IJCI V9-02(2022) 25–44
International Journal of Computers and Information (IJCI)
Available online at https://ijci.journals.ekb.eg/

Prediction of Cardiovascular Disease Using Machine Learning Techniques

Shaimaa Mahmoud, Mohamed Gaber, Gamal Farouk, Arabi Keshk

Computer Science Department, Faculty of Computers and Information, Menoufia University,Shebin Elkom 32511, Egypt
sh.mahmoud600@gmail.com, m.gmalhat@yahoo.com ,gamal.farouk@ci.menofia.edu.eg, arabikeshk@yahoo.com

Abstract

Cardiovascular disease is one of the most dangerous diseases that lead to death. It results from the lack of early detection of heart patients. Many researchers analyzed the risk factors of cardiovascular disease and proposed machine learning models for the early detection of heart patients. However, these models suffer from the high dimensionality of data and need to be improved to obtain highly accurate results. In this paper, a practical proposal is presented that can predict whether a patient has cardiovascular disease or not. The proposal was tested using five different standard data sets from the UCI repository. Our proposal consists of two main processes: the first is the data preprocessing process, and the second is the prediction process. In data preprocessing, the data is prepared for the prediction process, and three different feature selection methods (e.g., PCA) are applied to select the most relevant features from the data. In the prediction process, fourteen different prediction techniques (for example, Random Forest (RF) and Support Vector Classifier (SVC)) were applied to over-employed datasets. The techniques used were evaluated using four evaluation metrics: accuracy, precision, recall, and F1-score. The experimental results show that the LASSO method as a feature selection method with RF as a prediction technique produced the best accuracy (100%). Accuracy (99.57%) was obtained for Decision Tree (DT), Gradient Boosting (GB), AdaBoost (AB), Decision Tree Bagging Method (DTBM), Random Forest Bagging Method (RFBM), K-Nearest Neighbors Bagging Method (KNNBM), AdaBoost Boosting Method (ABBM), and Gradient Boosting Boosting Method (GBBM). The accuracy of SVC, Logistic Regression (LR), Naïve Bayes (NB), and Support Vector Classifier Bagging Method (SVCBM) was very similar to each other (98.73%).

Keywords: Cardiovascular Disease, Prediction, Machine Learning Algorithm ;

1. Introduction

Heart disease is a type of disease that affects the heart or blood vessels. According to WHO data, cardiovascular disease is responsible for about 30% of deaths worldwide. [1]. There are several types of heart disease, such as cardiovascular disease, arrhythmias, heart failure, heart valve disease, cardiomyopathy, and congenital heart disease. [2]. Cardiovascular disease is considered the most important heart disease as it is the first cause of death both in the United States (source: American Heart Association, 2013) and worldwide (Data of the World Health Organization, 2013). [3]). The incidence of cardiovascular disease and its high mortality rate pose significant risks and burdens to health care systems around the world. Early detection of cardiovascular disease can help to minimize the disease's effects, perhaps lowering mortality rates.

Machine learning is a data analysis method that automates the construction of an analytic model [4]. It is based on the idea that systems can learn from data, identify patterns, and make decisions with minimal human intervention [4]. There are three frameworks for machine learning algorithms: supervised, semi-supervised, and unsupervised [5, 6]. When mapping input labels to output labels, such as image classification, facial recognition, sales forecasting, disease prediction, and spam detection, supervised models are used. Semi-supervised models are used when a small amount of labelled data with a large amount of unlabelled data is combined during training. Unsupervised models are used when the output of the data isn't known. So, it is used to find relevant patterns. For example, customer segmentation, anomaly detection in network traffic, and content recommendation.

Data pre-processing methods are used to prepare data for machine learning algorithms. There are different pre-processing methods such as data cleaning, data transformation, feature selection, missing value imputation, and redundancy elimination. The performance of machine learning methods is highly correlated to the efficiency of data pre-processing methods. For example, the accuracy of a machine learning model is degraded if the dataset contains missing values.

The supervised models are the most popular because they are used in both the classification and prediction processes. The supervised models were used for the early diagnosis of many diseases such as heart diseases, breast cancer, diabetes, liver disease, lung cancer, Parkinson's disease, and stroke. These models are trained using real-existing patient data to classify new patients. These models were adapted and utilized to be used for the early prediction of cardiovascular disease. In [1], the authors used random forest and logistic regression methods for the early detection of heart diseases. However, the obtained results are not accurate enough to be a reliable model. The author [7] proposed improving accuracy by using deep learning models. However, this model suffers from the high dimensionality of data.

In this paper, a comparative and analytical study between fourteen different supervised models was performed to evaluate and validate their accuracy results for the prediction of cardiovascular disease and compare these results with the results of the related work [1] published in 2021. These models are tested using five standard datasets from the UCI and evaluated using four evaluation metrics: accuracy, precision, recall, and F1-score. Moreover, the high dimensionality of data is solved by using three different feature selection methods. The results of using the employed models with and without feature selection approaches are compared. The experimental results showed that RF as a classification technique with PCA as a feature selection method achieves an accuracy of 97.05%, SVC can achieve 98.31%, and DT achieves an accuracy of 97.89%. In the proposed work, our contributions include:

- Testing a huge amount of data.
- Performing the necessary data pre-processing and cleaning by dealing with missing values and using standardization.
- Appling various feature selection methods (e.g., PCA) to build an effective framework for the early detection of heart disease patients.
- Making a comprehensive test with fourteen classifiers that include traditional and hybrid ones on a dataset combined from five standard datasets from the UCI repository.

This paper is organized as follows: In section 2, some related works in heart disease prediction are introduced. The research approach is presented in section 3. The implementation and results of our approach are presented in section 4. Conclusion and future work are put forward in section 5

2. **Related Work**

Heart disease prediction is a difficult task that demands both experience and information [7]. Therefore, there are many researchers interested in solving this problem by building models for the early diagnosis of heart problems based on patient-related characteristics. In this section, the current research papers on this topic are listed and identify their strengths and weaknesses.

Mohan, et al [7], the authors proposed a hybrid machine learning technique for the effective prediction of heart disease. The proposed technique improves the accuracy of the prediction of the cardiovascular disease model for early diagnosis of the disease and protects people's lives. The Naïve Bayes (NB) [8], Generalized Linear Model (GLM) [7], Linear Model (LM) [9,10], Deep Learning (DL) [11,12], Decision Tree (DT) [13], Random Forest (RF) [14], Gradient Boosted Trees (GBT) [15], and Support Vector Machine (SVM) [16,17] methods were implemented and compared. The dataset used in this work was collected from the UCI machine learning repository. There are four databases (i.e., Cleveland, Hungary, Switzerland, and the VA Long Beach) [18]. Several standard performance metrics such as accuracy, precision, and classification error have been considered for the computation of the performance efficacy of these techniques. From the obtained results, the authors selected hybrid RF and LM to propose a new hybrid method called Hyper Random Forest Linear Model (HRFLM). The author compared his work using HRFLM with other researchers' methodologies. It was found that the proposed hybrid model achieved better results in the case of using the accuracy, and classification error criterion in the evaluation, but it decreased when using precision, F-Measure, sensitivity, and specificity. HRFLM achieved an accuracy of about 88.4%.

Khan [19], the authors proposed an IoT framework for improving heart disease prediction based on the Modified Convolution Neural Network (MDCNN) classifier. The authors compared the performance of MDCNN with that of Deep Learning Neural Network (DLNN) [19] and Logistic Regression (LR) [9,10]. Data from the UCI machine learning repository, Framingham, Public Health, and Sensor Data [20,21] were used to train and evaluate the disease. Accuracy, precision, sensitivity, recall, and F1 Score metrics were used to evaluate the performance of the MDCNN and the other employed methods. It was found that the proposed model achieved the best results compared with other methodologies. The MDCNN achieves 98.2% accuracy. In contrast, the existing LR and DLNN have lower accuracy of 88.3% and 81.6%, respectively. However, this model suffers from the high dimensionality of data.

Li et al. [22] proposed a heart disease identification method using machine learning classification in E-Healthcare. Researchers study the impact of using two feature selection methods (i.e., Relief [23] and LASSO [24]) on the performance of six standard machine learning techniques. These techniques are SVM [16,17], Artificial Neural Network (ANN) [25], NB [8], DT [26], LR [9,10], and K-Nearest Neighbors (KNN) [27]. The Cleveland heart disease dataset [18], which was extracted from the UCI machine learning repository and contains 303 instances and 75 attributes, is used in this work. Accuracy, sensitivity, specificity, precision, and Matthews Correlation Coefficient (MCC) metrics are used to evaluate the performance of the employed techniques. The accuracy of SVM with their feature selection algorithm was achieved at 92.37%. However, this model suffers from the small size of the data.

Mienye, et al[28], the authors proposed an ensemble learning approach for the prediction of heart disease risk using a weighted ageing classifier ensemble. The authors compared the performance of ensemble classifiers with machine learning algorithms including (KNN), (LR), linear discriminant analysis (LDA), (SVM), classification and regression tree (CART), gradient boosting, and random forest. Two heart disease datasets are used, the Cleveland dataset [29] obtained from the University of California, Irvine (UCI) repository, and the Framingham dataset obtained from the Kaggle website [30]. There are 303 instances and 14 attributes in the former, while there are 4238 instances and 16 attributes in the latter. The Framingham dataset has missing attributes and has been preprocessed to make its machine learning compatible. Age, sex, cholesterol level, blood pressure, alcohol consumption, and diabetes are all included in both databases. The models' performance is measured using accuracy, precision, sensitivity, and F1 Score. With 93% accuracy, 96% precision, 91% sensitivity, and a 93% F1 score, the proposed method performed best. This model, however, has a high dimensionality of data.

Ghosh et al. [1] proposed an efficient prediction of cardiovascular disease using machine learning algorithms with Relief and LASSO feature selection methods. The following methods are used: DT [26], Gradient Boosting (GB) [33], KNN [27], RF [26], Decision Tree Bagging Method (DTBM) [34], Random Forest Bagging Method (RFBM) [35], K-Nearest Neighbors Bagging Method (KNNBM) [36], AdaBoost Boosting Method (ABBM) [37], and Gradient Boosting Boosting Method (GBBM) [38]. Researchers used a combined dataset from Cleveland, Long Beach VA, Switzerland, Hungarian, and Stat log datasets [18,39]. Accuracy, Precision, Recall, F1-score, False Positive Rate, False Negative Rate, and Negative predictive value metrics were used to evaluate the employed algorithms. Based on the result analysis, using RFBM and the Relief feature selection method achieved an accuracy above 90%. However, the obtained results are not accurate enough to be a reliable model.

Neloy et al. [40] proposed a novel machine learning model called the Weighted Average Ensemble that achieves a superior result by combining three standard machine learning techniques (Random Forest, Decision Tree, and Naive Bayes). Researchers used a combined dataset from Cleveland, Long Beach VA, Switzerland, Hungarian, and Stat log datasets [18,39]. The performance of the Weighted Average Ensemble model was evaluated using the following metrics: Accuracy, Precision, Recall, F1-score, Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) metrics. It was found that the average ensemble model's precision, recall, and F1-score are all 0.93. And, when compared to the other six algorithms, MAE, MSE, and RMSE of 0.07, 0.07, and 0.27, respectively, are the best performance results. However, the dataset has a limited number of data points. And the obtained results are not accurate enough to be a reliable model for heart disease prediction.

Table. 1 list the discussed related work in summary. It shows the year of publication, algorithms used, employed datasets, and accuracy achieved for each related work. From this table, these works suffer from the small size of the used datasets. Researchers didn't implement all available datasets about cardiovascular disease using available classifiers, and most authors have relied on the use of standard datasets about heart disease. These models suffer from the high dimensionality of data. The achieved accuracy wasn't stable on all classifiers used by the authors and was not accurate enough to be a reliable model.

Table 1. *Summary of the related work*

Year	Reference Number	Algorithms Used	Dataset	Best Accuracy Achieved	Advantages	Disadvantages
2019	**[7]**	Naïve Bayes, Generalized Linear Model, Logistic Regression, Deep Learning, Decision Tree, Random Forest, Gradient Boosted Trees, Support Vector Machine, VOTE, and HRFLM.	Cleveland, Hungary, Switzerland, and the VA Long Beach	88.4%	accuracy, and classification error criterion achieved good results in the evaluation.	The proposed hybrid model performance decreased when using precision, F-Measure, Sensitivity, and Specificity.
2020	**[19]**	Deep Learning Neural Network (DLNN), Logistic Regression (LR), and Modified Deep Convolutional Neural Network (MDCNN).	UCI machine learning repository, Framingham, Public Health, and Sensor Data	88.3%, 81.6%, and 98.2% when using LR, DLNN, and MDCNN respectively.	The MDCNN achieves 98.2% accuracy which is the best results compared with other methodologies.	This model suffers from high dimensionality of data.
2020	**[22]**	Support vector machine, Artificial neural network, Naïve bays, Decision tree, Logistic regression, and K-nearest neighbor.	Heart disease contains 303 instances and 75 attributes.	92.37%	Researchers presented a comprehensive test using machine learning classifiers and Artificial neural network	This model suffers from small size of data.
2020	**[28]**	k-nearest neighbor (KNN), logistic regression (LR), linear discriminant analysis (LDA), support vector machine (SVM), classification and regression tree (CART), gradient boosting, random forest, and ensemble learning model.	Cleveland dataset obtained from the University of California, Irvine (UCI) repository, and the Framingham dataset obtained from the Kaggle website.	93%, 96%, 91%, and 93% using accuracy, precision, sensitivity, and F1_Score respectively.	A huge amount of data are used and implemented different machine learning classifiers.	The model suffers from high dimensionality of data.
2021	**[1]**	(AB), (DT), (GB), (KNN),(RF), (DTBM), (RFBM), (KNNBM), (ABBM), and (GBBM)	Cleveland, Long Beach VA, Switzerland, Hungarian and Stat log	Above 90%	Researchers presented a comprehensive test with ten classifiers that include traditional and hybrid classifiers in addition using a dataset combined from five datasets.	The obtained results are not accurate enough to be a reliable model.
2022	**[40]**	Random Forest, Decision Tree, and Naive Bayes, and Weighted Average Ensemble model.	A combined dataset from Cleveland, Long Beach VA, Switzerland, Hungarian, and Stat log datasets are used.	precision, recall, and F1-score are all 0.93%. MAE, MSE, and RMSE of 0.07, 0.07, and 0.27, respectively.	A huge amount of dataset from the UCI repository are used. In addition, using different evaluation metric to evaluate the performance of their model.	the dataset has a limited number of data. And the obtained results are not accurate enough to be a reliable model for heart disease prediction.

3. Proposed Approach

In this section, the overall steps of our proposed model, the most important evaluation metrics used to evaluate the performance of the proposed model, and the machine learning classifiers are discussed.

3.1. *Overview of The Proposed Model.*

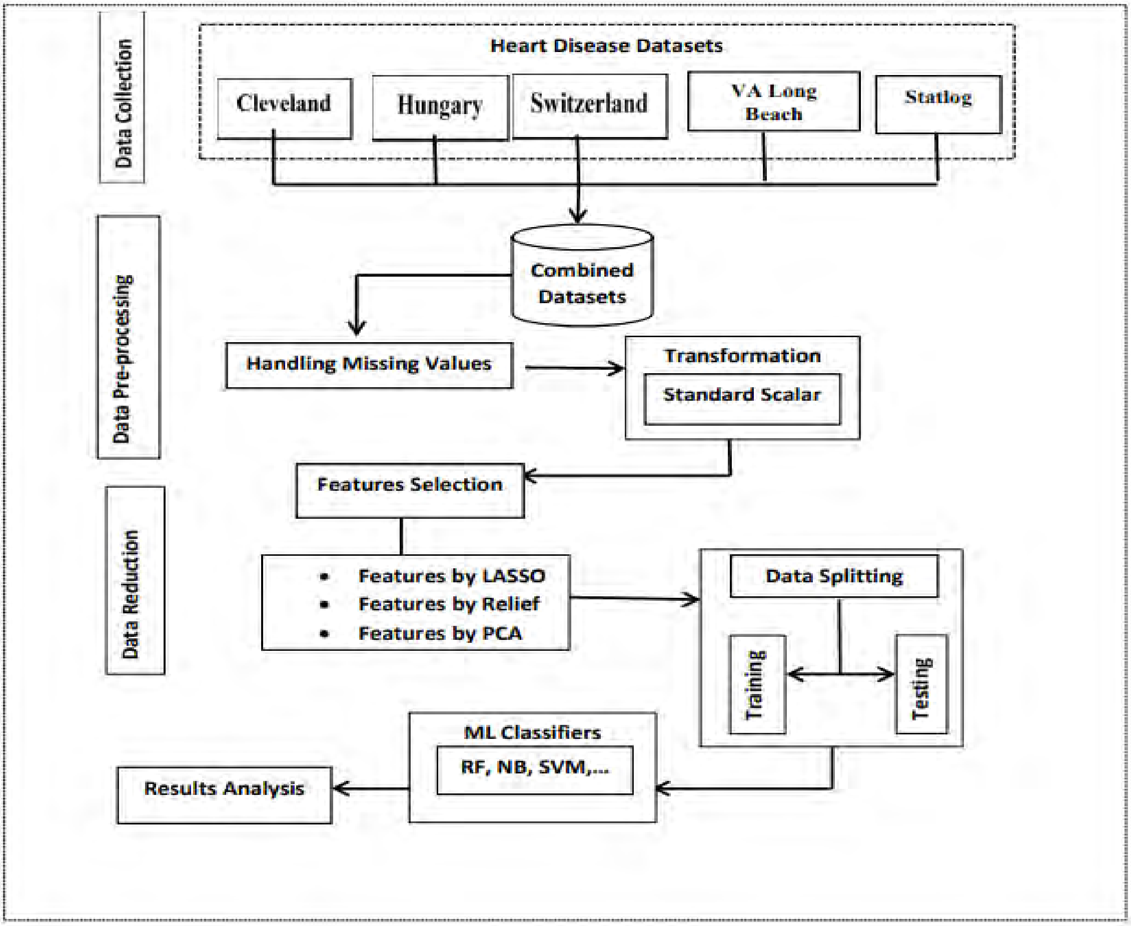

Fig.1. The architecture of our proposed model

As shown in fig.1. The data used consists of five data sets which are extracted from the UCI repository and then combined for processing [18,39]. In the data preprocessing stage, the collected data is analyzed to check the value of NaN and replace it with the best value. Missing values can be dealt with using a variety of strategies, including imputation and deletion. This problem is solved in our dataset by replacing all NaN values with the mean value.

After solving the missing values problem, standardization [41]is used by converting the data to a mean of 0 (μ) and a standard deviation (∑) of 1, and then the dataset is divided into two parts: training and testing. Where 80% of the data is allocated to the training phase, and the remaining 20% to the testing phase. To solve the overfitting problem, three different feature selection methods Relief, LASSO, and PCA are used to select the best features from the dataset by extracting the most relevant features based on rank values in medical references. Fourteen different machine learning classifiers are implemented such as Random Forest (RF) and Support Vector Machine (SVM).

Relief is a selection attribute approach that weights all of the dataset's features [42]. These weights can then be gradually increased. The goal is to make sure that the most significant elements have a large weight and that the other features have a small weight. To determine feature weights, Relief employs algorithms similar to those used by KNN. Kira and Rendell demonstrated this well-known method of feature selection approaches. Ri represents a randomly chosen instance. Relief looks for two of its closest neighbors: one from the same class, known as closest hit H, and one from the opposite class, known as closest miss M. The Ri, M, and H values are used to change the consistency calculation $W[A]$ for feature A. If there is a significant discrepancy between Ri and H, this is undesirable, and the performance value $W[A]$ is reduced. If the difference between Ri and M for attribute A is large, A can be utilized to differentiate various classes, and the weight $W[A]$ is increased. This operation will be repeated several times, where m is a variable that can be changed. Based on their ranking values, 10 features were selected: age in years (age), (sex), resting blood pressure in mm Hg (trestbps), serum cholesterol (chol), fasting blood sugar (fbs), resting electrocardiographic results (restecg), maximum heart rate (thalach), exercise induced angina (exang), chest pain type (cp), and the number of major vessels (0–3) coloured by fluoroscopy (ca).

Modifying the absolute value of the coefficient of functions is required for this operator's minimal selection and shrinkage functionality. Some of the features' coefficient values are zero, and features with negative coefficients can be excluded from the subset. For feature values with small coefficients, the LASSO performs well. The chosen subsets of features will include features with large coefficient values.

Unnecessary features can be found with LASSO [43]. Moreover, the reliability of this feature can be enhanced by repeating the above procedure many times, eventually taking the most frequently found features as the most important ones. This is called the randomized LASSO feature, which was introduced by Meinshausen and Buhlmann in 2010 and Wang in 2011.

Following the application of the LASSO feature selection algorithm to the used dataset, 11 features were chosen according to their ranking values: age in years (age), gender (sex), resting blood pressure in mm Hg (trestbps), serum cholesterol (chol), defect types (thal), slope of the peak exercise ST segment (slope), fasting blood sugar (fbs), resting electrocardiographic results (restecg), ST depression induced by exercise relative to rest (oldpeak), maximum heart rate (thalach), and chest pain type.

PCA is an unsupervised feature reduction method for projecting high-dimensional data into a new lower-dimensional representation that describes as much of the variation in the data as possible with the least amount of reconstruction error. A quantitatively accurate way of obtaining this reduction is Principal Component Analysis.

The primary component technique creates a new set of variables. The original variables are linearly combined in each main component. There is no redundant information because all of the primary components are orthogonal to one another. The principal components as a whole provide an orthogonal foundation for the data space. An unsupervised feature selection approach based on eigenvectors analysis is offered to find crucial original characteristics for the principal component.

The PCA feature selection technique is used with the help of the PCA class of the scikit-learn Python library. In the output, the number of principal components is selected. In our implementation, we used PCA to select the best 7 principal components from the used dataset.

3.2. *Performance Measure Indices.*

The evaluation metrics used for evaluating the employed classifiers are accuracy, precision, recall, and F1 score are discussed in this section. First, some terminologies are discussed:

1) *True Positive :-* Consider the time when the model's heart disease was accurately recognized.
2) *True Negative:-* When the model successfully identified the opposing class, such as patients who do not have any heart problems.
3) *False Positive:-* Refer to when the model incorrectly identified heart disease patients i.e., identifying non-heart disease patients as heart disease patients.
4) *False Negative:-* When the model wrongly identifies the opposite class, such as heart disease patients as normal patients.

- Accuracy refers to the proximity of the measurements to a specified value. The higher the accuracy value, the better the performance of the model used as defined in Eq. (1).

$$Accuracy = \frac{(True\ Positive + True\ Negative)}{(True\ Postive + True\ Negative + False\ Positive + False\ Negative)}\ [1] \qquad (1)$$

- Precision as defined in Eq. (2) quantifies the number of positive class predictions that actually belong to the positive class.

$$Precision = \frac{(True\ Positive\ )}{(True\ Postive + False\ Positive\ )}\ [1] \qquad (2)$$

- Recall quantifies the number of positive class predictions made out of all as defined in Eq. (3).

$$Recall = \frac{(True\ Positive\ )}{(True\ Postive + \ False\ Negative)}\ [1] \qquad (3)$$

In Eq. (4), the F1-score combines precision and recall in relation to a given positive class - The F1 score can be thought of as a weighted average of precision and recall, with 1 being the highest and 0 being the worst.

$$F1\text{-}score = \frac{2(Precision * Recall\ )}{(precision + Recall)}\ [1] \qquad (4)$$

3.3. *Overview of The Proposed Algorithms*

Machine learning is a type of data analysis that automates the creation of analytical models. It's a subset of artificial intelligence based on the concept that machines can learn from data, recognize patterns, and draw conclusions with little or no involvement from people [44].

The machine learning algorithms used in our suggested methodology are briefly explained in this section.

1) Random Forest

Random-forest is a supervised machine learning method that can be used for classification and regression [14]. The trees in the random forest run in a straight line. During the tree-building process, there is no interaction between these trees. It operates by training by constructing a huge number of decision trees. Then either the mean prediction (regression) or the category representing the mode of the categories (classification) is output. With certain useful adjustments, it aggregates the results of numerous predictions that aggregate many decision trees. As can be seen in fig.2, [45] By merging decisions from a set of basic models, the random forest can create predictions.

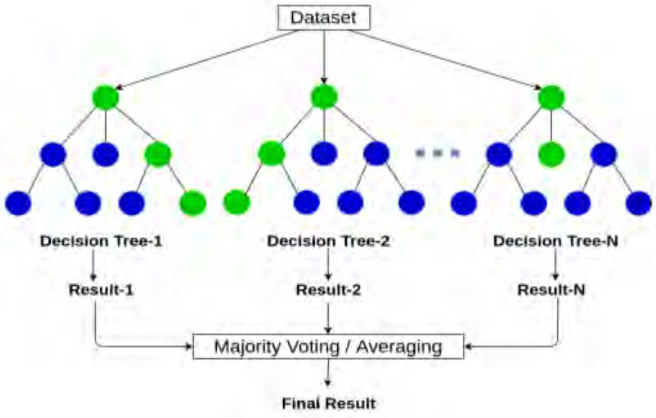

Fig. 2. Flow Chart of Random Forest Algorithm

2) Naïve Bayes

A Naive Bayes system is simple to create and does not require entangled iterative parameter estimation, making it particularly effective for large datasets [8]. From P(c), P(x), and P(x|c) (discussed below), Bayes' hypothesis provides a method for determining the returned likelihood, P (c|x) [15]. The impact of the estimation of an indicator (x) on a given class (c) is independent of the estimations of multiple predictors as specified in Eq. (5) [15], according to the Naive Bayes classifier.

- $P\ (c|x)$is the opportunity of class (target) given predictor (attribute).
- $P(c)$ is the preceding opportunity of class.
- $P(x|c)$ is the opportunity of predictor given class.
- $P(x)$ is the preceding opportunity of a predictor.

$$P\ (c|x) = \frac{P(x|c)P(c)}{P(x)} \quad [15] \qquad (5)$$

3) Decision Trees

One of the most significant supervised machine-learning classifiers used in both classification and regression applications is the decision tree. It reduces complex decision-making processes to simplified procedures. Fig. 3 shows an example of this. [46] From the root node, the tree develops by choosing a "Best Feature" or "Best Attribute" from a list of available attributes, then splitting. " The calculation of two metrics, "entropy" and "information gain," as defined in Eqs. (6) [13] and (7) [46], is usually used to select the "best attribute." The most useful information is provided by the 'best feature.' Entropy is a measure of a dataset's homogeneity, whereas information gain is the pace at which it increases or decreases.

$$E(D) = -P(positive)log_2P(positive) - P(negative)log_2P(negative) \ [13] \qquad (6)$$

Eq (6) calculates the Entropy E, of a dataset D, which holds the positive and negative ‘Decision Attributes’.

$$Gain(Attribute\ X) = Entropy(Decision\ Attribute\ Y) - Entropy(X, Y) \ [46] \qquad (7)$$

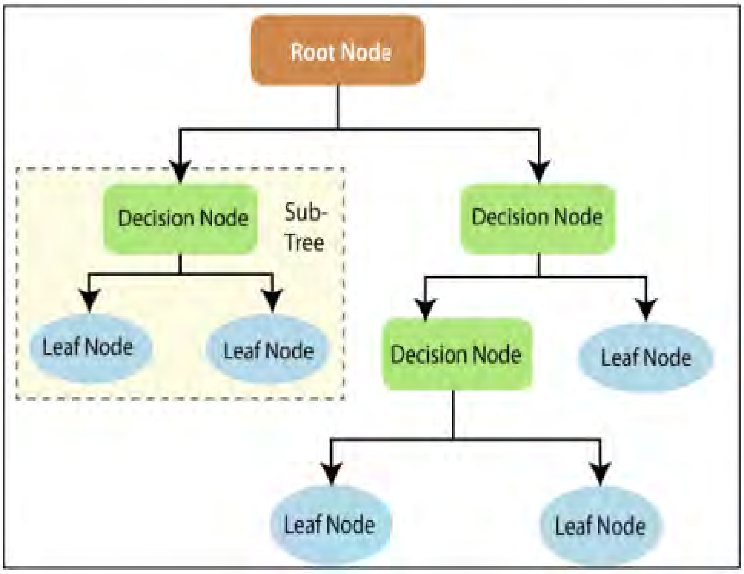

Fig. 3. Flow Chart of Decision Tree Classifier

4) *K-Nearest Neighbors*

In the field of machine learning, KNN is one of the most commonly used classification classifiers. It is nonparametric since it does not rely on data distribution assumptions. It takes into account the new data's equation with the old data and assigns the new data to the class that is closest to the existing classes. It uses Eq. 8. [47] to calculate the Euclidean distance between new A (x1, y1) data and previously accessible B (x2, y2) data.

$$d= \sqrt{(x_2 - x_1)^2 + (y_2 - y1)^2} \quad [47] \tag{8}$$

5) *Support Vector Machine*

SVM is a type of supervised machine learning classifier that is commonly used in classification problems [16,17]. As shown in fig.4, each data item is represented as a point in n-dimensional space (where n is the number of features), with the value of each feature being the value of a given coordinate. Then, using the information collected from the dataset [48], a prediction is produced.

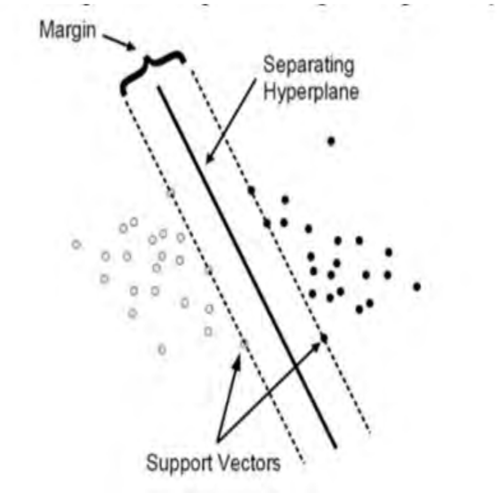

Fig.4. Example of support vector machine working

6) *Logistic Regression*

Logistic regression is a type of applied math analysis in which an information value is predicted based on previous data set observations [9,10]. Eq. (9). [49] and (10) [50] defines a logistic regression model to analyze the relationship between one or more existing independent factors to predict a dependent data variable. The following equation represents the model:

$$p(x) = e^{b_0+b_1x}/(1 + e^{b_0+b_1x}) \; [49] \tag{9}$$

It can be transformed into - $$\ln\left(\frac{p(x)}{1-p(x)}\right) = b_0 + b_1x \quad [50] \tag{10}$$

Where p(x), b_0, and b_1 refer to the predicted output, bias, or intercept term, and the coefficient for the single input value (x) respectively. The goal is to reduce the difference between predicted and actual data by using the values of coefficients b_0 and b_1 from the training data.

7) Gradient Boosting

Gradient boosting is a boosting approach for classification and regression problems that only takes 100 samples [51]. The improved loss function, the weak learner to generate predictions, and an additive model for merging the weak learners to minimize the loss function are the three essential components of graded reinforcement [52].

It enhances the efficiency of an algorithm by reducing overshoot. When the numbers in each class are out of balance, adding gradient tree augmentation to the Tobit model, dubbed the "Grabit" model, improves accuracy.

8) AdaBoost

Adaptive boosting, or AdaBoost, is a binary classification boosting strategy that combines multiple weak classifiers to generate a more robust classifier [53]. This algorithm calculates the anticipated accuracy based on 1000 samples. The training dataset instances are weighted with a beginning weight, as shown in Eq. (11) [54].

$$Weight\ (xi) = {}^{1}\!/_{N} \quad [54] \tag{11}$$

Where N is the training instance frequency, and xi is the training instance. For each input variable, the decision stump produces an output Eq. (12) is then used to get the misclassification rate.

$$Error = \frac{Correct-N}{N} \quad [54] \tag{12}$$

Where N represents the number of training instances. Combining numerous simple trainers to generate a more accurate prediction is known as boosting.

9) Bagging Method

To reduce the variance of decision tree classifiers, the bagging approach is applied. The goal is to use the training samples to divide the data into various subsets. [55] They employ randomly selected subsets of data to train their decision tree. After that, the average of all the forecasts from the various trees is applied. This is more powerful than a single decision tree classifier since it reduces overfitting while also handling higher-dimensional data appropriately. It resolves difficulties with data loss while keeping accuracy.

10) Boosting Method

Boosting is a repetitive technique that modifies the weight based on the previous prediction. Boosting creates effective predictive models in most cases [56]. It generates several loss functions and improves the performance of weak models by mixing them.

In this work, all the above algorithms are implemented to provide a comprehensive review of the results. Analyzing this review provides critical detail that helps us determine the most important algorithms that achieve good results in predicting heart disease using numerical data.

4. **Implementation and Results**

4.1. Machine Learning Libraries

The models in this research have been implemented using the Python language on the Jupiter notebook and many machine learning libraries such as Numpy, Pandas, Pyplot, and Sk-learn [57].

As mentioned before, the standard heart disease dataset collected from the UCI machine learning repository is used. As shown in Table.2, there are five databases: Cleveland, Hungary, Switzerland, VA Long Beach, and Statlog, that contain 303, 294, 123, 200, and 270 instances, respectively. Then these datasets are combined by concatenating them using Python. As shown in Table.3, attributes information in the dataset are mentioned.

Table 2. *Description of the datasets*

Dataset Name	**Number of Instances**	**Number of Attributes**	**Source of The Data**
Cleveland	303	14 including the predicted attribute	Cleveland Clinic Foundation
Hungary	294	14 including the predicted attribute	Hungarian Institute of Cardiology, Budapest
Switzerland	123	14 including the predicted attribute	V.A. Medical Center, Long Beach, CA
VA Long Beach	200	14 including the predicted attribute	University Hospital, Zurich, Switzerland
Statlog	270	14 including the predicted attribute	German Credit data

Table 3. *Attributes Information in dataset [46]*

No.	Attributes	Data Type	Description	Value Range
1	Age	Integer	Age in years	29 to 79
2	Sex		Gender instance	0 and 1
3	Cp		Chest pain type	1,2,3, and 4
4	Trestbps		Resting blood pressure in mm Hg	94 to 200
5	Chol		Serum cholesterol in mg/dl	126 to 564
6	Fbs		Fasting blood sugar > 120 mg/dl	0,1
7	Restecg		Resting ECG results	0,1, and 2
8	Thalach		Maximum heart rate achieved	71 to 202
9	Exang		Exercise induced angina	0,1
10	Oldpeak	Real	ST depression induced by exercise relative to rest	1 to 3
11	Slope	Integer	Slope of the peak exercise ST segment	1,2,3
12	Ca		Number of major vessels colored by fluoroscopy	0 to 3
13	Thal		Defect types	3,6,7
14	Num		Diagnosis of heart disease	0,1,2,3, and 4

After collecting data, some data preprocessing is performed. The data set is cleaned by replacing all the missing values with the best value after studying the data set. Before applying machine learning classifiers, data must also be standardized. Then feature selection methods (Relief, LASSO, and PCA) are applied to the data. The dataset is split with a ratio of 80:20, where 80% is used for training (952 records) while 20% is used for testing (238 records). Finally, traditional and hybrid machine learning classifiers such as RF, NB, DT, KNN, SVC, LR, GB, AB, DTBM, RFBM, KNNBM, SVCBM, ABBM, and GBBM are applied to the dataset.

4.2. Comparison Between Different Classifiers Using Accuracy

Accuracy is considered the most important technique to evaluate machine learning classifiers. As mentioned above, fourteen different machine learning classifiers are applied to the original 13 input features, then to the 11 input features selected by the LASSO method, 10 features selected with the Relief method, and finally to 7 features selected with the PCA method. Table 4. shows the accuracy of the different types of classifiers. Considering 13 features, the most accurate prediction (98.31%) was obtained from the use of all classifiers except KNN, LR, and KNNBM. KNNMB achieved an accuracy of 96.63%. KNN and LR had accuracy very similar to each other (96.21%).

Table 4. *Machine Learning Classifiers Accuracy*

Accuracy (%)				
Models	**Using ALL features (13)**	**Using 7 features (PCA)**	**Using 11 features (LASSO)**	**Using 10 features (Relief)**
RF	98.31	97.05	100	96.21
DT	98.31	97.89	99.57	94.53
KNN	96.21	96.63	96.63	90.33
SVC	98.31	98.31	98.73	95.79
LR	96.21	95.37	98.73	91.59
NB	98.31	98.31	98.73	91.59
GB	98.31	98.31	99.57	95.37
AB	98.31	98.31	99.57	94.53
DTBM	98.31	97.05	99.57	97.89
RFBM	98.31	98.31	99.15	96.21
KNNBM	96.63	97.05	96.63	91.17
SVCBM	98.31	98.31	98.73	94.53
ABBM	98.31	98.31	99.57	94.53
GBBM	98.31	98.31	99.57	95.37

The accuracy increased after decreasing the number of features to 11 (LASSO) features. All classifiers had better accuracy than for 13 features, except KNNBM, which had a similar result. When only evaluating 11 selected features (LASSO), the RF Classifier achieved the best accuracy (100%). Accuracy (99.57%) was obtained for DT, GB, AB, DTBM, RFBM, KNNBM, ABBM, and GBBM. The accuracy of SVC, LR, NB, and SVCBM was very similar to each other (98.73%). For the 10 (Relief) features, the accuracy decreased compared with 13 features using all classifiers. For the 7 (PCA) features, all classifiers achieved similar results compared to the use of 13 features except RF, DT, LR, DTBM, and KNNBM. (97.05%) accuracy was obtained in RF, DTBM, and KNNBM. (97.89%) and (95.37%) accuracy was obtained for DT and LR classifiers, respectively, which were less than compared with using 13 features.

4.3. *Comparison Between Different Classifiers Using Precision*

Precision has also been used to evaluate the performance of machine learning classifiers, as shown in Table. 5., the outcomes for precision are depicted. Considering 13 input features, a noticeable result of (98%) was obtained for precision with all classifiers except KNN, LR, and KNNBM. Both the KNN and LR achieved a precision score of 96%. KNNBM had a precision score of 97%. When applied to the 11 (LASSO) features, the best precision was obtained with both the RF, DT, GB, AB, BTBM, ABBM, and GBBM (100%). Both the KNN and KNNBM had the lowest precision (97%). Both the SVC, LR, NB, RFBM, and SVCBM achieved a precision score of (99%). For the 10 (relief) features, all classifiers except RF, KNN, LR, DTBM, and KNNBM had a precision score of 98%. Both the RF, KNN, DTBM, and KNNBM achieved a precision score of 97%. LR produced the lowest precision score (95%). In the case of PCA as a feature selection method, a noticeable result of 98% was obtained for precision with all classifiers except RF, KNN, LR, DTBM, and KNNBM. Both the RF, KNN, DTBM, and KNNBM had a precision score of 97%. LR produced the lowest precision score (95%).

Table. 5. *Machine Learning Classifiers Precision*

Precision (%)												
Models	**Using ALL features (13)**			**Using 7 features (PCA)**			**Using 11 features (LASSO)**			**Using 10 features (Relief)**		
	٠	**١**	**avg**	**٠**	**١**	**avg**	**٠**	**١**	**avg**	**٠**	**١**	**avg**
RF	0.96	1.00	0.98	0.94	1.00	0.97	1.00	1.00	1.00	0.94	1.00	0.97
DT	0.96	1.00	0.98	0.96	0.99	0.98	1.00	0.99	1.00	0.96	0.99	0.98
KNN	0.93	0.99	0.96	0.94	0.99	0.97	0.94	0.99	0.97	0.94	0.99	0.97
SVC	0.96	1.00	0.98	0.96	1.00	0.98	0.98	0.99	0.99	0.96	1.00	0.98
LR	0.92	1.00	0.96	0.91	0.99	0.95	0.97	1.00	0.99	0.91	0.99	0.95
NB	0.96	1.00	0.98	0.96	1.00	0.98	0.97	1.00	0.99	0.96	1.00	0.98
GB	0.96	1.00	0.98	0.96	1.00	0.98	0.99	1.00	1.00	0.96	1.00	0.98
AB	0.96	1.00	0.98	0.96	1.00	0.98	0.99	1.00	1.00	0.96	1.00	0.98
DTBM	0.96	1.00	0.98	0.95	0.99	0.97	0.99	1.00	1.00	0.95	0.99	0.97
RFBM	0.96	1.00	0.98	0.96	1.00	0.98	0.98	1.00	0.99	0.96	1.00	0.98
KNNBM	0.93	1.00	0.97	0.94	1.00	0.97	0.94	0.99	0.97	0.94	1.00	0.97
SVCBM	0.96	1.00	0.98	0.96	1.00	0.9 8	0.98	0.99	0.99	0.96	1.00	0.98
ABBM	0.96	1.00	0.98	0.96	1.00	0.98	0.99	1.00	1.00	0.96	1.00	0.98
GBBM	0.96	1.00	0.98	0.96	1.00	0.98	0.99	1.00	1.00	0.96	1.00	0.98

4.4. *Comparison Between Different Classifiers Using Recall*

The recall or sensitivity score is an important performance matrix because accurately classifying people with heart disease is critical. Table 6. shows the recall scores for the different classifiers and feature sets. Considering 13 input features, a noticeable result of 98% was obtained for recall with all classifiers except KNN, LR, and KNNBM. Both the KNN and LR achieved a recall score of 96%. KNNBM had a recall score of 97%. When applied to the 11 (LASSO) features, the best recall was obtained with both the RF, DT, GB, AB, DTBM, ABBM, and GBBM (100%).

Both the KNN and KNNBM had the lowest recall (97%). Both the SVC, LR, NB, RFBM, and SVCBM achieved a recall score of (99%). For the 10 (relief) features, the best recall was obtained using DTBM. Both the RF, SVC, and RFBM achieved a recall score of 96%. DT, GB, AB, SVCBM, ABBM, and GBBM had a recall score of 95%. Both the LR and NB had a recall score of 92%. KNNBM achieved

a recall score of 91%. KNN produced the lowest recall score (90%). In the case of PCA as a feature selection method, a noticeable result of 98% was obtained for recall with all classifiers except RF, KNN, LR, DTBM, and KNNBM. Both the RF, KNN, DTBM, and KNNBM had a recall score of 97%. LR produced the lowest recall score (95%).

Table. 6. *Machine Learning Classifiers Recall*

Recall (%)												
Models	**Using ALL features (13)**			**Using 7 features (PCA)**			**Using 11 features (LASSO)**			**Using 10 features (Relief)**		
	٠	**١**	**avg**	**٠**	**١**	**avg**	**٠**	**١**	**avg**	**٠**	**١**	**avg**
RF	1.00	0.97	0.98	0.94	1.00	0.97	1.00	1.00	1.00	1.00	0.93	0.96
DT	1.00	0.97	0.98	0.99	0.97	0.98	0.00	1.00	1.00	0.90	0.98	0.95
KNN	0.99	0.94	0.96	0.99	0.95	0.97	0.99	0.94	0.97	0.92	0.89	0.90
SVC	1.00	0.97	0.98	1.00	0.97	0.98	0.99	0.98	0.99	0.98	0.94	0.96
LR	1.00	0.93	0.96	0.99	0.92	0.95	1.00	0.98	0.99	0.98	0.85	0.92
NB	1.00	0.97	0.98	1.00	0.97	0.98	1.00	0.98	0.99	1.00	0.84	0.92
GB	1.00	0.97	0.98	1.00	0.97	0.98	1.00	0.99	1.00	0.99	0.92	0.95
AB	1.00	0.97	0.98	1.00	0.97	0.98	1.00	0.99	1.00	0.98	0.91	0.95
DTBM	1.00	0.97	0.98	0.99	0.95	0.97	1.00	0.99	1.00	0.99	0.97	0.98
RFBM	1.00	0.97	0.98	1.00	0.97	0.98	1.00	0.98	0.99	1.00	0.93	0.96
KNNBM	1.00	0.94	0.97	1.00	0.95	0.97	0.99	0.94	0.97	0.93	0.90	0.91
SVCBM	1.00	0.97	0.98	1.00	0.97	0.98	0.99	0.98	0.99	0.96	0.93	0.95
ABBM	1.00	0.97	0.98	1.00	0.97	0.98	1.00	0.99	1.00	0.98	0.91	0.95
GBBM	1.00	0.97	0.98	1.00	0.97	0.98	1.00	0.99	1.00	0.99	0.92	0.95

4.5. *Comparison Between Different Classifiers Using F1-Score*

The F1-score is the harmonic mean of the precision and recall scores. For the 13 features as shown in Table 7., all classifiers achieved an f1-score (98%). KNN and LR had the lowest f1-score (96%), and the result for the KNNBM was 97%. After decreasing the number of features using 11 (LASSO) features, the f1-score increased. All classifiers had a better f1-score than for 13 features. For the 10 (relief) features, the f1-score decreased compared with 13 features using all classifiers except DTBM, which had a similar result. For the 7 (PCA) features, all classifiers achieved comparable results compared to the use of 13 features.

Table. 7. M*achine Learning Classifiers F1-Score*

F1-Score (%)												
Models	**Using ALL features (13)**			**Using 7 features (PCA)**			**Using 11 features (LASSO)**			**Using 10 features (Relief)**		
	٠	**١**	**avg**	**٠**	**١**	**avg**	**٠**	**١**	**avg**	**٠**	**١**	**avg**
RF	0.98	0.98	0.98	0.97	0.97	0.97	1.00	1.00	1.00	0.96	0.96	0.96
DT	0.98	0.98	0.98	0.98	0.98	0.98	1.00	1.00	1.00	0.94	0.95	0.95
KNN	0.96	0.97	0.96	0.96	0.97	0.97	0.97	0.97	0.97	0.90	0.91	0.90
SVC	0.98	0.98	0.98	0.98	0.98	0.98	0.99	0.99	0.99	0.96	0.96	0.96
LR	0.96	0.96	0.96	0.95	0.96	0.95	0.99	0.99	0.99	0.92	0.91	0.92
NB	0.98	0.98	0.98	0.98	0.98	0.98	0.99	0.99	0.99	0.92	0.91	0.92
GB	0.98	0.98	0.98	0.98	0.98	0.98	1.00	1.00	1.00	0.95	0.95	0.95
AB	0.98	0.98	0.98	0.98	0.98	0.98	1.00	1.00	1.00	0.95	0.95	0.95
DTBM	0.98	0.98	0.98	0.97	0.97	0.97	1.00	1.00	1.00	0.98	0.98	0.98
RFBM	0.98	0.98	0.98	0.98	0.98	0.98	0.99	0.99	0.99	0.96	0.96	0.96
KNNBM	0.96	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.91	0.91	0.91
SVCBM	0.98	0.98	0.98	0.98	0.98	0.98	0.99	0.99	0.99	0.94	0.95	0.95
ABBM	0.98	0.98	0.98	0.98	0.98	0.98	1.00	1.00	1.00	0.95	0.95	0.95
GBBM	0.98	0.98	0.98	0.98	0.98	0.98	1.00	1.00	1.00	0.95	0.95	0.95

Table. 8. Confusion matrix for all classifiers

confusion matrix																
Models	**Using ALL features (13)**				**Using 7 features (PCA)**				**Using 11 features (LASSO)**				**Using 10 features (Relief)**			
	TP	TN	FP	FN	TP	TN	FP	FN	TP	TN	FP	FN	TP	TN	FP	FN
RF	105	0	129	4	105	0	126	7	114	0	124	0	114	0	115	9
DT	105	0	129	4	104	1	129	4	113	1	124	0	103	11	122	2
KNN	104	1	125	8	104	1	126	7	113	1	117	7	105	9	110	114
SVC	105	0	129	4	105	0	129	4	113	1	122	2	112	2	116	8
LR	105	0	124	9	104	1	123	10	114	0	121	3	112	2	106	18
NB	105	0	129	4	105	0	129	4	114	0	121	3	114	0	104	20
GB	105	0	129	4	105	0	129	4	114	0	123	1	113	1	114	10
AB	105	0	129	4	105	0	129	4	114	0	123	1	112	2	113	11
DTBM	105	0	129	4	104	1	127	6	114	0	123	1	113	1	120	4
RFBM	105	0	129	4	105	0	129	4	114	0	122	2	114	0	115	9
KNNBM	105	0	125	8	105	0	126	7	113	1	117	7	106	8	111	13
SVCBM	105	0	129	4	105	0	129	4	113	1	122	2	110	4	115	9
ABBM	105	0	129	4	105	0	129	4	114	0	123	1	112	2	113	11
GBBM	105	0	129	4	105	0	129	4	114	0	123	1	113	1	114	10

Table. 9. *A comparison of accuracy between our work and previous work*

Accuracy(%)							
Models	**Our Work**				**Previous work [1]**		
	Using ALL features (13)	**Using 7 features (PCA)**	**Using 11 features (LASSO)**	**Using 10 features (Relief)**	**Using ALL features (13)**	**Using 11 features (LASSO)**	**Using 10 features (Relief)**
RF	98.31	97.05	100	96.21	88.97	86.97	97.89
DT	98.31	97.89	99.57	94.53	86.97	88.6	89.12
KNN	96.21	96.63	96.63	90.33	83.61	93	94.11
SVC	98.31	98.31	98.73	95.79	-	-	-
LR	96.21	95.37	98.73	91.59	-	-	-
NB	98.31	98.31	98.73	91.59	-	-	-
GB	98.31	98.31	99.57	95.37	86.97	92.85	96.22
AB	98.31	98.31	99.57	94.53	89.07	90.75	92.85
DTBM	98.31	97.05	99.57	97.89	87.97	88.65	90.22
RFBM	98.31	98.31	99.15	96.21	92.65	97.65	99.05
KNNBM	96.63	97.05	96.63	91.17	89.63	96.6	98.05
SVCBM	98.31	98.31	98.73	94.53	-	-	-
ABBM	98.31	98.31	99.57	94.53	89.07	90.75	95.38
GBBM	98.31	98.31	99.57	95.37	90.97	97.85	98.32

As shown in fig. 5., 6, 7, and 8, the results of heart disease prediction accuracy are presented using traditional and hybrid machine learning classifiers as RF, NB, DT, KNN, SVC, LR, GB, AB, DTBM, RFBM, KNNBM, SVCBM, ABBM, and GBBM using all features and with three different feature selection techniques: Relief, LASSO, and PCA.

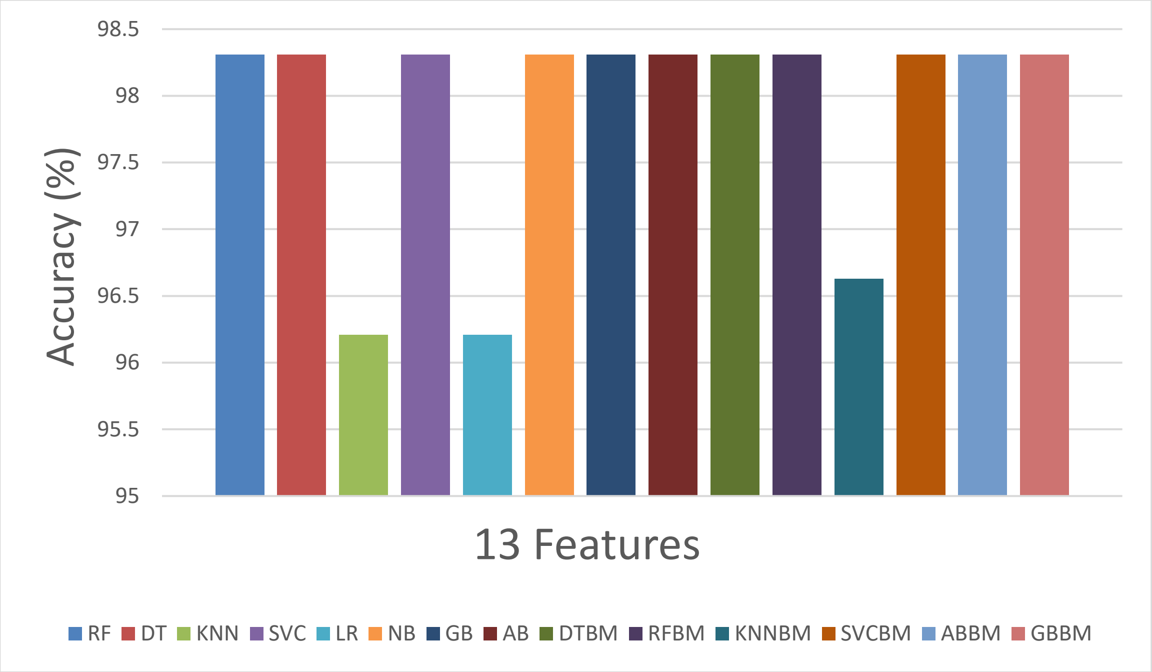

Fig.5. Classifiers accuracy using all features.

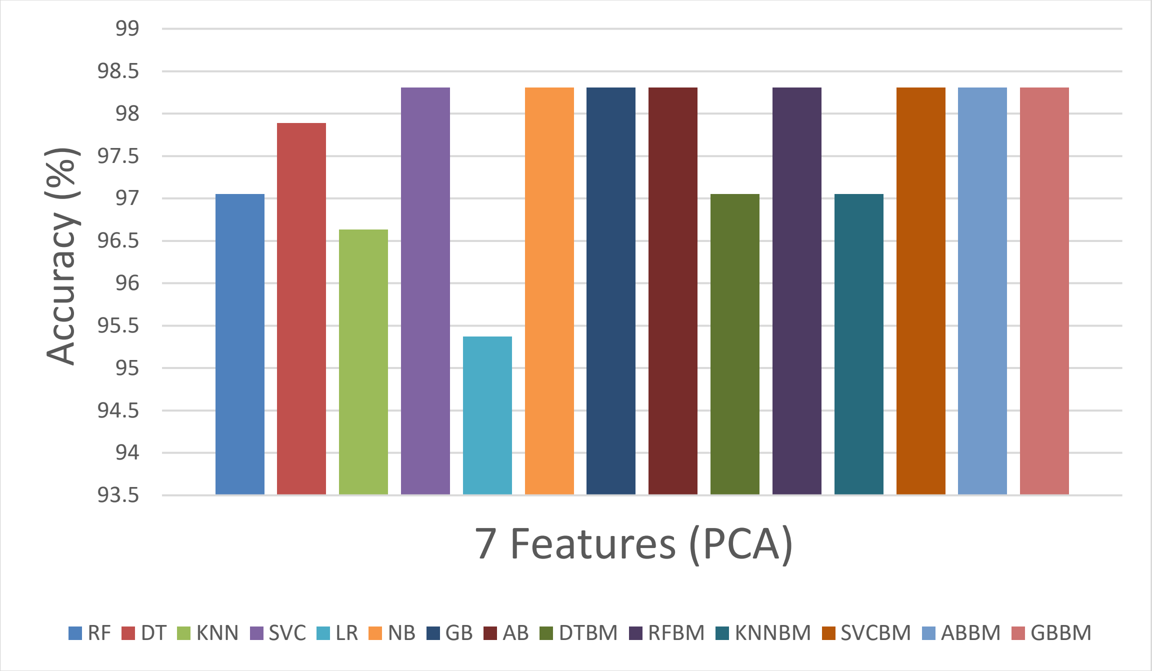

Fig.6. Classifiers accuracy using 7 features (PCA).

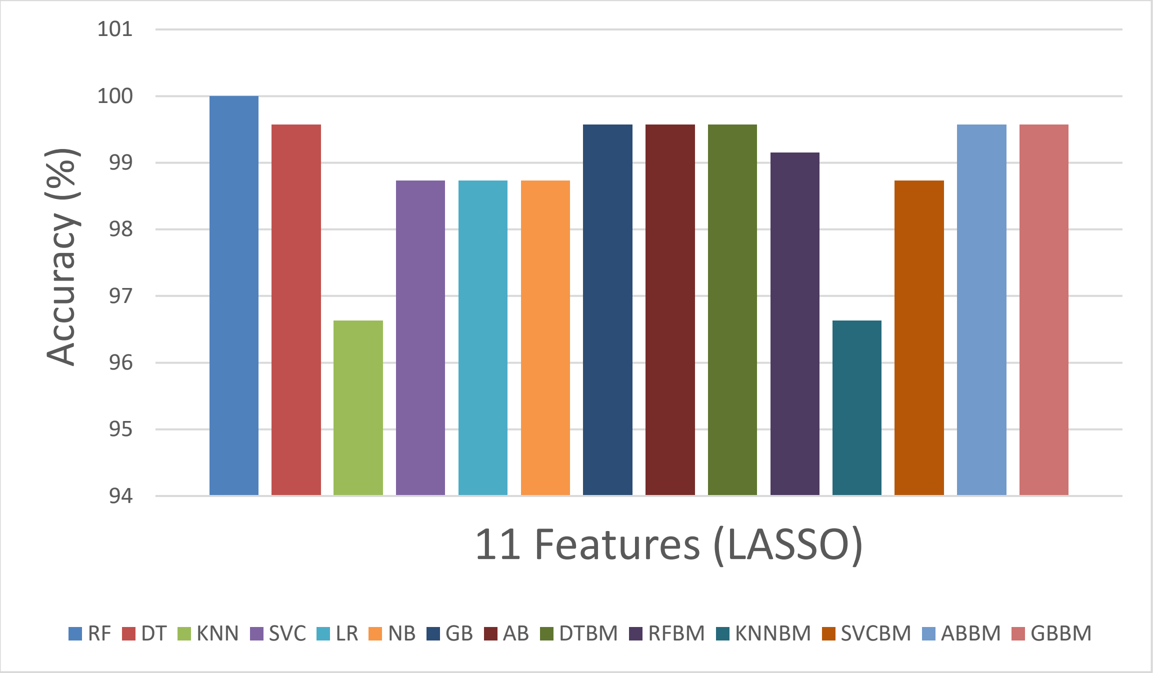

Fig.7. Classifiers accuracy using 11features (LASSO).

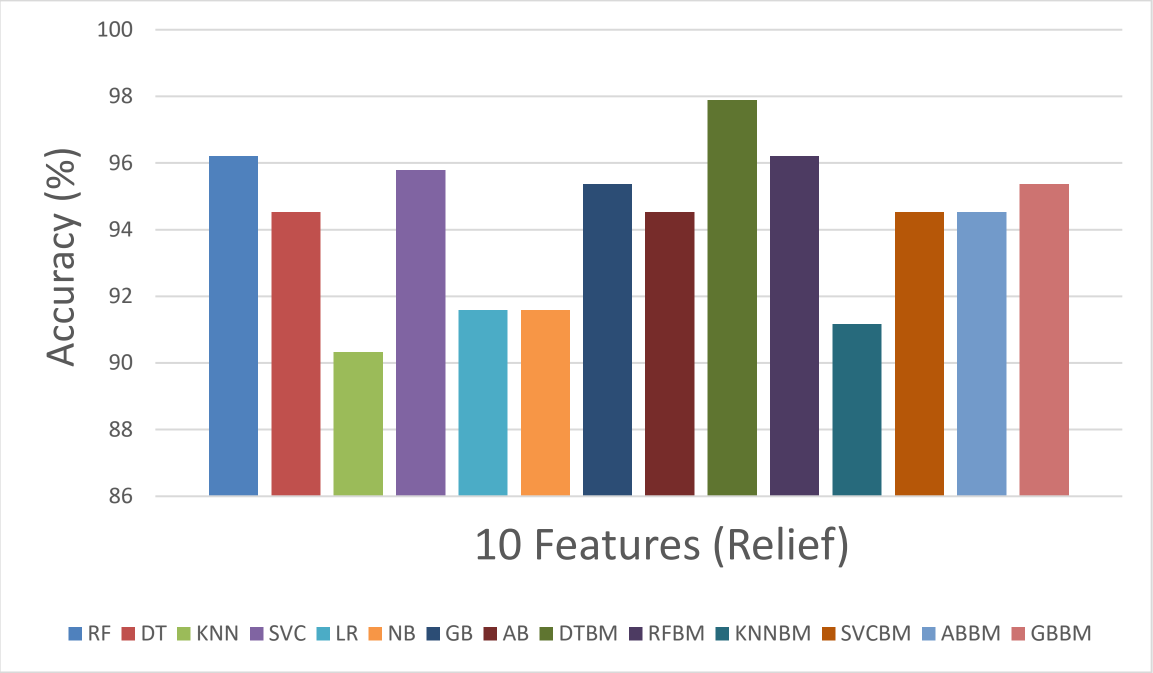

Fig.8. Classifiers accuracy using 10features (Relief).

GridSearchCV is a tuning procedure that determines the best parameters for a given model by allocating hyper parameters. GridSearchCV has been used in our proposed framework to get improved accuracy. The following parameters were used on the examined algorithms, which were determined using sklearn.model_selection.GridSearchCV (see Table VIII). For the ensemble technique, the default parameters were used with base classifiers.

Table. 10. *Parameters used*

Applied Algorithms	Parameters
RF	criterion='gini',max_depth=None,max_features='auto', max_leaf_nodes=None,min_samples_leaf=1, min_samples_split=2,n_estimators=10, random_state=None
DT	criterion='gini', max_depth=5, random_state=0
KNN	algorithm='auto', leaf_size=30, metric='minkowski', metric_params=None, n_jobs=1, n_neighbors=5, p=2,weights='uniform'
SVC	C=8.0, random_state=1, kernel='rbf' , degree=3
LR	C=0.1, class_weight=None, max_iter=100, multi_class='warn', n_jobs=None, random_state=None, tol=0.0001
NB	priors=None, var_smoothing=1e-09
GB	max_depth': 10, 'max_features': 'sqrt', 'min_samples_leaf': 6, 'n_estimators': 200
AB	algorithm='SAMME.R', base_estimator=None, learning_rate=1.0,n_estimators=100,random_state=0
DTBM	base classifier=’DT’, parameter= ‘default’
RFBM	base classifier=’RF’, parameter= ‘default’
KNNBM	base classifier=’KNN’, parameter= ‘default’
SVCBM	base classifier=’SVC’, parameter= ‘default’
ABBM	base classifier=’AB’, parameter= ‘default’
GBBM	base classifier=’GB’, parameter= ‘default’

5. Conclusion and Future Work

In this work, a framework for predicting cardiovascular disease was built using machine learning classifiers. The dataset used is collected from five data sets (Cleveland, Hungary, Switzerland, VA Long Beach, and Statlog), which are extracted from the UCI repository. Different feature selection methods like Relief, LASSO, and PCA are used. The dataset is split with a ratio of 80:20, where 80% is used to train the algorithms while 20% is used for testing. Then traditional and hybrid machine learning classifiers such as (RF), (NB), (DT), (KNN), (SVC), (LR), (GB), (AB), (BDTBM), (RFBM), (KNNBM), (SVCBM), (ABBM), and (GBBM) are applied to the dataset. Our improvements are made by (a) modifying the data preprocessing phase to clean it up by dealing with NaN values; applying standardization to the data; and (b) using different traditional and hybrid machine learning classifiers with different feature selection methods. This study shows that the LASSO feature selection technique may generate a tightly linked feature set that can be used with a variety of machine learning algorithms and achieve the best results in recall, accuracy, etc. The study also discovered that RF performs particularly well with high-impact features (as determined by the LASSO feature selection technique) and has significantly higher accuracy than other models. Finally, an accuracy of 100% was achieved when using a Random Forest classifier with the LASSO feature selection method, which is better than the existing approach's accuracy.

In the future, we will plan to implement our proposed approach with the Egyptian heart disease dataset.

References

[1] Ghosh, Pronab, et al. "Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms With Relief and LASSO Feature Selection Techniques." IEEE Access 9 (2021): 19304-19326.

[2] https://www.webmd.com/heart-disease/heart-disease-types-causes-symptoms “Last accessed” January 2022”

[3] Santulli, Gaetano. "Epidemiology of cardiovascular disease in the 21st century: Updated updated numbers and updated facts." Journal of Cardiovascular Disease Research 1.1 (2013).

[4] Marouli, Georgios. "Comparison between Maximum Entropy and Naïve Bayes classifiers: Case study; Appliance of Machine Learning Algorithms to an Odesk’s Corporation Dataset." (2014).

[5] Li, Shoushan, et al. "Employing personal/impersonal views in supervised and semi-supervised sentiment classification." Proceedings of the 48th annual meeting of the association for computational linguistics. Association for Computational Linguistics, 2010.

[6] Uddin, Shahadat, et al. "Comparing different supervised machine learning algorithms for disease prediction." BMC medical informatics and decision making 19.1 (2019): 1-16.

[7] Mohan, Senthilkumar, Chandrasegar Thirumalai, and Gautam Srivastava. "Effective heart disease prediction using hybrid machine learning techniques." IEEE access 7 (2019): 81542-81554

[8] Jackins, V., et al. "AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes." The Journal of Supercomputing 77.5 (2021): 5198-5219.

[9] Rahman, Nor Azziaty, Abdul, Kian Lam Tan, and Chen Kim Lim. "Supervised and unsupervised learning in data mining for employment prediction of fresh graduate students." Journal of Telecommunication, Electronic and Computer Engineering (JTEC) 9.2-12 :155-161 (2017).

[10] Westreich, Daniel, Justin Lessler, and Michele Jonsson Funk. "Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression." Journal of clinical epidemiology 63.8 :826-833 (2010)

[11] C. Raju, ‘‘Mining techniques,’’ in Proc. Conf. Emerg. Devices Smart Syst. (ICEDSS), Mar. 2016, pp. 253–255.

[12] B. Tarle and S. Jena, ‘‘An artificial neural network based pattern classification algorithm for diagnosis of heart disease,’’ in Proc. Int. Conf. Comput., Commun., Control Automat. (ICCUBEA), Aug. 2017, pp. 1–4.

[13] S. Hegelich, ‘‘Decision trees and random forests: Machine learning techniques to classify rare events,’’ Eur. Policy Anal., vol. 2, no. 1, pp. 98–120, 2016.

[14] Liu, Kailong, et al. "Feature analyses and modelling of lithium-ion batteries manufacturing based on random forest classification." IEEE/ASME Transactions on Mechatronics (2021).

[15] An Overview of Gradient Boosting Algorithm. Accessed: Jun. 31, 2020. [Online]. Available: https://machinelearningmastery.com/gentleintroduction-gradient-Boosting-algorithm-machine-learning/

[16] Suthaharan, Shan. "Support vector machine." Machine learning models and algorithms for big data classification. Springer, Boston, MA, 2016. 207-235.

[17] Shaimaa Mahmoud, Mahmoud Hussein, and Arabi Keshk. "Predicting Future Products Rate using Machine Learning Algorithms." International Journal of Intelligent Systems & Applications 12.5 (2020).

[18] Heart Disease Datasets From UCI Machine Learning Repository. Accessed: May 31, 2020. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Heart+Disease

[19] Khan, Mohammad Ayoub. "An IoT framework for heart disease prediction based on MDCNN classifier." IEEE Access 8 (2020): 34717-34727

[20] H. A. El Zouka and M. M. Hosni, ‘‘Secure IoT communications for smart healthcare monitoring system, Internet of Things,’’ Res. Paper, 2019, doi: 10.1016/j.iot.2019.01.003.

[21] Kaggle open dataset, Accessed: Jan. 15, 2020. [Online]. Available: https://www.kaggle.com/datasets

[22] Li, Jian Ping, et al. "Heart disease identification method using machine learning classification in e-healthcare." IEEE Access 8 (2020): 107562-107582.

[23] A. M. D. Silva, Feature Selection, vol. 13. Berlin, Germany: Springer, 2015, pp. 1–13

[24] R. Tibshirani, ‘‘Regression shrinkage and selection via the lasso,’’ J. Roy. Stat. Soc., B, Methodol., vol. 58, no. 1, pp. 267–288, Jan. 1996.

[25] E. O. Olaniyi, O. K. Oyedotun, and K. Adnan, ‘‘Heart diseases diagnosis using neural networks arbitration,’’ Int. J. Intell. Syst. Appl., vol. 7, no. 12, p. 72, 2015.

[26] Trevisan, G. Sergi, S. J. B. Maggi, and H. Dynamics, ‘‘Gender differences in brain-heart connection,’’ in Brain and Heart Dynamics. Cham, Switzerland: Springer, 2020, p. 937.

[27] C. Yadav and S. Pal, ‘‘Prediction of heart disease using feature selection and random forest ensemble method,’’ Int. J. Pharmaceutical Res., vol. 12, no. 4, 2020.

[28] Mienye, Ibomoiye Domor, Yanxia Sun, and Zenghui Wang. "An improved ensemble learning approach for the prediction of heart disease risk." Informatics in Medicine Unlocked 20 (2020): 100402.

[29] UCI machine learning repository: heart disease data set. http://archive.ics.uci.edu /ml/datasets/Heart+Disease. accessed Apr. 09, 2022.

[30] Framingham Heart study dataset. https://kaggle.com/amanajmera1/framingham heart-study-dataset. accessed Apr. 09, 2022.

[31] K. Uyar and A. Ilhan, ‘‘Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks,’’ Procedia Comput. Sci., vol. 120, pp. 588–593, Jan. 2017.

[32] A. U. Haq, J. P. Li, M. H. Memon, S. Nazir, and R. Sun, ‘‘A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms,’’ Mobile Inf. Syst., vol. 2018, pp. 1–21, Dec. 2018.

[33] R. Bhuvaneeswari, P. Sudhakar, and G. Prabakaran, ‘‘Heart disease prediction model based on gradient boosting tree (GBT) classification algorithm,’’ Int. J. Recent Technol. Eng., v

[34] R. Alizadehsani, J. Habibi, Z. A. Sani, H. Mashayekhi, R. Boghrati, A. Ghandeharioun, F. Khozeimeh, and F. Alizadeh-Sani, ‘‘Diagnosing coronary artery disease via data mining algorithms by considering laboratory and echocardiography features,’’ Res. Cardiovascular Med., vol. 2, no. 3, pp. 133–139, Aug. 2013.

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

[36] K. C. Tan, E. J. Teoh, Q. Yu, and K. C. Goh, ‘‘A hybrid evolutionary algorithm for attribute selection in data mining,’’ Expert Syst. Appl., vol. 36, no. 4, pp. 8616–8630, May 2009.

[37] A. A. Shetty and C. Naik, ‘‘Different data mining approaches for predicting heart disease,’’ Int. J. Innov. Sci. Eng. Technol., vol. 5, pp. 277–281, May 2016.

[38] C. A. Cheng and H. W. Chiu, ‘‘An artificial neural network model for the evaluation of carotid artery stenting prognosis using a national-wide database,’’ in Proc. 39th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC), Jul. 2017, pp. 2566–2569.

[39] Heart Disease Statlog Dataset of UCI Machine Learning Repository. Accessed: May 31, 2020. [Online]. Available: http://archive. ics.uci.edu/ml/datasets/statlog+(heart)

[40] Neloy, Md, et al. "A Weighted Average Ensemble Technique to Predict Heart Disease." Proceedings of the Third International Conference on Trends in Computational and Cognitive Engineering. Springer, Singapore, 2022.

[41] A. Acharya, ‘‘Comparative study of machine learning algorithms for heart disease prediction,’’ M.S. thesis, Helsinki Metropolia Univ. Appl. Sci., Helsinki, Finland, Apr. 2017. [Online]. Available: https:// www.theseus.fi/bitstream/handle/10024/124622/Final%20Thesis.pdf? sequence=1&isAllowed=y

[42] A. M. D. Silva, Feature Selection, vol. 13. Berlin, Germany: Springer, 2015, pp. 1–13.

[43] R. Tibshirani, ‘‘Regression shrinkage and selection via the lasso: A retro-spective,’’ J. Roy. Stat. Soc. B, Stat. Methodol., vol. 73, no. 3, pp. 273–282, Jun. 2011.

[44] Mahesh, Batta. "Machine Learning Algorithms-A Review." International Journal of Science and Research (IJSR).[Internet] 9 (2020): 381-386.

[45] Abdulkareem, Nasiba Mahdi, and Adnan Mohsin Abdulazeez. "Machine learning classification based on Radom Forest Algorithm: A review." International Journal of Science and Business 5.2 (2021): 128-142.

[46] Ghosh, Pronab, et al. "Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms With Relief and LASSO Feature Selection Techniques." IEEE Access 9 (2021): 19304-19326.

[47] An Overview of K_Nearest Neighbors Algorithm. Accessed: Jun. 31, 2020. [Online]. Available: https://www.javatpoint. com/k-nearest-neighbor algorithm- for-machine-learning

[48] D. Meyer, “Support Vector Machines – The Interface to libsvm in package e1071”, August 2015

[49] Li, Jian Ping, et al. "Heart disease identification method using machine learning classification in e-healthcare." IEEE Access 8 (2020): 107562-107582.

[50] Atiglo, D. Yaw, et al. "Sense of community and willingness to support malaria intervention programme in urban poor Accra, Ghana." Malaria journal 17.1 :289 (2018).

[51] M. Almasoud and T. E. Ward, ‘‘Detection of chronic kidney disease using machine learning algorithms with least number of predictors,’’ Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 8, pp. 89–96, 2019.

[52] Gradient Boosting Algorithm. Accessed: Jun. 31, 2020. [Online]. Available: https://data-flair.training/blogs/gradient-Boosting-algorithm/

[53] H. Ripon, ‘‘Rule induction and prediction of chronic kidney disease using boosting classifiers, Ant-Miner and J48 Decision Tree,’’ in Proc. Int. Conf. Elect., Comput. Commun. Eng. (ECCE), Cox’s Bazar, Bangladesh, 2019, pp. 1–6.

[54] A. Karim, S. Azam, B. Shanmugam, K. Kannoorpatti, and M. Alazab, ‘‘A comprehensive survey for intelligent spam email detection,’’ IEEE Access, vol. 7, pp. 168261–168295, 2019.

[55] Ensemble Techniques of Bagging. Accessed: Jun. 31, 2020. [Online]. Available: https://quantdare.com/what-is-the-difference-betweenBagging-and-Boosting

[56] An Explanation of Ensemble Bagging Techniques. Accessed: Jun. 31, 2020. [Online]. Available: https://towardsdatascience. com/ensemble-methods-Bagging-Boosting-and-stacking-c9214a10a205/

[57] G. Singh, ‘‘Breast cancer prediction using machine learning,’’ Int. J. Sci. Res. Comput. Sci., Eng. Inf. Technol., vol. 8, no. 4, pp. 278–284, Jul. 2020.