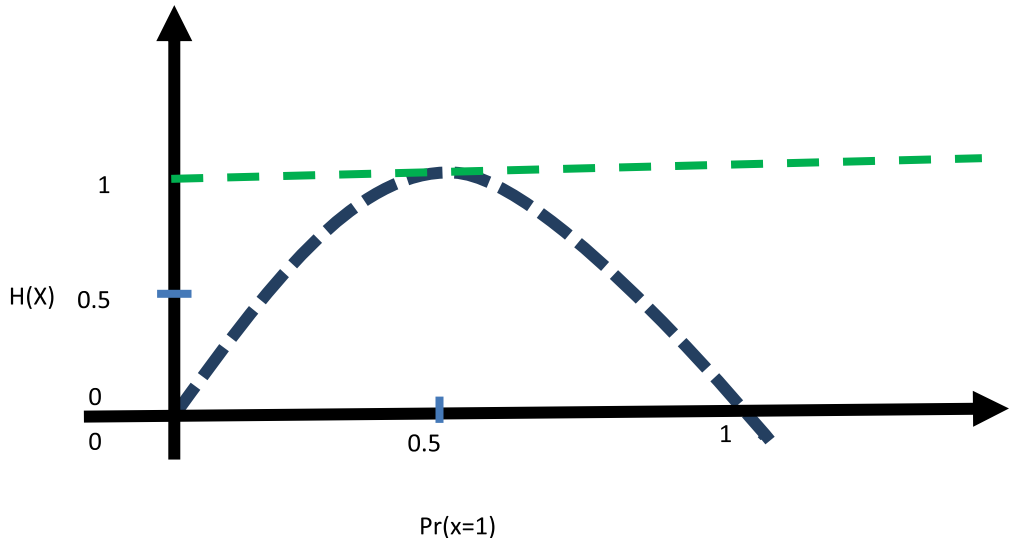


FIGURE 4. Entropy graph.

The features are kept in the internal nodes, whereas the branches indicate the outcomes of each test on each node. DT is frequently used for classification applications since it does not need considerable field experience or parameter setting [46]. The flowchart of Decision Tree is given in **Figure-3.**

E. ENTROPY(H)

The unpredictability of the evidence existence managed is measured by its entropy. The higher the entropy, the harder it is to draw any conclusions from the data. When the probability is either 0 or 1, H(X) has nil entropy. At what time the likelihood is either 0 or 1, it is said to be a 0 or 1 probability, the entropy H(X) is zero, as illustrated in **Figure-3**. When the probability is 0.5, the Entropy is greatest because it shows that the data is completely random and that there is no chance of properly deciding on the results.

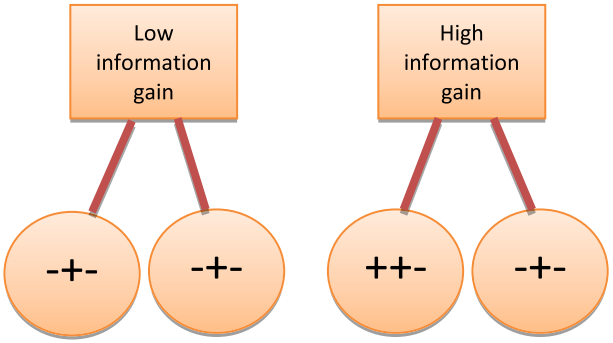


FIGURE 5. Information gain.

A leaf node is entropy-zero branches, whereas a branch with entropy larger than zero has to be divided further. Mathematically the entropy of a single property is expressed as:

$$Entropy = E\,(S) = \sum_{i=1}^{c} -p_i log_2\,(p_i) \tag{4}$$

In a state node, the probability of class i is pi S, where S denotes the present condition. Entropy for many characteristics is expressed mathematically as:

$$E(T, X) = \sum_{c \in X} P(c)E(c) \tag{5}$$

T denotes the current state, while X is the selected property. The Entropy (H) is given in **Figure-4**

F. INFORMATION GAIN (IG)

Information gain (IG) in **Figure-5** is a metric that measures how much information is gained and how successfully based on certain characteristics, training samples can be distinguished from its categorization goals. The key for creating the purpose of a decision tree is to select a characteristic that provides the most information with the least entropy [47].

As more knowledge is gained, entropy drops. Based on the provided attribute values, it estimates the change in entropy between, before and after dividing the dataset.

1) MATHEMATICALLY

$$\text{Information Gain (T, X)} = \text{Entropy (T)} - \text{Entropy (T, X)} \tag{6}$$

$$Information\ Gain = Entropy\,(before) - \sum_{j=1}^{K} Entropy(j, after) \tag{7}$$

"Before" refers to the dataset that existed before to the split, "K" to the quantity of subsets created (j, after) to subset j after the split, and (j before) to subset j before the split.

G. COST FUNCTION (GINI INDEX)

A cost function is the Gini index that may be used toward evaluate dataset splits. It is computed by taking one from the sum of each class's squared probability. Greater dividers are preferred because they remain mediators to construct, but smaller divisions with unique values are preferable for information gain.

$$Gini = 1 - [P_{(Class1)^2} + P_{(Class2)^2} + P_{(Class3)^2} + \ldots\ldots\ldots\ldots\ldots + P_{(ClassC)^2}]$$

$$Gini = 1 - \sum_{i=1}^{c} (P_i)^2 \tag{8}$$

The Gini Index is based on a categorical variable to be measured called "Success" or "Failure." It is the only way to performs binary divisions [48].

H. SUPPORT VECTOR MACHINE (SVM)

Support Vector Machine (SVM) is based on statistical learning theory [33] and it can classify both linear and nonlinear data. It divides the data into two groups by using support vectors and margins to create a linear optimal dividing Hyperplane inside a higher dimension (or classes). Using a suitable mapping that isn't linear, the original training data is transferred to a greater extent. In this context, a hyperplane may always be used to separate data into two classes.

If f is a support vector machine classification function then, f : P → Q, P stands for the domain (here i.e. data set),

$$P = [X, Y], \quad X = [x_i / 1 \le i \le n], \ Y = [y_i / 1 \le i \le n] \tag{9}$$

x_i is a set of n training tuples with yi as the associated class label. Each yi can have one of two values: +1 or -1, which indicate the first and second classes, respectively.

$$y_i \in \{+1, -1\} \tag{10}$$

The term "output set" can be used to describe a group of results.

$$Q_i = \{u / 1 \le i \le n\} \tag{11}$$

The main principle of SVM is to locate the hyperplane with the greatest margin to distinguish a collection from a series of bad examples, good instances, as shown in **Figure-5.**

A linear classification of the type is computed using a support vector machine:

$$f(x) = b + wx \tag{12}$$

A categorization that follows a straight line of the form f(x) = b + wx is computed using the support vector machine. Where w is a vector of weights, x represents an example of training and b represents bias. F(x) = 0 can be represented as the hyperplane separation. As a result, every point from

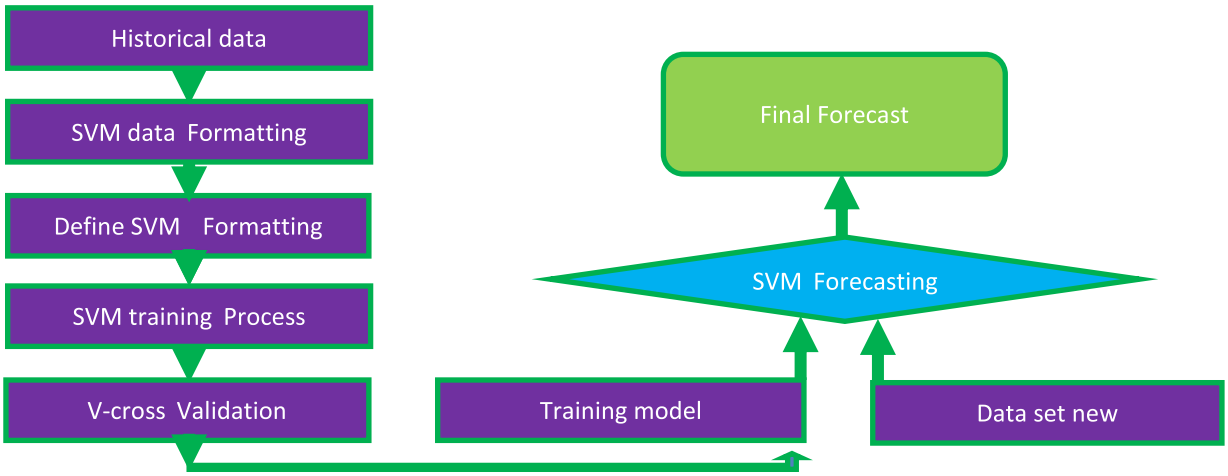


FIGURE 6. Support vector machine flowchart.

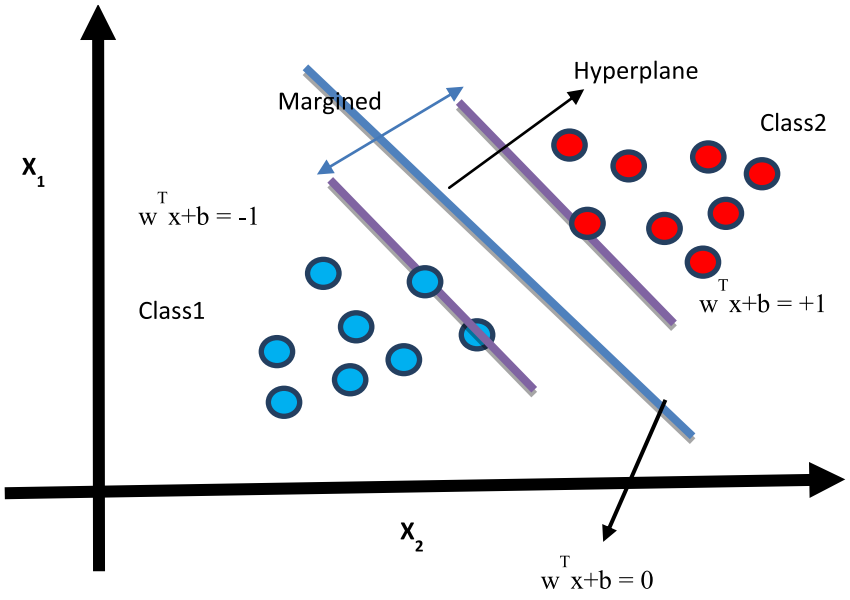


FIGURE 7. Optimal hyperplane with maximum margin.

one class that is above the separating hyperplane fulfills the condition f(x)>0. f(x)>0 is satisfied at the same time that any point from a different subject lies below the hyperplane that separates the two hyperplanes. The above equations were used to form the Set D is linearly separable, which satisfied the disparity,

$$y_i\,(f\,(x)) \geq 1, \quad \forall i \tag{13}$$

m is the margin in this case., $m = \frac{1}{\|w\|_2}$

Increasing profit margins may be stated in the shape of an issue of optimization using the above equation:

$$Min_{w,b}\,\frac{1}{2}\,\|w\|^2\,, \quad Subject\ to\ y_i\,(w.x+b) \geq 1,\ \forall i \tag{14}$$

The dual Lagrange multiplier can be used to address this optimization challenge.

$$\underset{\tilde{\alpha}}{\text{Min}}\,\Psi(\tilde{\alpha}) = \underset{\tilde{\alpha}}{\text{Min}}\,\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j(\vec{x}_i.\vec{x}_j)\alpha_i\alpha_j - \sum_{i=1}^{N}\alpha_i \tag{15}$$

The linearly split data support vectors are a subset of the actual training tuples. There is a dot product between the support vector x_i and the dot product between the support vector x_i and the dot product and the test tuple x_j in the Lagrangian formulation of the aforementioned optimization problem. Each Lagrange multiplier and each training tuple have a one-to-one connection. Not all data sets can be separated in a linear fashion. There may not be the positive and negative instances are separated by a hyperplane. SVMs may

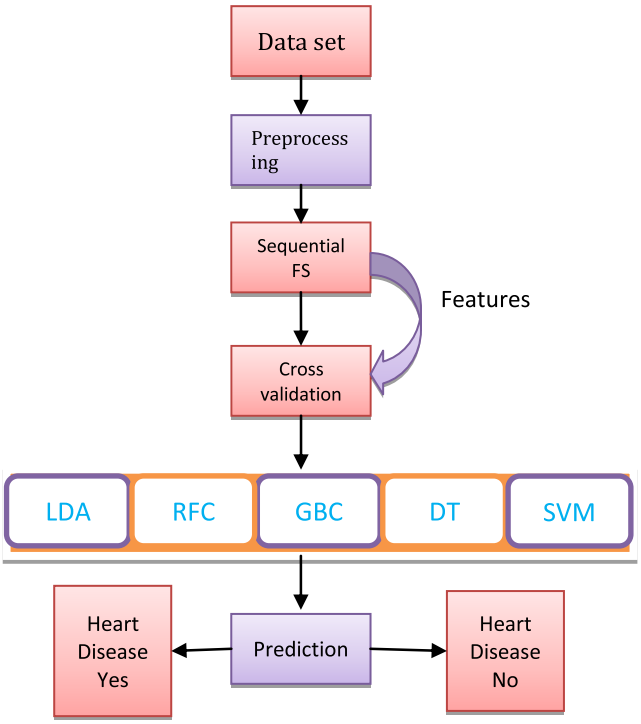


FIGURE 8. System framework for predicting disease (yes/no) from heart disease.

also be used to create non-linear classifiers. The Lagrange multiplier is used to compute the output of a non-linear SVM.

$$u = \sum_{j=1}^{N} y_i \, \alpha_j K\left(\vec{x_j}, \vec{x}\right) - b \tag{16}$$

K is the kernel function. In this case, we utilized the Radial Basis Kernel Function (RBF), which is written as follows:

$$\mathrm{K}\left(\mathrm{x_i}, \mathrm{x_j}\right) = \exp\left(-\gamma \left\|\mathrm{x_i} - \mathrm{x_j}\right\|^2\right), \quad \gamma > 0 \tag{17}$$

The quadratic form is altered by non-linearity, but the dual goal function Ψ i remain quadratic in α,

$$\begin{aligned} \mathrm{Min}_{\tilde{\alpha}}\, \Psi\left(\tilde{\alpha}\right) = \mathrm{min}_{\tilde{\alpha}}\, \frac{1}{2} \sum_{\mathrm{i}=1}^{\mathrm{N}} \sum_{\mathrm{j}=1}^{\mathrm{N}} \mathrm{y_i y_j} (\vec{\mathrm{x}}_i . \vec{\mathrm{x}}_j) \alpha_\mathrm{i} \alpha_\mathrm{j} \\ - \sum_{\mathrm{i}=1}^{\mathrm{N}} \alpha_\mathrm{i}\; 0 \le \alpha_\mathrm{i} \le \mathrm{C}, \quad \forall \mathrm{i} \\ \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{y_i} \alpha_\mathrm{i} = 0 \end{aligned} \tag{18}$$

optimization algorithm solves above quadratic programming problem. The flowchart Support Vector Machine (SVM) is given in **Figure-6**

III. MEDICAL DIAGNOSIS USING MACHINE LEARNING TECHNIQUE

Humans could never have envisioned the possibilities that machine learning has given to them. Machine learning is an artificial intelligence branch that allows machines to learn from examples [49] in order to evaluate how various models without relying on human judgment. The functioning of ML is discussed step by step, as shown in **Figure-8**.

A. PERFORMANCE METRICS

Different performance assessment criteria are employed in this investigation to assess the classifier's performance.

B. CORRELATION MATRIX

The correlation coefficients between factors can be represented by correlation matrix in a tabular form. The correlation between the two factors is shown in each cell in the table. The aggregate information is obtained by a correlation matrix, [50] as a sign of cutting-edge exams. Similar factors are shown in lines and segments of the correlation matrix, which is "square." In concept, the 1.00 line from the top left corner to the bottom right corner represents the corners, indicating that each component in any circumstance is inextricably linked to itself. The grid is balanced, and the main tilt, which is the same depiction of the tilt under the major tilt, shows a similar connection.

C. CORRELATION WITH TARGET VARIABLE

Feature determination is one of the most significant breakthroughs in any machine learning work. If a data set appears, an element will appear just one portion of the data set will be handled. Not every segment in any data set will have an effect on the yield variable. These superfluous elements are quite likely to be included in the model. This necessitates the determination of features. It is possible to describe embedded technology as iterative. It is capable of handling an individually cycle of model research and measurement, as well as carefully separating the functions that provide the largest contribution to specific attention research [51]. The regularization method, which penalizes components within a specific coefficient limit, is the most often utilized installation technique. The usage of lasso regularization features will be discussed here. When an element is unimportant, the lasso penalizes it by lowering the coefficient to zero, the feature is eliminated and the remaining features are used.

D. VALIDATION ACCURACY METRICS

The following is a description of validation to check the classifier's accuracy: assessment measures are employed to assess the classifier's performance. In this investigation, multiple performances each observation in the test set in the correct container. It is a 22 network since there are two rest categories. It also provides two valid classifier predictions and two non-benchmark predictions. The confusion matrix [52] is shown in **Table 3**.

The following conclusions may be drawn from the confusion matrix: True Positive (TP) return is far higher than expected (TP). We may deduce that the characteristics of people with heart diseases have been accurately stated and that the patients have heart disease. True Negative (TN) outcome is a large negative number (TN). We believe the individual is healthy and has been correctly recognized. We assume a person has been misdiagnosed with heart disease (a level 1 error) False Negative (FN). The expected outcome is a false

TABLE 2. Dataset description cardiac heart disease (CHD).

S.NO	Attribute Name	Description
1	Age	Age in years
2	Sex	Male/ Female
3	Cp	Pericarditis with a positive outcome
4	restbps	On admission to the hospital, resting blood pressure was measured in millimetres of mercury (mm of Hg).
5	Chol	cholesterol levels in the blood in milligrams per deciliter
6	Fbs	Blood sugar levels in the fasting state (higher than 120 mg/dl) 1 = true, 0 = false
7	restecga	Electrocardiographic findings during rest. 0 indicates that the waveform is normal, whereas 1 indicates that the waveform is abnormal.
8	thalach	Attained maximum heart rate.
9	exang	Exercise including angina. value: 1=yes, 0=no
10	oldpeak	Exercise-induced depression compared to rest
11	slope	The slope of the ST segment at maximal activity. 1 equals up sloping, 2 equals flat, and 3 equals down sloping.
12	Ca	Number of main vessels coloured by fluoroscopy (0-3)
13	Thal	Inherited blood disease in which your body produces less HB than it should. 3 indicate normal, 6 indicate a permanent abnormality, and 7 indicate a reversible fault.

negative (FN). We believe the diagnosis of heart disease is incorrect since the person does not have heart disease. The following is the classifiers accuracy that how accuracy reflects the categorization system's overall performance:

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP} \tag{19}$$

$$Recall = \frac{TP}{TP + FN}, \tag{20}$$

$$Precision = \frac{TP}{TP + FP} \tag{21}$$

$$f - measure = \frac{2 * Recall * Precision}{Recall + Precision} \tag{22}$$

We use assessment measures including accuracy, sensitivity, specificity, F1-score, and the area under the curve (AUC) of ROC charts to assess the efficacy of the suggested strategy. The percentage of correctly identified subjects is known as accuracy. The fraction of those who have the condition and test positive is known as sensitivity. The fraction of those who do not have the disease who test negative is known as specificity. The terms recall and sensitivity are interchangeable. The fraction of participants accurately recognized as positive

TABLE 3.

confusion matrix		
	Heart disease (have disease yes)	Heart disease (have disease=no)
Have disease (yes)	TP	FN
Have disease (no)	FP	FN

out of the total number of subjects identified as positive is known as precision. A harmonic mean of accuracy and recall is the f-measure. [53].

IV. DATA SET

The "Heart Disease Clinical Record Data Set 2020" is utilized by several researchers [54] and may be acquired via UCI machine learning online information mining archives. This data collection was utilized in this inspection research to develop a machine-learning-based heart disease framework. The UCI heart disease data collection comprises 1025 patients, 13 characteristics, and no missing values as an example. To identify heart disease, more relevant autonomous information functions and target yield markers are retrieved and applied. During the follow-up period, there are two types of objective classes to classify the patient's disease (yes/no) or alive from cardiac disease. As a result, there are $1025 * 13$ feature matrices in the retrieved data set. **Table-2** shows the complete data as well as descriptions of 1025 examples from the data set's 13 attributes.

A. DATA PRE-PROCESSING

Information pre-processing is necessary for successfully summarizing data and developing machine learning classifiers, and it should be created and tested as soon as possible. The data set has been pre-processed (for example, missing quality removal, standard scalar, and Minimax scalar) and may be utilized in the classifier [55]. The standard scalar ensures that each element's mean is 0, its variance is 1, and all of the elements' coefficients are comparable. Similarly, the ultimate aim of shift information in Minimax Scalar is that all functions be in the 0–1 range.

V. EXPERIMENTAL METHODOLOGY

The suggested framework was established to distinguish people who have the cardiac disease (yes/no). We attempted to demonstrate multiple machine learning models that fully determine the distribution and chosen properties of the heart disease data set in the suggested model. SFS is used to pick essential features and attempt to present classifiers on these characteristics for feature selection. The framework processes model approval and execution assessment metrics using the well-known machine learning classifiers LDA,

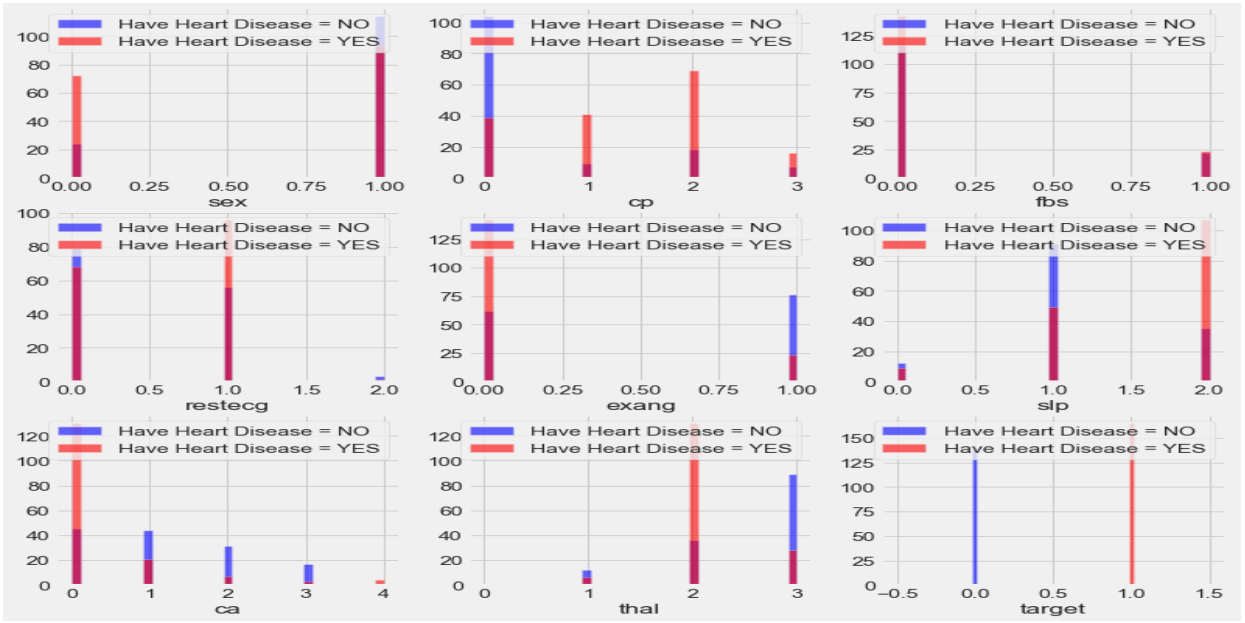


FIGURE 9. Attributes with Boolean values.

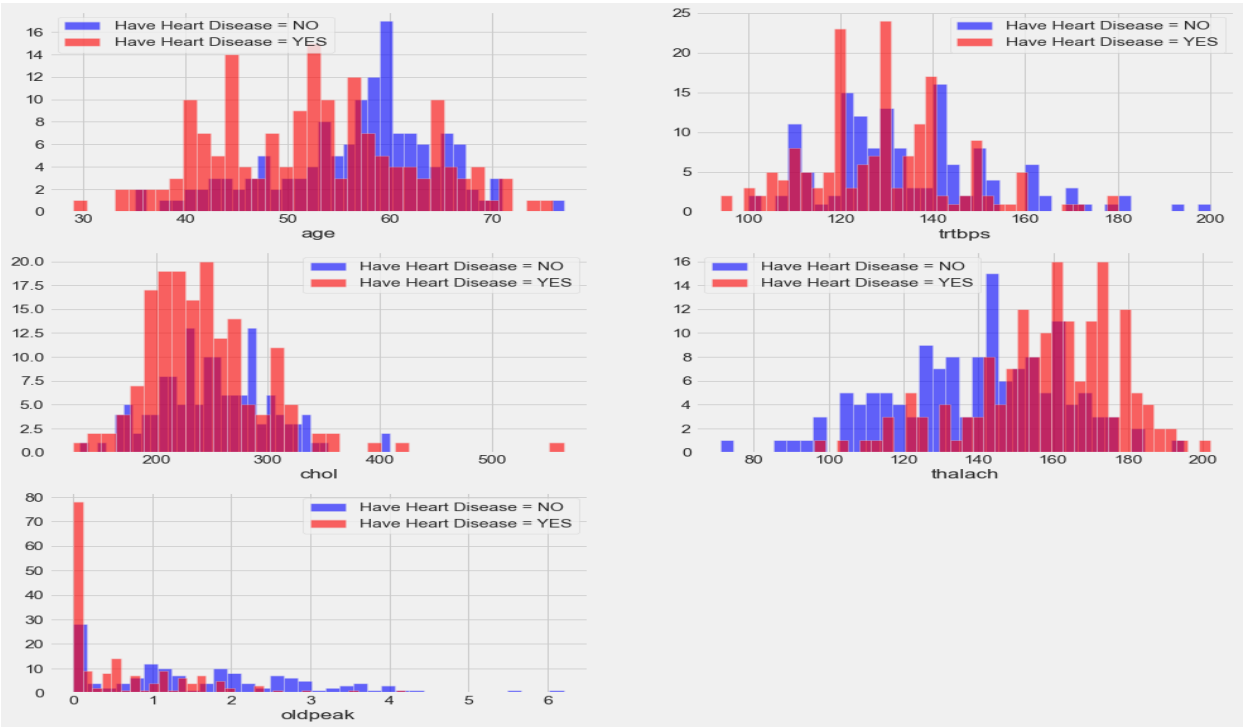


FIGURE 10. Attributes with continuous values.

RF, DT, GBC, and SVM. The experimental framework for predicting diseases cases owing to heart disease is shown in **Figure-8**. The suggested framework's strategy is broken into five stages, including Pre-processing of data sets, feature selection, cross-validation methods, machine learning classifiers, and assessment approaches for classifier representation.

A. EXPERIMENTAL SETUP

The study materials and techniques of the work are briefly discussed in the subsections that follow. All computations were done on an Intel(R) CoreTMi3-1800CPU @ 2.93 GHz PC using Python 3.8.8.

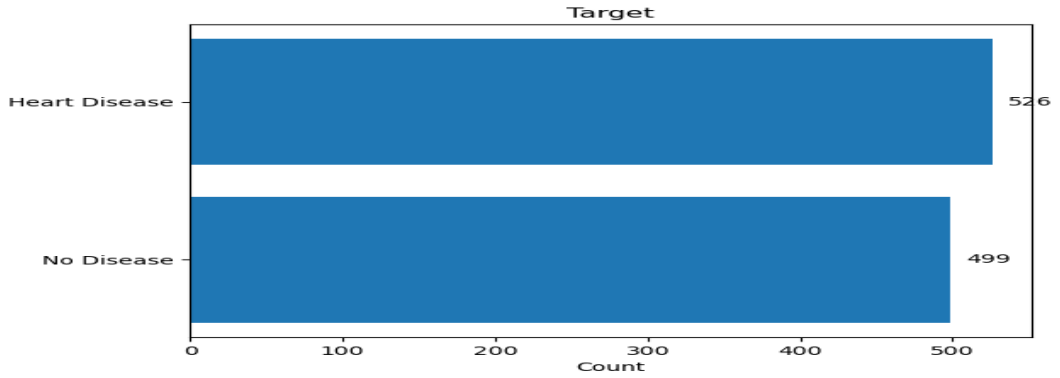


FIGURE 11. Target (heart disease patient or not disease patient).

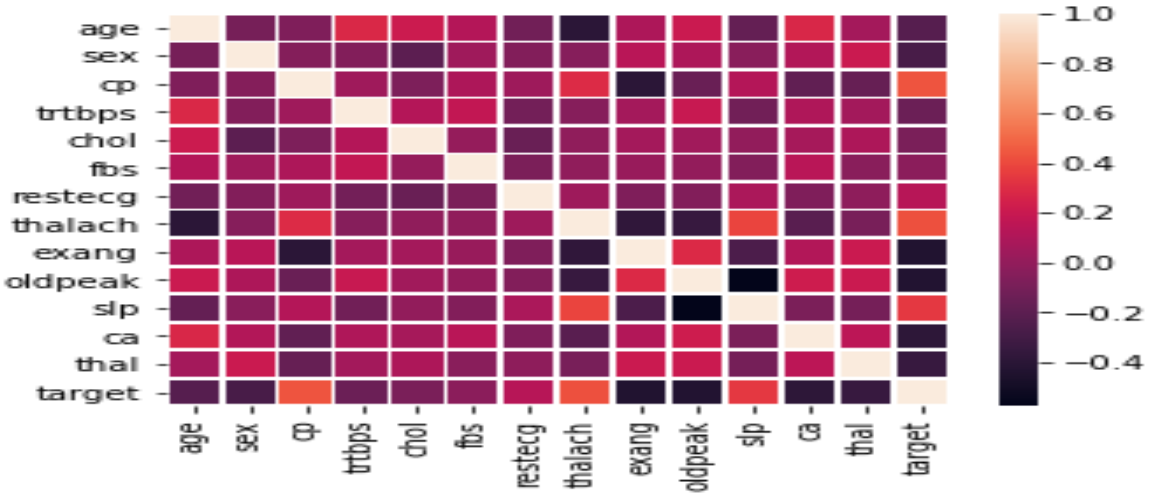


FIGURE 12. Correlation matrix.

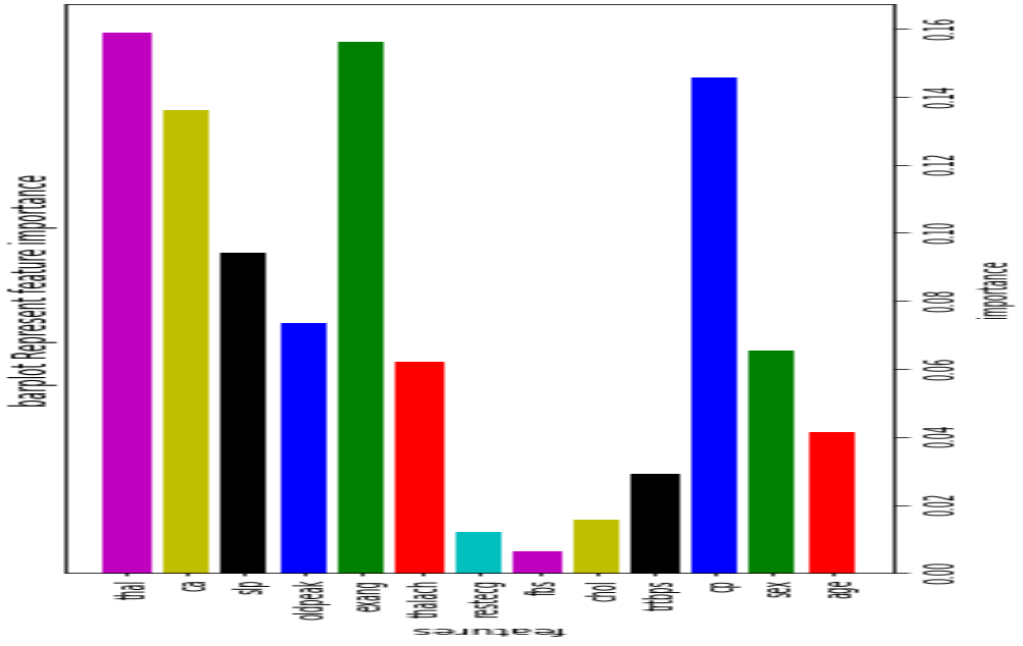


FIGURE 13. Correlation with the variable of interest.

TABLE 4. Attribute information.

Attribute name	Attribute Description	Unit measurement	Range
sex	Woman or man	binary	0-1
cp	There are four different forms of chest pain: (0) asymptomatic (1) non-angina discomfort (2) atypical angina (3) typical angina.	Boolean	0-1
fbs	The blood sugar level, which is displayed as 1 when the blood sugar level is 120 mg/dl, and 0 otherwise.	Boolean	0-1
restecg	According to Estes' criteria, this feature pertains to the reading of ECG value, which is 0 if normal, 1 if ST-T wave abnormalities, and 2 if definite or probable left ventricular hypertrophy.	Boolean	0-1
exang	During exercise agnosia detected	Boolean	0-1
slp	In the ST portion, the slope of the peak exercise is given as (0) to indicate uphill, (1) to indicate at, and (2) to indicate downhill.	Boolean	0-1
ca	With uroscopy, the number of major vessels (0–3) is coloured.	Boolean	0-1
thal	The heart status sign is a 3 to indicate _ne, a 6 to signal a permanent problem, and a 7 to suggest a reversible abnormality.	Boolean	0-1
target	If the patient died during the time period after that	Boolean	0-1
age	Age of the patient	years	35-75
trtbps	When a patient is admitted to the hospital, their blood pressure is measured in millimeters of mercury (mm of Hg).	mm of Hg	130-180
chol	A patient's serum cholesterol level in milligram's per deciliter.	mg /dl	100-600
thalach	This is the highest heart rate possible.	beat per minute (bpm)	150-190
oldpeak	When contrast to rest, activity causes ST depression.	mm	25-80

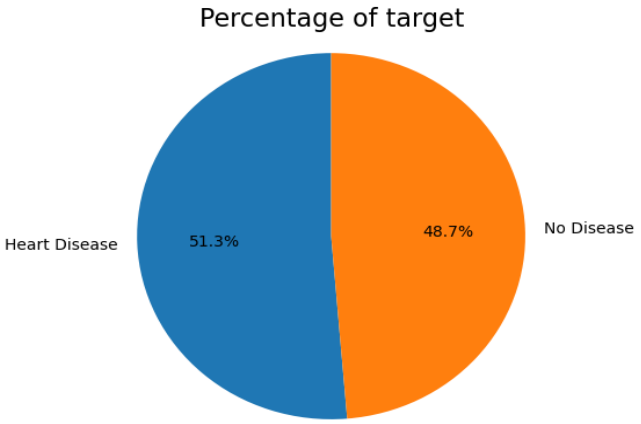


FIGURE 14. Percentage of target.

B. CROSS VALIDATION

In k-fold cross validation, the data is divided into k equal-sized portions, with k – 1 collection used to develop the classifier and the rest used to assess performance in each step [56]. The approval cycle is now known as the k times cycle, with the classifier running based on the k outcomes. Various k estimations are chosen for CV. We chose k=5 in our analysis since its appearance is appropriate. 70% of the information in the fivefold CV measurement is utilized for training, whereas 30 % is used for assessment. Before deciding to create and the loop are redefined several times, and all of the conditions in the training and test strings are randomly divided into the whole data set to test the new set for the new loop. Finally, the midway of the fivefold measurement was reached after conclusion of the fivefold measurement.

VI. RESULTS AND DISCUSSION

This section of the article discusses categorization models and their outcomes (from other perspectives). For the whole function of the heart disease clinical record data set, we first

TABLE 5. Classifiers accuracy with and without SFS.

S.NO	Algorithm name	Acc _test	Acc_train
1	Random Forest Classifier_	100.00	100.00
2	Random Forest Classifier_FS	94.81	94.70
3	SVC_sfs	90.58	87.73
4	Linear Discriminant Analysis_sfs	84.09	81.87
5	Random Forest Classifier_sfs	100.00	100.00
6	Gradient Boosting Classifier_	100.00	100.00
7	SVC_rbf	95.13	93.72
8	LinearDiscriminantAnalysis_	86.04	82.43
9	SVC_linear	85.71	83.26
10	Decision Tree Classifier_sfs	98.70	100.00
11	Gradient Boosting Classifier_sfs	97.73	99.16
12	SVC_poly	95.78	94.56
13	Decision Tree Classifier_	100.00	100.00

TABLE 6. k-fivefold cross validation.

K- fold cross validation (ROC_AUC)							
TECH-NIQUE	K=0	K=1	K=2	K=3	K=4	Mean (ROC_AUC)	Std. dev
LDA	0.86	0.82	0.82	0.82	0.74	0.81± 0.04	±1
RFC	0.96	0.97	0.94	0.96	0.93	0.95± 0.01	±1
DTC	0.94	0.97	0.93	0.97	0.93	0.95± 0.02	±1
GBC	0.99	0.99	0.97	0.99	0.98	0.98± 0.01	±1
SVM	0.86	0.79	0.82	0.77	0.73	0.80± 0.04	±1

examined the representations of various machine learning computations, such as linear feature analysis, random forest, decision tree, gradient boosting classifier, and support vector machine. Second, we compute SFS using element selection to determine relevant characteristics. Exhibitions are regarded as chosen features in the third category. The k-cross-validation approach is also utilized. Execution assessment measures are used to check the exhibition's classification. Before being applied to the classifier, all functions are standardized.

A. RESULT OF IMAGE ANALYSIS

People who have experienced a heart attack and died or survived during follow-up are included in this study [57].

TABLE 7. Two different heart disease dataset for models Classifier training & testing accuracy with and without SFS.

Datasets for Cleveland, Hungary, Switzerland & Long Beach V		Dataset Heart_ Statlog_ Cleveland_ Hungary		
Algorithm name	Acc _test	Acc_train	Acc _test	Acc_train
RF Classifier	100.00	100.00	91.04	100.00
SVC_sfs	90.58	87.73	83.75	86.79
LD Analysis_sfs	84.09	81.87	82.91	84.63
RF Classifier_sfs	100.00	**100.00**	88.52	**99.40**
GB Classifier	100.00	100.00	88.24	97.72
SVC_sf	95.13	93.72	85.15	90.88
LD Analysis	86.04	82.43	81.51	84.63
SVC_linear	85.71	83.26	80.95	85.11
Decision Tree Classifier_sfs	98.70	**100.00**	85.71	**99.76**
Gradient Boosting Classifier_sfs	97.73	99.16	83.75	93.40
SVC_poly	95.78	94.56	86.55	90.52
Decision Tree Classifier	100.00	100.00	83.19	100.00

Figure-3 shows a list of properties that have binary values of 1 or 0 (with or without). The qualities of sex, cp, fbs, restecg, exang, slp, ca, thal, and target are contained in this category. **Figure-9** shows the Attributes with Boolean values. **Figure-10** shows the Attributes with continuous values. **Figure-11** shows the target (heart disease patient or not disease patient). **Figure-12** shows the Correlation matrix. **Figure-13** shows **c**orrelation with the variable of interest. **Figure-14** shows the Percentage of the target.

- **Sex (woman or man)** Male patients (1) have a greater chance of mortality than female patients (0).
- **Cp** The discomfort in the chest can be classified into four types: asymptomatic (0), non-angina pain(1), atypical angina(2), and classic angina(3).
- **fbs:** if the blood sugar level is 120 mg/dl, it is represented as 1,and when it is not, it is shown as0.
- **restecg:** According to Estes' criteria, this feature pertains to the reading of ECG value, which is 0 if normal, 1 if ST wave abnormalities, and 2 if definite or probable left ventricular hypertrophy.
- **exan**: Agnosia was discovered after exercising..
- **slp:** The slope of the peak workout in the ST portion is denoted by the numbers 0 for uphill, 1 for at, and 2 for downhill.
- **ca:** With uroscopy, the number of major vessels (0–3) is coloured.
- **thal** The heart status sign is a 3 to indicate normal, a 6 to signal a permanent abnormality, and a 7 to suggest a reversible fault.

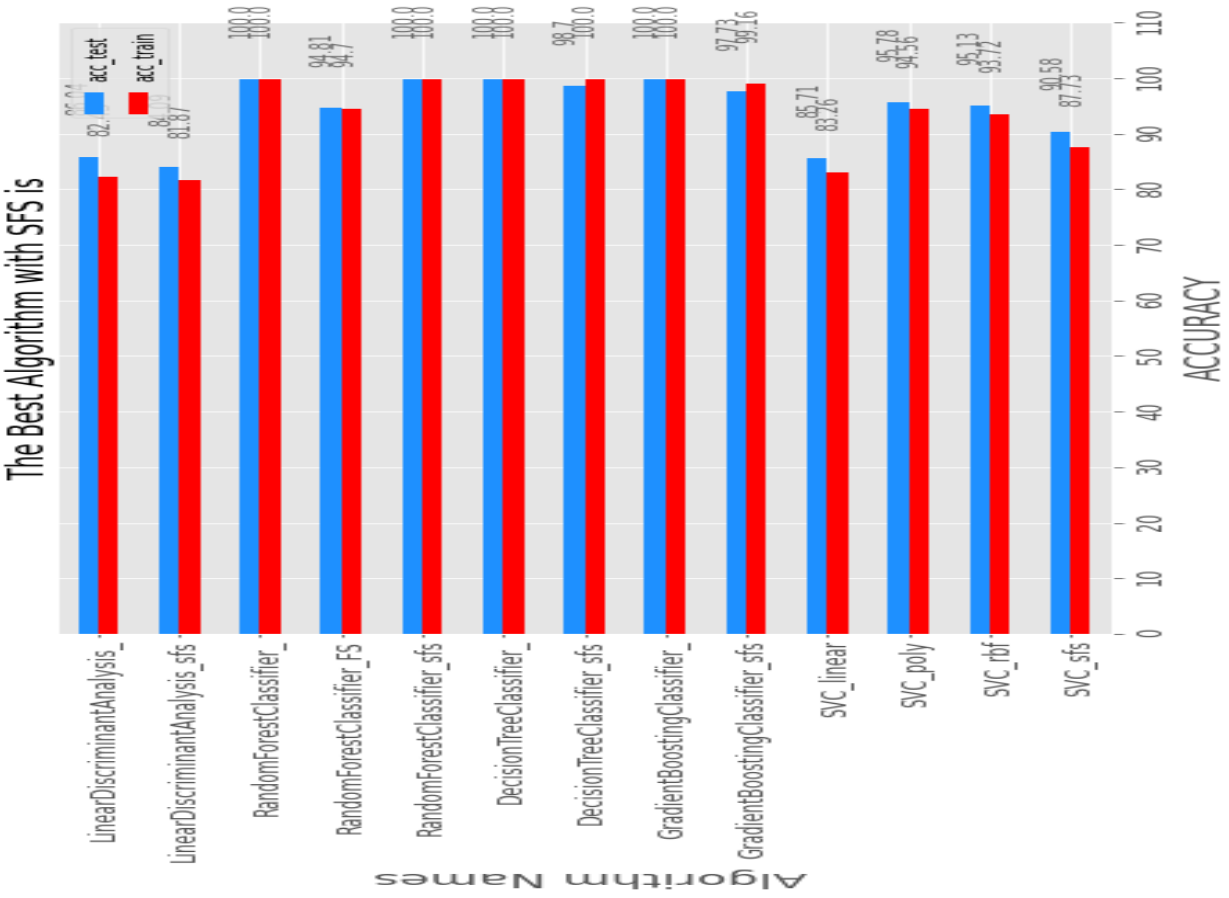
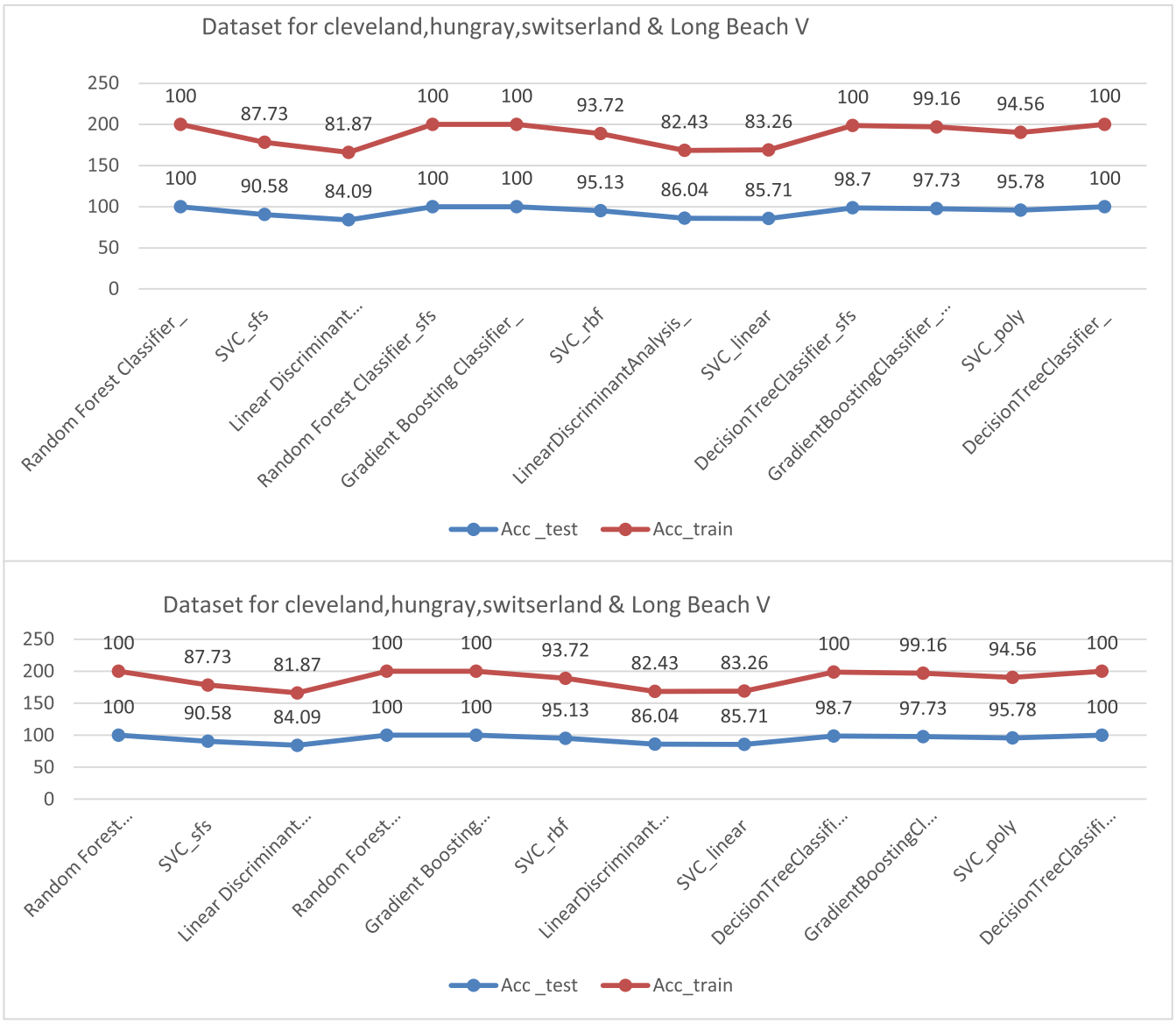


FIGURE 15. Performance of the classifiers with and without SFS.

- **target** If this value is 0, there is no risk of heart disease, but if it is between 0 and 1, there is a chance of heart disease.
- **Figure-4** depicts a continuous value characteristic that includes age, trtbps, chol, thalach, and oldpeak.
- **Age (years): In Figure-**4, the majority of patients died between the ages of 40 and 55, which is more risky during the follow-up period, and the majority of heart disease patients are above the age of 60.
- **trtbps** When a patient is admitted to the hospital, their blood pressure is measured in millimeters of mercury (120-140).
- **chol** patient's serum cholesterol level in milligram per deciliter (mg/dl)
- **thalach:** This is the highest heart rate possible.
- **oldpeak:** When compared to rest, exercise causes ST depreion.
- The link between qualities is shown by the correlation matrix in **Figure-5**.
- The higher the Positive number approaches 1, the more closely associated the feature is, whereas the negative value shows the negative correlation between characteristics, i.e., if one feature grows, other features drop, and vice versa. There is no relationship between the properties if the value is 0. Age, sex, trtbps, chol, and fbs are substantially connected with heart disease events in the graph below, whereas slp, ca, and thal are adversely correlated with target variables. Another indicator of feature relevance is the connection with the target variable. The four properties of restecg, chol, trtbps, and fbs, exhibit a low correlation with the target variable. Using various machine learning algorithms, it was determined that 526 people, or 51.3%, had a cardiac condition, whereas 499 people, or 48.7%, have no such abnormalities.The other variables, on the other hand, have a strong relationship with the target variable.

B. RESULT OF CLASSIFIERS (Fivefold CROSS Validation) WITH ALL FEATURES (n=13) AND WITH SELECTED FEATURES (SFS)

In this study, all features of the data set are focused on five machine learning classifiers using the fivefold cross-validation approach. Only 30% of the fivefold CV was used to train the classifier, whereas 70% was graded. Finally, the usual measurement result of the fivefold approach is achieved. The classifier has also passed several boundary

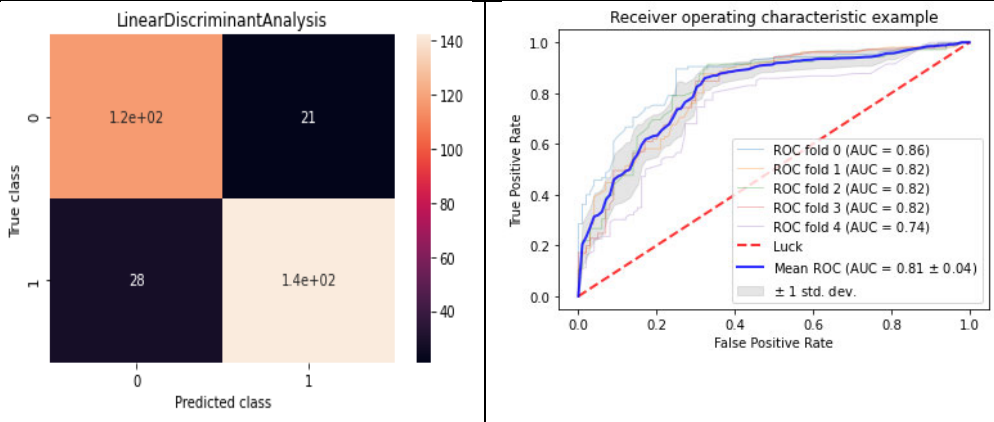
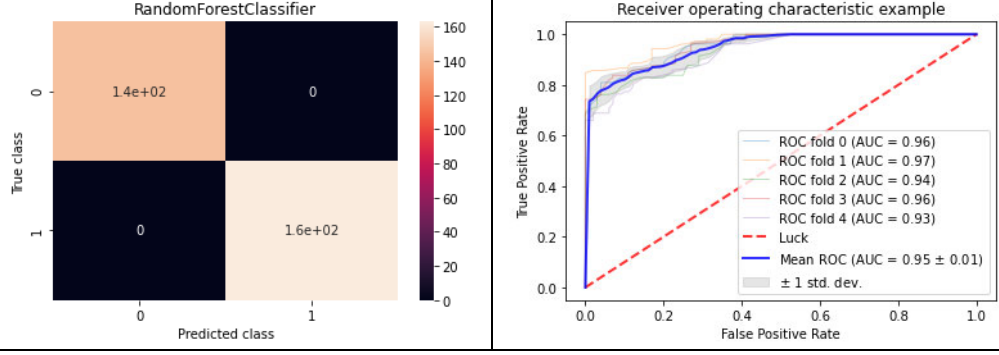


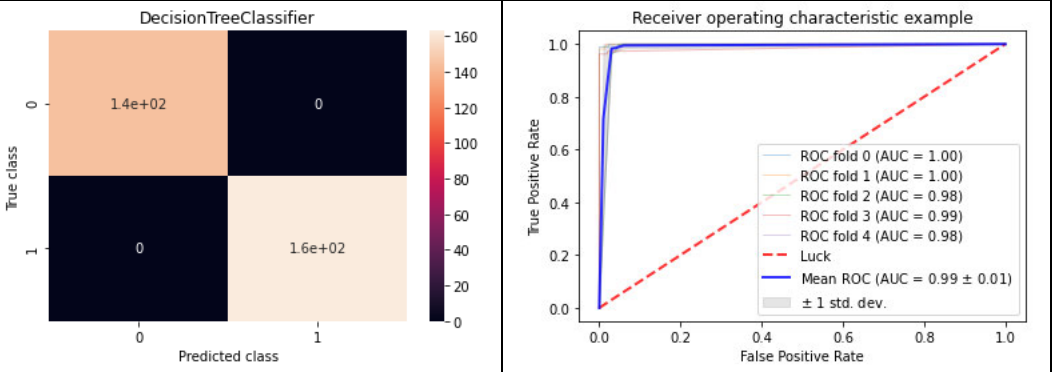
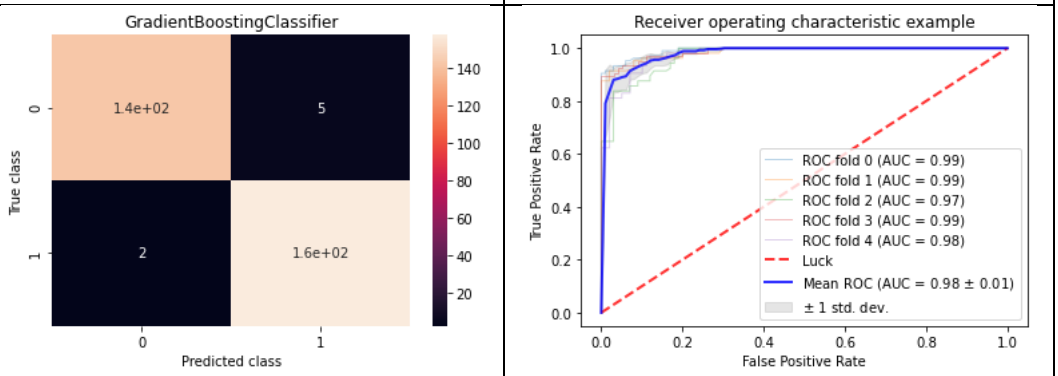
checks. **Table-3** shows the results of Cross-validation five times and sequential feature selection for five full-featured classifiers. **Table-3** indicates that the random forest classifier sequential feature selection both perform well, with 100% accuracy. The "random forest" and "Decision Tree" classifiers are the next two numbers, and their accuracy in all functions is 100%. We can distinguish between classifiers that use feature selection and those that don't, Random Forest Classifiers sfs, Decision Tree Classifier sfs, Gradient Boosting Classifier sfs.. As a result, support vector classifier sfs and Linear Discriminant Analysis sfs have declining accuracy of 100% 98.70%, 97.73%, 94.81%, 90.58%, and 84.09% respectively. Random Forest Classifier, Gradient Boosting Classifier, Decision Tree Classifier., SVC poly SVM rbf, Linear Discriminant Analysis, and SVC linear, on the other hand, have descending accuracy of 100%, 100%, 100%, 95.78%, 95.13%, 86.0%, and 85.71% respectively. Only random forest and gradient boosting have adequate invisibility for all fea for all characteristics, only random forest and gradient boosting provide enough invisibility elements, except for the random forest classifier and Decision Tree Classifier with feature selection. In the same sequence as the training and test data sets, **Figure-**7 demonstrates the presence of several classifiers.

The performance of the classifiers with and without SFS in which Random Forest Classifier Fs is the best among all the classifiers.

C. RESULTS OF VALIDATION METRICS

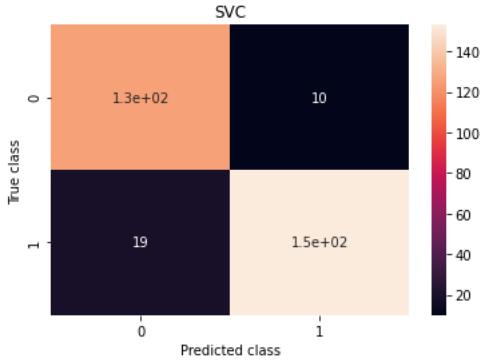
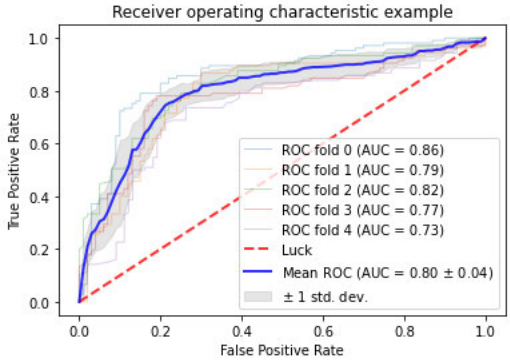
To compile and verify the results from the six classifiers, we have **Table-1**. In this table, it consists of different classifiers with selected important features. By observing the table, all five classifiers. Precision, recall, and f1-score have the

<table>
<tr><td>S.N.</td><td>Features</td><td></td></tr>
<tr><td>1</td><td rowspan="4">LDA</td><td>Sequential feature selection (SFS): sex , cp , trestbps , oldpeak , ca</td></tr>
<tr><td>Classification report</td></tr>
<tr><td>
<table>
<tr><td></td><td>Precision</td><td>recall</td><td>f1-score</td><td>support</td></tr>
<tr><td>0</td><td>0.85</td><td>0.81</td><td>0.83</td><td>145</td></tr>
<tr><td>1</td><td>0.84</td><td>0.87</td><td>0.85</td><td>163</td></tr>
<tr><td>Accuracy</td><td></td><td></td><td>0.84</td><td>308</td></tr>
<tr><td>macro avg</td><td>0.84</td><td>0.84</td><td>0.84</td><td>308</td></tr>
<tr><td>weighted avg</td><td>0.84</td><td>0.84</td><td>0.84</td><td>308</td></tr>
</table>
</td></tr>
<tr><td>Confusion matrix | ROC curve</td></tr>
<tr><td> 
</td></tr>
<tr><td>2</td><td rowspan="3">RFC</td><td>Sequential feature selection (SFS): Age , sex , cp , chol , thalach</td></tr>
<tr><td>Classification report</td></tr>
<tr><td>
<table>
<tr><td></td><td>precision</td><td>recall</td><td>f1-score</td><td>support</td></tr>
<tr><td>0</td><td>1.00</td><td>1.00</td><td>1.00</td><td>145</td></tr>
<tr><td>1</td><td>1.00</td><td>1.00</td><td>1.00</td><td>163</td></tr>
<tr><td>accuracy</td><td></td><td></td><td>1.00</td><td>308</td></tr>
<tr><td>macro avg</td><td>1.00</td><td>1.00</td><td>1.00</td><td>308</td></tr>
<tr><td>weighted avg</td><td>1.00</td><td>1.00</td><td>1.00</td><td>308</td></tr>
</table>
</td></tr>
<tr><td></td><td>Confusion matrix</td><td>ROC curve</td></tr>
<tr><td></td><td> 
</td></tr>
</table>

<table>
<tr><td>3</td><td>DTC</td><td>Sequential feature selection (SFS): age ,cp , chol , thalach , slope</td></tr>
<tr><td></td><td></td><td>Classification report
<table>
<tr><td></td><td>Precision</td><td>recall</td><td>f1-score</td><td>support</td></tr>
<tr><td>0</td><td>1.00</td><td>1.00</td><td>1.00</td><td>145</td></tr>
<tr><td>1</td><td>1.00</td><td>1.00</td><td>1.00</td><td>163</td></tr>
<tr><td>accuracy</td><td></td><td></td><td>1.00</td><td>308</td></tr>
<tr><td>macro avg</td><td>1.00</td><td>1.00</td><td>1.00</td><td>308</td></tr>
<tr><td>weighted avg</td><td>1.00</td><td>1.00</td><td>1.00</td><td>308</td></tr>
</table></td></tr>
<tr><td></td><td></td><td>Confusion matrix | ROC curve</td></tr>
<tr><td></td><td></td><td> 
</td></tr>
<tr><td>4</td><td>GBC</td><td>Sequential feature selection (SFS): age , cp , chol, oldpeak , thal</td></tr>
<tr><td></td><td></td><td>Classification report
<table>
<tr><td></td><td>precision</td><td>recall</td><td>f1-score</td><td>support</td></tr>
<tr><td>0</td><td>0.97</td><td>0.99</td><td>0.98</td><td>145</td></tr>
<tr><td>1</td><td>0.99</td><td>0.97</td><td>0.98</td><td>163</td></tr>
<tr><td>Accuracy</td><td></td><td></td><td>0.98</td><td>308</td></tr>
<tr><td>macro avg</td><td>0.98</td><td>0.98</td><td>0.98</td><td>308</td></tr>
<tr><td>weighted avg</td><td>0.98</td><td>0.98</td><td>0.98</td><td>308</td></tr>
</table></td></tr>
<tr><td></td><td></td><td>Confusion matrix | ROC curve</td></tr>
<tr><td></td><td></td><td> 
</td></tr>
</table>

same meanings as before, and these numbers authenticate the classifier's findings. The predictions of TP, FN, and NF are shown in the confusion matrix.

Different classifiers of heart patient disease during the follow-up period are TN and FP. The area under the true positive rate and the false positive rate is the ROC curve. The ROC

<table>
<tr><td>5</td><td>SVM</td><td colspan="2">Sequential feature selection (SFS): cp , restecg, thalach , ca , thal</td></tr>
<tr><td></td><td></td><td colspan="2">Classification report</td></tr>
<tr><td></td><td></td><td colspan="2">
<table>
<tr><th></th><th>precision</th><th>recall</th><th>f1-score</th><th>support</th></tr>
<tr><td>0</td><td>0.93</td><td>0.87</td><td>0.90</td><td>145</td></tr>
<tr><td>1</td><td>0.89</td><td>0.94</td><td>0.91</td><td>163</td></tr>
<tr><td>accuracy</td><td></td><td></td><td>0.91</td><td>308</td></tr>
<tr><td>macro avg</td><td>0.91</td><td>0.90</td><td>0.91</td><td>308</td></tr>
<tr><td>weighted avg</td><td>0.91</td><td>0.91</td><td>0.91</td><td>308</td></tr>
</table>
</td></tr>
<tr><td></td><td></td><td>Confusion matrix</td><td>ROC curve</td></tr>
<tr><td></td><td></td><td> 
</td><td> 
</td></tr>
</table>

AUC curve under fivefold cross validation is shown below. It produces diverse outcomes depending on the fold. To clear up any ambiguity, the average accuracy is determined. The higher ROC_ AUC obtained by the GBC classifier has an average accuracy of 98%.

VII. CONCLUSION

In this paper, a prediction strategy based on hybrid intelligent machine learning was developed to diagnose mortality during follow-up. Data from a database of heart disease clinical records were used to evaluate the approach. To choose significant characteristics, one of the most challenging difficulties in medicine is predicting disease sickness. Researchers used a range of algorithms; including LDA, RF, GBC, DT, SVM, and KNN, as well as the feature selection approach SFS, to predict cardiac illness. The system uses a K-fold cross-validation technique for verification. These six approaches were used in the comparison study. The Datasets for Cleveland, Hungary, Switzerland, and Long Beach V, as well as the Dataset Heart Statlog Cleveland Hungary, are used to assess the models' performance. For the Hungary, Switzerland & Long Beach V and Heart Statlog Cleveland Hungary Datasets, Random Forest Classifier sfs and Decision Tree Classifier sfs achieved the highest and very identical accuracy ratings (100%, 99.40% and 100%, 99.76% respectively). The findings were compared to previous research that focused on cardiac prediction. **Table- 4** & **7** shows that Gradient Boosting Classifier (GBC) findings have greater accuracy of 98% in terms of average ROC_AUC. To increase the classifer's classifcation accuracy, feature selection procedures should be utilized before classification, as shown in **Table-4**. We found two essential characteristics (restecg and chol) from which disease episodes may be predicted using feature selection (SFS). As a result, the SFS method can minimize computation time while also improving the classifier's classification accuracy. The SFS algorithm selects key features that help distinguish people who have disease events from those who are healthy.

The area of this exploratory effort is to generate a discovery framework that can expect when disease occurrences may occur. SFS computations, six classifiers, a cross-approval approach, and execution assessment measures are all used in the system. The analysis of heart disease will be more reasonable thanks to a machine learning technique for planning the choice of a decision support network. Furthermore, certain unrelated features degrade the model's performance and lengthen the computation time. As a result, this study's use of feature determination computations to choose the best qualities is another novel component. These top characteristics can shorten the classification model's execution time and enhance classification accuracy. Later, we'll do additional inspections to create these exhibitions of necessary classifiers for identifying heart diseases, employing various aspects (such as feature selection and streamlined methods).

We plan to expand the model in the future so that it may be used with a variety of feature selection strategies; another option is to use a random forest classifier. The main purpose of this research is to build on past work by inventing a new and distinctive model-creation approach, as well as to make the model relevant and easy to utilize in real-world settings.

REFERENCES

[1] J. L. Scully, "What is a disease," *EMBO Rep.*, vol. 5, no. 7, pp. 650–653, 2004.

[2] K. S. Reddy, V. Patel, P. Jha, V. K. Paul, A. S. Kumar, and L. Dandona, "Towards achievement of universal health care in India by 2020: A call to action," *Lancet*, vol. 377, no. 9767, pp. 760–768, Feb. 2011.

[3] V. D. Steen and T. A. Medsger, "Changes in causes of death in systemic sclerosis," *Ann. Rheumatic Diseases*, vol. 66, no. 7, pp. 940–944, Jul. 2007.

[4] P. Ponikowski, A. A. Voors, S. D. Anker, H. Bueno, J. G. Cleland, and A. J. Coats, "ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure: The Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC) Developed with the special contribution of the Heart Failure Association (HFA) of the ESC," *Eur Heart J.*, vol. 37, no. 27, pp. 200–2129, 2016.

[5] A. J. Aljaaf, D. Al-Jumeily, A. J. Hussain, T. Dawson, P. Fergus, and M. Al-Jumaily, "Predicting the likelihood of heart failure with a multi level risk assessment using decision tree," in *Proc. 3rd Int. Conf. Technol. Adv. Electr., Electron. Comput. Eng. (TAEECE)*, Apr. 2015, pp. 101–106.

[6] P. Croft, D. G. Altman, J. J. Deeks, K. M. Dunn, A. D. Hay, H. Hemingway, L. LeResche, G. Peat, P. Perel, S. E. Petersen, R. D. Riley, I. Roberts, M. Sharpe, R. J. Stevens, D. A. Van Der Windt, M. Von Korff, and A. Timmis, "The science of clinical practice: Disease diagnosis or patient prognosis? Evidence about 'what is likely to happen' should shape clinical practice," *BMC Med.*, vol. 13, no. 1, p. 20, Dec. 2015.

[7] E. Choi, M. T. Bahadori, A. Schuetz, W. F. Stewart, and J. Sun, "Doctor AI: Predicting clinical events via recurrent neural networks," in *Proc. Mach. Learn. Healthcare Conf.*, Los Angeles, CA, USA, Aug. 2016, pp. 301–318.

[8] K. Srinivas, B. Rani, and D. Govardhan, "Applications of data mining techniques in healthcare and prediction of heart attacks," *Int. J. Comput. Sci. Eng.*, vol. 2, pp. 250–255, Jan. 2010.

[9] S. Kaur, J. Singla, L. Nkenyereye, S. Jha, D. Prashar, G. P. Joshi, S. El-Sappagh, M. S. Islam, and S. M. R. Islam, "Medical diagnostic systems using artificial intelligence (AI) algorithms: Principles and perspectives," *IEEE Access*, vol. 8, pp. 228049–228069, 2020.

[10] D. Ntiloudi, G. Giannakoulas, D. Parcharidou, T. Panagiotidis, M. A. Gatzoulis, and H. Karvounis, "Adult congenital heart disease: A paradigm of epidemiological change," *Int. J. Cardiol.*, vol. 218, pp. 269–274, Sep. 2016.

[11] L. Yahaya, N. D. Oye, and E. J. Garba, "A comprehensive review on heart disease prediction using data mining and machine learning techniques," *Amer. J. Artif. Intell.*, vol. 4, no. 1, pp. 20–29, Apr. 2020.

[12] H. Eyre, R. Kahn, and R. M. Robertson, N. G. Clark, C. Doyle, Y. Hong, T. Gansler, T. Glynn, R. A. Smith, K. Taubert, and M. J. Thun, "Preventing cancer, cardiovascular disease, and diabetes: A common agenda for the American Cancer Society, the American Diabetes Association, and the American Heart Association," *Circulation*, vol. 109, no. 25, pp. 3244–3255, Jun. 2004.

[13] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, "Comparing different supervised machine learning algorithms for disease prediction," *BMC Med. Informat. Decis. Making*, vol. 19, no. 1, pp. 1–16, Dec. 2019.

[14] I. Olaronke and O. Oluwaseun, "Big data in healthcare: Prospects, challenges and resolutions," in *Proc. Future Technol. Conf. (FTC)*, Dec. 2016, pp. 1152–1157.

[15] A. L'Heureux, K. Grolinger, H. F. Elyamany, and M. A. M. Capretz, "Machine learning with big data: Challenges and approaches," *IEEE Access*, vol. 5, pp. 7776–7797, 2017.

[16] V. Sharma, S. Yadav, and M. Gupta, "Heart disease prediction using machine learning techniques," in *Proc. 2nd Int. Conf. Adv. Comput., Commun. Control Netw. (ICACCCN)*, Dec. 2020, pp. 177–181.

[17] W. Zhang and J. Han, "Towards heart sound classification without segmentation using convolutional neural network," in *Proc. Comput. Cardiol. Conf. (CinC)*, Sep. 2017, pp. 1–4.

[18] V. Chaurasia and S. Pal, "Skin diseases prediction: Binary classifcation machine learning and multi model ensemble techniques," *Res. J. Pharmacy Technol.*, vol. 12, no. 8, pp. 3829–3832, 2019.

[19] M. Marimuthu, M. Abinaya, K. S., K. Madhankumar, and V. Pavithra, "A review on heart disease prediction using machine learning and data analytics approach," *Int. J. Comput. Appl.*, vol. 181, no. 18, pp. 20–25, Sep. 2018.

[20] B. Brahmi and M. H. Shirvani, "Prediction and diagnosis of heart disease by data mining techniques," *J. Multidisciplinary Eng. Sci. Technol.*, vol. 2, no. 2, pp. 164–168, Feb. 2015.

[21] A. Javeed, S. Zhou, L. Yongjian, I. Qasim, A. Noor, and R. Nour, "An intelligent learning system based on random search algorithm and optimized random forest model for improved heart disease detection," *IEEE Access*, vol. 7, pp. 180235–180243, 2019.

[22] C. Yang, B. An, and S. Yin, "Heart-disease diagnosis via support vector machine-based approaches," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2018, pp. 3153–3158.

[23] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE Access*, vol. 7, pp. 81542–81554, 2019.

[24] M. Alex and P. S. P. Shaji, "Predictionand diagnosis of heart disease patients using data mining technique," in *Proc. Int. Conf. Commun. Signal Process. (ICCSP)*, Apr. 2019, pp. 0848–0852.

[25] S. Sandhiya and U. Palani, "An effective disease prediction system using incremental feature selection and temporal convolutional neural network," *J. Ambient Intell. Humanized Comput.*, vol. 11, no. 11, pp. 5547–5560, Nov. 2020.

[26] S. Raschka, J. Patterson, and C. Nolet, "Machine learning in Python: Main developments and technology trends in data science, machine learning, and artificial intelligence," *Information*, vol. 11, no. 4, p. 193, Apr. 2020.

[27] R. Aggrawal and S. Pal, "Sequential feature selection and machine learning algorithm-based patient's death events prediction and diagnosis in heart disease," *Social Netw. Comput. Sci.*, vol. 1, no. 6, pp. 1–16, Nov. 2020.

[28] D. Kumar, P. Carvalho, M. Antunes, R. P. Paiva, and J. Henriques, "Heart murmur classification with feature selection," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol.*, Aug. 2010, pp. 4566–4569.

[29] H. Shaheen, S. Agarwal, and P. Ranjan, "MinMaxScaler binary PSO for feature selection," in *Proc. 1st Int. Conf. Sustain. Technol. Comput. Intell.*, Singapore: Springer, 2020, pp. 705–716.

[30] S. Mokeddem, B. Atmani, and M. Mokaddem, "Supervised feature selection for diagnosis of coronary artery disease based on genetic algorithm," 2013, *arXiv:1305.6046*.

[31] A. M. Usman, U. K. Yusof, and S. Naim, "Cuckoo inspired algorithms for feature selection in heart disease prediction," *Int. J. Adv. Intell. Inform.*, vol. 4, no. 2, pp. 95–106, Jul. 2018.

[32] A. U. Haq, J. Li, M. H. Memon, M. Hunain Memon, J. Khan, and S. M. Marium, "Heart disease prediction system using model of machine learning and sequential backward selection algorithm for features selection," in *Proc. IEEE 5th Int. Conf. Converg. Technol. (I2CT)*, Mar. 2019, pp. 1–4.

[33] A. Javeed, S. S. Rizvi, S. Zhou, R. Riaz, S. U. Khan, and S. J. Kwon, "Heart risk failure prediction using a novel feature selection method for feature refinement and neural network for classification," *Mobile Inf. Syst.*, vol. 2020, pp. 1–11, Aug. 2020.

[34] 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, pp. 56–66, Oct. 2020.

[35] R. Aggrawal and S. Pal, "Sequential feature selection and machine learning algorithm-based patient's death events prediction and diagnosis in heart disease," *Social Netw. Comput. Sci.*, vol. 1, no. 6, pp. 1–16, Nov. 2020.

[36] V. Chaurasia and S. Pal, "Skin diseases prediction: Binary classifcation machine learning and multi model ensemble techniques," *Res. J. Pharmacy Technol.*, vol. 12, no. 8, pp. 3829–3832, 2019.

[37] Q. Ren, H. Cheng, and H. Han, "Research on machine learning framework based on random forest algorithm," in *Proc. AIP Conf.*, Mar. 2017, vol. 1820, no. 1, p. 080020.

[38] J. Son, I. Jung, K. K. Park, and B. Han, "Tracking-by-segmentation with online gradient boosting decision tree," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 3056–3064.

[39] D. Landgrebe, "A survey of decision tree classifier methodology," *IEEE Trans. Syst., Man Cybern.*, vol. 21, no. 3, pp. 660–674, May 1991.
[40] K. Mathan, P. M. Kumar, P. Panchatcharam, G. Manogaran, and R. Varadharajan, "A novel Gini index decision tree data mining method with neural network classifiers for prediction of heart disease," *Design Autom. Embedded Syst.*, vol. 22, no. 3, pp. 225–242, Sep. 2018.
[41] V. N. Vapnik, "An overview of statistical learning theory," *IEEE Trans. Neural Netw.*, vol. 10, no. 5, pp. 988–999, Sep. 1999.
[42] S. Marsland, *Machine. Learning: Algorithmic Perspective*. Boca Raton, FL, USA: CRC Press, 2015.
[43] N. J. Higham, "Computing the nearest correlation matrix—A problem from finance," *IMA J. Numer. Anal.*, vol. 22, no. 3, pp. 329–343, Jul. 2002.
[44] K. A. McColl, J. Vogelzang, A. G. Konings, D. Entekhabi, M. Piles, and A. Stofelen, "Extended triple collocation: Estimating errors and correlation coefcients with respect to an unknown target," *Geophys. Res. Lett.*, vol. 41, no. 17, pp. 6229–6236, Sep. 2014.
[45] W. L. Oberkampf and M. F. Barone, "Measures of agreement between computation and experiment: Validation metrics," *J. Comput. Phys.*, vol. 217, no. 1, pp. 5–36, Sep. 2006.
[46] A. I. Pritom, M. A. R. Munshi, S. A. Sabab, and S. Shihab, "Predicting breast cancer recurrence using effective classification and feature selection technique," in *Proc. 19th Int. Conf. Comput. Inf. Technol. (ICCIT)*, Dec. 2016, pp. 310–314.
[47] B. Chapman, A. D. DeVore, R. J. Mentz, and M. Metra, "Clinical profles in acute heart failure: An urgent need for a new approach," *ESC Heart Failure*, vol. 6, no. 3, pp. 464–674, Jun. 2019.
[48] H. Shaheen, S. Agarwal, and P. Ranjan, "MinMaxScaler binary PSO for feature selection," in *Proc. 1st Int. Conf. Sustain. Technol. Comput. Intell.*, Singapore: Springer, 2020, pp. 705–716.
[49] V. Chaurasia, S. Pal, and B. B. Tiwari, "Chronic Kidney disease: A predictive model using decision tree," *Int. J. Eng. Res. Technol.*, vol. 11, no. 11, pp. 1781–1794, Dec. 2018.
[50] T. Tantimongcolwat, T. Naenna, C. Isarankura-Na-Ayudhya, M. J. Embrechts, and V. Prachayasittikul, "Identification of ischemic heart disease via machine learning analysis on magnetocardiograms," *Comput. Biol. Med.*, vol. 38, no. 7, pp. 817–825, Jul. 2008.
[51] K. Schöttle and R. Werner, "Improving the most general methodology to create a valid correlation matrix," *Manage. Inf. Syst.*, vol. 9, pp. 701–710, Aug. 2004.
[52] R. Alzubi, N. Ramzan, H. Alzoubi, and A. Amira, "A hybrid feature selection method for complex diseases SNPs," *IEEE Access*, vol. 6, pp. 1292–1301, 2018.
[53] A. U. Haq, J. P. Li, M. H. Memon, J. Khan, A. Malik, T. Ahmad, A. Ali, S. Nazir, I. Ahad, and M. Shahid, "Feature selection based on L1-norm support vector machine and effective recognition system for Parkinson's disease using voice recordings," *IEEE Access*, vol. 7, pp. 37718–37734, 2019.
[54] G. G. N. Geweid and M. A. Abdallah, "A new automatic identification method of heart failure using improved support vector machine based on duality optimization technique," *IEEE Access*, vol. 7, pp. 149595–149611, 2019.
[55] 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.
[56] G. T. Reddy, M. P. K. Reddy, K. Lakshmanna, D. S. Rajput, R. Kaluri, and G. Srivastava, "Hybrid genetic algorithm and a fuzzy logic classifier for heart disease diagnosis," *Evol. Intell.*, vol. 13, no. 2, pp. 185–196, Jun. 2020.
[57] S. Koppu, P. K. R. Maddikunta, and G. Srivastava, "Deep learning disease prediction model for use with intelligent robots," *Comput. Electr. Eng.*, vol. 87, Oct. 2020, Art. no. 106765.
[58] C. Konstantinos and Siontis, "How will machine learning inform the clinical care of atrial fibrillation," *Circulat. Res.*, vol. 127, no. 1, pp. 155–169, Jun. 2020.