Perspectives in Communication, Embedded-Systems and Signal-Processing (PiCES) – An International Journal
ISSN: 2566-932X, Vol. 4, Issue 10, January 2021

A Comparative Study of Machine Learning Techniques in Heart Disease Detection

Praneeth Kumar T
Department of CSE, BNM Institute of Technology, Bangalore, India,
praneeth.kumt@gmail.com

Sriram Praveen V A
Department of CSE, RV College of Engineering, Bangalore, India,
srirampraveenva@gmail.com

Rohan Maheshwari
Department of CSE, RV College of Engineering, Bangalore, India,
therohanm@gmail.com

Sahana D Gowda
HOD, Department of CSE, BNM Institute of Technology, Bangalore, India,
sahanadgowda@bnmit.in

Abstract: Cardiovascular diseases (CVD) are the world's leading cause of mortality accounting for an estimated 31% of all deaths worldwide. Out of 17.9 million deaths per year due to CVDs, three-fourths of these deaths have occurred as there are no systems in place to predict the occurrence of a heart attack and warn the patient or doctor to take appropriate action. Data generated by clinical reports and examination reports by doctors are available for prediction through ERP models. Data science and reliable algorithms powered by AI can be used to develop medical devices that can predict such incidents of CVDs. In this paper, seven common classifiers are implemented that are computationally inexpensive and easily implementable and their performance metrics are compared. Two feature selection techniques are implemented and Grid Search is used for hyper-parameter tuning. Using k-fold cross-validation, classifiers are then evaluated, which generates classification metrics such as accuracy, f1-score, recall, and precision. It is evident from the study that the combination of Random Forest Classifier and SelectKBest feature selector has the highest accuracy of 89.706% and precision of 89.655%.

Keywords: CVD; Machine Learning; Feature Selection; Grid Search; K-Fold Validation; SVM; Decision Tree; Random Forest; KNN; Multi-Layer Perceptron; Gaussian Naive-Bayes; Bagging

I. INTRODUCTION

According to statistics [1], Cardiovascular diseases (CVDs) are the major cause of death globally, taking an estimation of 17.9 million lives every year. Four out of five CVD deaths are due to heart attacks and strokes, and one-third of these deaths occur prematurely in people under 70 years of age.[2]

Primary prevention of CVD is beneficial but most often, people do not get diagnosed before its occurrence [3]. As a precautionary in today's medical practices people with CVD are called for a periodic checkup and are monitored. This system is an offline process where the patient and the doctor are not alarmed by the symptoms before any mortality. Many a times, symptoms may not be precisely identified because the cause of the symptom could be for multiple reasons and goes unrecognized. This states that many of the systems are mutually exclusive for many diagnoses.

The major symptoms of CVDs include Chest discomfort, heartburn, pain that spreads to the arm, irregular heartbeat. In the present situation when the patient goes with such symptoms the first prerequisite is the ECG. Based on ECG and other data, decisions are taken.

Using historical data, Traditionally, doctors would inspect the medical images of ECG to find irregularities but are often prone to errors due to the microscopic patterns prevailing, which go unidentified in the scanned images. With the advent of data science and the power of AI, it is possible to find even the microscopic pattern irregularities and alarm the patients and doctors about the occurrences of CVD.

Classifiers with low computational complexity but high accuracy are chosen for comparative study and implementation in this paper. The seven classifiers considered are Support Vector Machines (SVM) - which uses the concept of hyperplane and classifies the data and creates classes. Decision Tree - compares with the logical expressions defined at the subnodes and defines the class, Random Forest - which generates subclass of Decision Trees, and uses a voting method to classify the unseen data. K-Nearest Neighbour - which finds the distance between the given samples and samples are classified, Multi-Layer Perceptron - which uses nonlinear function approximator for classification, Gaussian Naive Bayes - which classifies the data based on Bayes theorem, and Bootstrap Aggregating (Bagging) using KNN - which reduces variance and helps to avoid overfitting. K-Fold validation

has been employed to test the models, which divides the dataset into k groups where each group gets to become a test dataset, and performance metric is obtained as an average of these tests.

UCI repository [4] contains CVD datasets from Cleveland, Hungary, Switzerland, and the VA Long Beach. The UCI Cleveland dataset contains 76 features. The additional datasets that are relevant to this research are - Heart Disease Mortality Data Among US Adults (35+) by State/Territory and County is a dataset made available by the U.S. Department of Health & Human Services [5]. Heart Failure Prediction is a dataset created in 2015 [6]. Variations of CITED2 Are Associated with Congenital Heart Disease (CHD) in Chinese Population.[7]

In this paper, the UCI dataset is preprocessed using Standard Scaler and One Hot Encoder followed by feature selection. Each of these models is then trained on the selected features, in parallel with hyperparameter tuning. The performance metrics obtained are presented in the paper and validated using K-Fold cross-validation.

The paper is organized as follows - In section I, an introduction to various classifiers and datasets available on CVD are covered. In section II, a detailed literature survey of classifiers utilized in the research study on the health care dataset is presented. In section III, preprocessing, feature selection techniques, classifiers used for the comparative study, and the validation and evaluation metrics of various classifiers are explained. In section IV, results, discussion and conclusion are presented.

II. LITERATURE REVIEW

The advent of Data science revolutionized the healthcare sector. It reduced the risk of treatment failure with its accurate predictions and made it possible to observe even the microscopic patterns. It helped not only to predict diseases but also to improve the quality and integrity of medical records. Reduction in medical expenses is one of the accomplishments of data science. Few available techniques include using a mobile phone and a 2 electrode 1-lead ECG [8], Artificial Neural Network and Genetic algorithm, which uses clinical and ECG data [9], advanced ensemble machine learning technology, utilizing an adaptive Boosting algorithm [10], and various other Machine Learning Algorithms.

Several studies have attempted to apply machine learning methods to identify heart attacks. A popular dataset to apply them to has been the UCI Cleveland dataset [11].

Haq et al. [12] made a comparative study of various classifiers and found that the best accuracy was achieved by logistic regression using relief feature selection. The paper also concluded that SVM with mRMR showed a specificity of 100% and best accuracy of 89%.

Meshref, Hossam [13] found that Multilayer Perceptron (MLP) produced the best accuracy with 84.25%. The major contribution of this paper is a new metric called Feature Cost Index (FCI) which is a measure of how easily interpretable the model is. Random Forests (RF) were found to be 5 times more interpretable than MLP.

Jabbar [14], used Heart Statlog dataset [15]. The paper proposes to apply discretization and IQR filters followed by hidden Naive Bayes and achieves an accuracy of 100% which is likely because of overfitting.

Peter John [16] makes a comparative study on the Cleveland dataset. The paper concludes that Correlation-based feature selection and Bayes Theorem for feature selection is best for KNN. It achieved an accuracy of 85.55% with MLP and J48 coming in a close second with 85.18%.

Thirumalai et al. [17] focused on minimizing the number of features by using Pearson's r values, box plots, and linear regressions and formulated a table to depict the relations. Through that table this paper pointed out that BPS is usually twice as cholesterol, therefore either one of the two can be ignored.

L. Ali et al. [18] used SVM as both a feature selection and prediction model. The paper utilized L1 regularized linear SVM stacked with L2 regularized linear SVM and L1 regularized linear SVM stacked with L2 regularized SVM using RBF kernel. The first combination produced an accuracy of 91.11% using 11 and 12 features. The second combination produced an accuracy of 92.22%.

H. Yang et al. [19] provides an interesting alternative as instead of using classic datasets of cardiac data, the paper uses clinical notes and records and applies Natural Language Processing (NLP). It uses different approaches like rule-based approaches and dictionary-based keyword spotting. The system achieved an overall micro-average F-measure of 0.915.

Rajamhoana et al. [20] made a review of various methods and found that maximum accuracy is achieved by Principal Component Analysis (PCA) followed by classification by Feed-Forward Neural Network giving an accuracy of 95.2%.

Y. Li et al. [21], uses datasets of RR intervals from physionet. In this paper, DDM's (Distance Distribution Matrix) are generated using FuzzyGMEn and FuzzyLMEn. These are then fed into 3 CNN's (Convolutional Neural Network) namely Inception_V4, AlexNet, and DenseNet. The highest accuracy of 81.85% is achieved by FuzzyGMEn and Inception_V4.

Shivendra Kaura et al. [22], relied on ECG data mainly. ECG signals were through several filters to obtain signals without noise [R-R peak detection method]. A total of used 14 attributes were used, which included 6 categorical attributes. Stress was also considered as a factor in this paper. The neural network was trained over 1000 iterations

which is a disadvantage. An accuracy of 95% was obtained.

E. O. Olaniyi et al. [23], used UCI Cleveland Dataset with 13 attributes. The results obtained were 85% for feedforward multilayer perceptron and 87.5% for support vector machine. Despite the small difference, the paper concludes that SVM is the best algorithm for heart disease diagnosis and claims to reduce the chances of misdiagnosis on the part of medical practitioners.

Archana Singh et al [24], used a dataset from the UCI repository. As part of pre-processing all that was done was to convert categorized data by dummy value. 73% of the data was used for training and 27% for testing to obtain a maximum accuracy of 83% with SVM.

From the survey, it is evident that many works have been reported in the literature where ML and AI algorithms are applied to health care for identification, classification, and prediction.

Most papers as per the literature review utilized the subset of 14 features mentioned in the UCI repository. The surveyed papers have not made any attempt to improve classifiers performance by tuning the hyper-parameters. This paper further improved the classifiers performance by utilizing Grid-Search cross-validation for hyper-parameter tuning and also selected the features based on feature selection techniques.

III. MATERIALS AND METHODS

A. Methodology

The methods used in this research paper have been designed to make an accurate prediction of whether a person has suffered a heart attack or not. As mentioned, the Cleveland dataset is utilized to train the model. The dataset is pre-processed, and feature selection is done using 2 different techniques.

Grid search is used for hyperparameter tuning for the models on the features that have been selected. This is followed by training and testing. The results are obtained using 10-fold cross-validation, which produces various metrics.

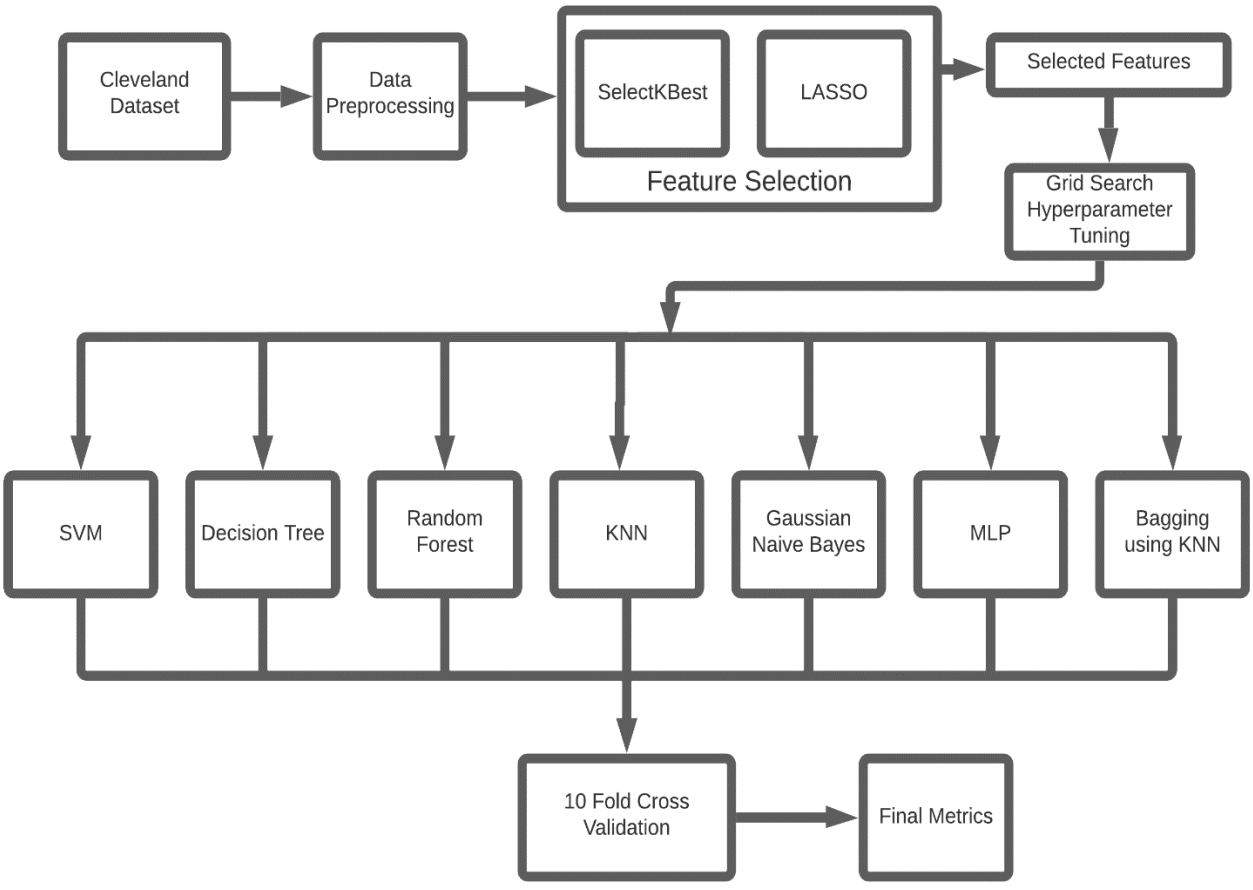

Fig 1. Flow of methodology

B. *Dataset Description*

The UCI Cleveland dataset [11] has been used to an exhaustive extent by different papers on this topic. It has 303 instances with 76 attributes each. Most papers use only the following 14 attribute age, sex, chest pain type, resting blood pressure, cholesterol, fasting blood sugar, resting ECG, maximum heart rate, exercise-induced angina, old peak, slope, number of vessels colored, and thalassemia. This paper attempts to make an analysis of the original 76 features and select the most appropriate ones using the aforementioned feature selection techniques. The output variable takes on values from 0-4, where 0 represents no heart attack and the values 1-4 represent the varying intensity of the heart attack.

Sl No.	Feature	Description	Domain
1.	Age	Age of patient in years	29-77
2.	Sex	Sex of patient (0: Female, 1: Male)	0: male 1: female
3.	cptype	Chest Pain	1: typical angina 2: atypical angina 3: non-anginal pain 4: asymptomatic
4.	restecg	Resting Electrographic Result	0: normal 1: having ST-T wave abnormality 2: hypertrophy
5.	thaldur	Duration of exercise test in minutes	1.8 - 15
6.	chol	Serum Cholesterol in mg/dl	126 - 564
7.	Mets	Mets achieved in the exercise	3 - 18
8.	thalach	Maximum heart rate achieved	71 - 202
9.	exang	Exercise-Induced Angina	0: No 1: Yes
10.	oldpeak	ST depression induced by exercise relative to rest	0 - 6.2
11.	slope	The slope of the peak exercise ST segment	1: Upsloping 2: Flat 3: Downsloping
12.	ca	number of major vessels colored by fluoroscopy	0 - 3
13.	thal	Thallium Scan	3: normal 6: fixed defect 7: reversible defect
14.	thalrest	Resting heart rate	40 - 119

Table 1. Description of features selected by feature selection techniques

Features Selected by SelectKBest	Age, Sex, cptype, restecg, thaldur, Mets, thalach, exang, oldpeak, slope, ca, thal
Features Selected by LASSO Selection	Sex, cptype, restecg, chol, thalach, exang, oldpeak, ca, thal, thalrest

Table 2. Features selected by feature selection techniques

C. *Preprocessing Data*

To get the best results from the classifiers it is important to preprocess the data. Features having more than 90% of its values as null are removed. Further, records having any null values are removed. The two techniques of data preprocessing used are One Hot Encoding and Standard Scaler. The standard scaler scales the features to a mean of 0 and variance of 1.

In One Hot Encoding is applied to variables with discrete values, it converts these features into integer values using a one of k scheme where for a given discrete value only one of the k values is 1.

D. *Feature Selection*

a) *Least Absolute Shrinkage and Selection Operator (LASSO)*

The 'lasso' minimizes the residual sum of squares subject to the sum of the absolute value of the coefficients being less than a constant. Some analysis of the constraint curve shows that LASSO shrinks coefficients to zero, thus making it perfect for feature selection. [25]

b) *SelectKBest*

SelectKBest is a method made available in the sci-kit-learn package in python. It takes a scoring function and ranks the features by these scores and retains the top 10 features. The scoring function used in this paper is ANOVA F-value. Analysis of Variance (ANOVA) first computes the ratio of the sum of squares between groups and its degrees of freedom (i.e., number of groups - 1) and then the ratio of the sum of squares within groups to its degrees of freedom (number of observations - number of groups) and then the F-value is computed by calculating the ratio of the first value to the second.

E. *Grid Search*

Grid Search is a hyperparameter tuning technique. Hyperparameters are those parameters that cannot be obtained from data for example the value of k in KNN or the number of neurons in each layer of MLP. This is a simplistic exhaustive search through the cartesian product of a given list of hyperparameters that are selected based on some cross-validation parameter on the training set [26]

F. *Classifiers Used*

a) *Support Vector Machine (SVM)*

Support vector machines (SVMs) are a set of supervised learning methods popularly used for classification. They use the concept of the hyperplane. It

classifies the data by creating a hyperplane between them. The points closest to this hyperplane are called the support vectors. In the case of nonlinear data, Kernel is used. It is a method of using linear classifiers to solve nonlinear problems. Several parameters can be tuned while using SVM, to get desired results, such as,

- C - parameter: It controls the tradeoff between smooth decision boundary and classifying training points correctly.
- Gamma parameter: It defines how far the influence of a single training example, with lower values implying far, reach and higher values implying close reach.

b) Decision *Tree (DT)*

Decision Tree algorithm can either be a Classification Tree or Regression Tree (referred to as CART). Classification Tree is used for classifying categorical data whereas Regression Tree is used to predict the outcome which can be a real number or so.

Classification Tree is used in this paper. This algorithm creates a tree-like structure by computing independent features and a target. For predicting a class label of data, the root node of a tree is accessed, and based on the comparison between the root attribute and data's attribute, the branch corresponding to that value is followed and jump to the next node takes place. This algorithm compares various features from various training examples and information gain to decide which variable to split on and how to make the splits.

Some terms associated with the working of this algorithm.

- Entropy: It measures the quality of a split.

$$E(S) = \sum_{i=1}^{c} -p_i log_2 p_i \qquad (1)$$

Where p_i = fraction of data points in class *i.*

An entropy of 0 would imply all data points belong to the same class and entropy of 1 implies data points are evenly split between classes.

- Information Gain: It is a property that measures how well a given attribute separates the training examples according to their target classification. This algorithm maximized the information gain and uses this to choose which feature to make a split on.

Information Gain =
*E(parent) - [weighted average] * E(children)* (2)

c) Random Forest (RF)

Random forests are a supervised learning algorithm that utilizes ensemble learning for problems like classification or regression. It works as an ensemble of multiple decision trees in which randomly selected data samples are used by each tree during training time. Predictions for unseen samples are made by the method of voting for the most popular class in the ensemble in the case of classification, or the average probability of prediction of each tree in the case of regression. [27]

There are high variance and low bias in a single decision tree model, which gives an inconsistent output and leads to overfitting. Every possible feature in a normal decision tree is considered when it is time to split a node and the one that produces the most separation between the observations in the left node versus the right node is chosen. Each tree in a random forest, by comparison, can only select from a random subset of features. This forces more variation among the trees and consequently results in low correlation across trees and therefore more diversification in the model. Usually, in each split, √p (rounded down) features are used for a classification problem with p features. Depending on the problem statement at hand, this parameter can be tuned.

d) K-Nearest Neighbour (KNN)

K-Nearest Neighbours is a non-parametric supervised learning algorithm commonly used for classification tasks. [28] This classifier finds the distance between the given points and the points in the dataset. This paper uses the Euclidean distance algorithm.

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \qquad (3)$$

A parameter called *k* is taken in and the classes of the k nearest points according to Euclidean distance are recorded. The point is classified as the class with the greatest number of nearest points.

e) Multi-Layer Perceptron (MLP)

Multi-layer Perceptron (MLP) is a supervised learning algorithm.[29] It consists of an input layer, one or multiple hidden layers, and an output layer to classify the given data. The training process broadly consists of three steps

1. Feedforward

Feedforward is the process that neural networks use to turn the given input into the output. In this step, the input of the model is multiplied with weights, and bias is added to it, to get a prediction. The output of the first layer now becomes the input of the 2nd layer (if it exists) and the same process of multiplying with weights and adding bias occurs. This step takes place at all the subsequent layers (if it exists).

2. Error calculation

The predicted output of the model is compared with the expected output and the loss is calculated. This loss can be determined by using various loss functions such as the cross-entropy loss function. This loss is then used to backpropagate, using the backpropagation algorithm.

3. Backpropagation

It runs the feedforward operation backward to spread the error to each of the weights to get a better predicting model. This step continues until we have a good model.

It is essential to find the optimal number of epochs, to know when to stop training our model, else the model will either be overfitted or under fitted.

f) Gaussian Naive Bayes (GNB)

Naive Bayes is a supervised learning classifier that uses Bayes theorem to classify any given data point [30]. It uses conditional probabilities to make the final classification. The prior probability for a given class [P(X) and P(Y)] is determined by taking the ratio of the number of data points in a class to the total number of data points. The probability of a given data point [P(X|Y)] given the hypothesis is determined by assuming a Gaussian distribution hence the name.

$$P(Y|X) = \frac{P(Y)\prod_{i=1}^{n} P(X_i|Y)}{P(X)} \quad (4)$$

Equation (4) is used to calculate the conditional probability of the given data point being in class Y.

$$y = argmax_y P(y)\prod_{i=1}^{n} P(x_i|y) \quad (5)$$

Equation (5) shows the final naive Bayes classifier which calculates the conditional probability for all classes and picks the highest probability as the predicted class.

g) Bootstrap Aggregating (Bagging) using KNN

In this method, the training dataset of size n is divided into m bags of size n' (In this paper 75% of n is used for SelectKBest and 25% for LASSO as determined by Grid Search). Then, n elements are chosen from the initial dataset with a replacement for each of the bags. In this manner, m classifiers are produced. The classifier of choice is KNN. Thus, an ensemble of learners is created where the output is the class with maximum predictions from the m classifiers.

G. K-fold Validation

This is a validation technique in which the dataset is divided into k groups and k-1 groups are used to train the classifier and the last group is used to test the dataset. In this paper, k is set to 10. The process of training is done k times such that each of the groups becomes a test set. The parameters are then computed by taking the average after each test. This popular technique has been used to get a better sense of the generalizability of the model.

H. Performance Metrics Used:

To evaluate our model 4 metrics have been used, the following terminology is used to explain each of them:

- True Positive (TP) - Number of patients who had a heart attack and the model correctly classified them as having one
- True Negative (TN) - Number of patients who didn't have a heart attack and the model correctly classified them as not having one.
- False Positive (FP) - Number of patients who didn't have a heart attack and the model incorrectly classified them as having one.
- False Negative (FN) - Number of patients who had a heart attack and the model incorrectly classified them as not having one.

a) Accuracy

Accuracy is the ratio of the number of TP to the total number of patients. Accuracy is a great measure but only when you have symmetric datasets where values of false positives and false negatives are almost the same, thus accuracy may not be the best metric for the dataset used.

$$Accuracy = \frac{TP}{TP + TN + FP + FN} \quad (6)$$

b) Precision

Precision is the ratio of the number of TP to the total actual positive observations.

$$Precision = \frac{TP}{TP + FP} \quad (7)$$

c) Recall

The recall is the ratio of the number of TP to the number of positive observations predicted by the model.

$$Recall = \frac{TP}{TP + FN} \quad (8)$$

d) F1 score

F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account.

$$F1\ score = \frac{2 * Precision * Recall}{Precision + Recall} \quad (9)$$

e) AUC-ROC Curve

Area Under Curve (AUC) Receiver Operating Characteristic (ROC) curve is constructed by plotting True Positive Rate (TPR) vs False Positive Rate (FPR).

$$FPR = \frac{FP}{FP + TN} \quad (10)$$

$$TPR = \frac{TP}{TP + FN} \quad (11)$$

The area under the curve is a measure of how easy it is for a classifier to differentiate between different classes.

IV. RESULTS AND DISCUSSION

A. *Results of selected features by SelectKBest*

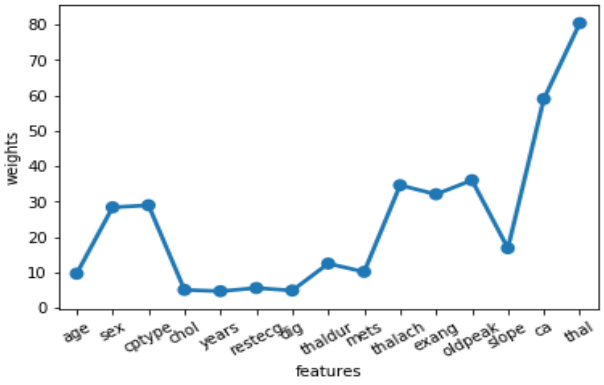

Fig 2. Weights of Selected Features by SelectKBest

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F-Score (%)	AUC-ROC Curve (%)
Support Vector Machine	80.882	77.419	80.000	78.689	91.404
Decision Tree	83.824	88.000	73.333	80.000	86.009
Random Forest Classifier	89.706	89.655	86.667	88.136	94.649
K-Nearest Neighbour	85.294	88.462	76.667	82.143	91.930
Multi-Layer Perceptron Classifier	82.353	82.143	76.667	79.310	91.667
Gaussian Naive Bayes	86.765	86.207	83.333	84.746	93.158
Bagging using KNN	85.294	83.333	83.333	83.333	92.895

Table 3. Performance Metrics table using SelectKBest

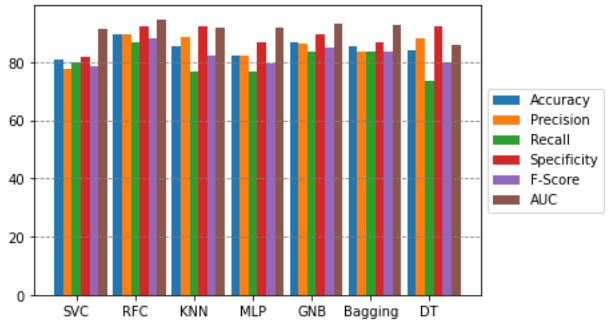

Fig 3. Performance metrics chart using SelectKBest

As seen from Fig 3., SelectKBest placed very high importance on the Thallium scan values and number of vessels from fluoroscopy which are also selected by LASSO.

From Table 3., it is evident that Random Forest outshines the other algorithms in every metric with an accuracy of 89.706%, precision of 89.655%, recall of 86.667%, an F-1 Score of 88.136%, and an AUC-ROC Curve value of 94.649%.

Another point of note is that recall and precision are key factors outside accuracy. This is because it is not as dangerous for a person who doesn't have heart disease to be diagnosed with one. But it is of extreme importance that some with CVD not be misdiagnosed.

B. *Results of selected features by LASSO Feature Selection Algorithm*

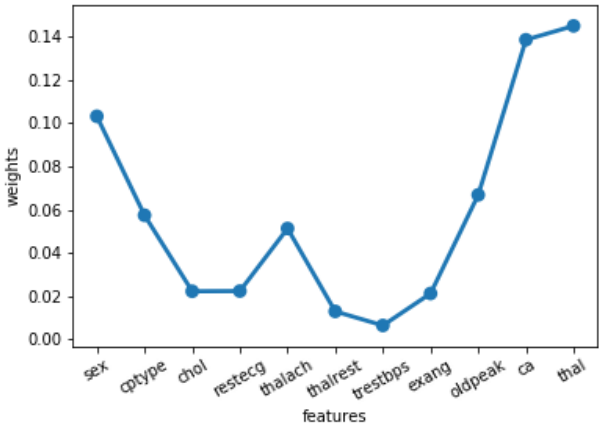

Fig 4. Weights of Selected Features by LASSO

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC Curve (%)
Support Vector Machine	85.294	88.462	76.667	82.143	92.281

Decision Tree	83.824	88.000	73.333	80.000	85.877
Random Forest Classifier	88.235	86.667	86.667	86.667	93.202
K-Nearest Neighbour	85.294	88.462	76.667	82.143	89.474
Multi-Layer Perceptron Classifier	82.353	84.615	73.333	78.571	88.333
Gaussian Naive Bayes	86.765	83.871	86.667	85.246	92.456
Bagging using KNN	85.294	85.714	80.000	82.759	89.035

Table 4. Performance Metrics table using LASSO

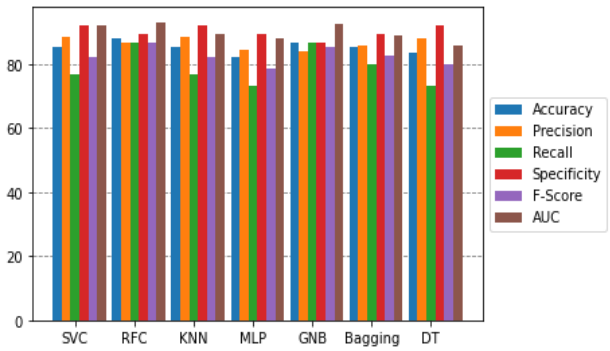

Fig 5. Performance metrics chart using LASSO

The second feature selection method used is LASSO. It is a commonly used shrinkage method. It is used to provide the weights in a linear function. These weights were then used as scoring and Thallium scan values and number of vessels from fluoroscopy were scored the highest. Thus, it these two features have little noise.

Random Forests had the highest accuracy of 88.235%. The precision is however lower. The highest precision is achieved by SVM and KNN with 88.462%. Random Forests also achieves the highest recall and F-1 Score of 86.667%. The Random Forest also achieves the highest AUC-ROC Curve value of 93.202%.

Thus, Random Forests provided better results for heart attack detection.

V. CONCLUSION

In this paper Support Vector Machine, Random Forest Classifier, K Nearest Neighbour, Multi-Layer Perceptron Classifier, Gaussian Naive Bayes, Decision Tree, and Bagging using KNN have been applied on Cleveland Heart Disease Dataset from UCI data repository. With as many as 75 attributes available in the dataset, it is essential to pick only those attributes which are of importance for predicting the results. It is also required to bring down the number of features being used. Lasso feature selection algorithm and select K Best has been employed to perform this task. Using Lasso feature selection, 10 features have been picked and the SelectKBest algorithm has picked 12 features. The combination of Random Forest Classifier and SelectKBest feature selector has the highest accuracy of 89.706% and precision of 89.655%.

Detection of CVDs at early stages is of vital importance and requires picking the right parameters for the models to work efficiently. To achieve this, Grid Search has been employed for tuning the right parameters for each classifier. For validation, 10-fold Cross Validation has been applied to the dataset for partitioning the data into training and test sets. Among all the classifiers used, Random Forest outshined all of them.

REFERENCES

[1] World Health Organization. 2021. Cardiovascular diseases. [Online] Available: https://www.who.int/health-topics/cardiovascular-diseases

[2] Centers for Disease Control and Prevention. Underlying Cause of Death, 1999–2019. CDC WONDER Online Database. Atlanta, GA: Centers for Disease Control and Prevention [Online] Available: https://wonder.cdc.gov/ucd-icd10.html

[3] Finegold JA, Shun-Shin MJ, Cole GD, et al. Distribution of lifespan gain from primary prevention intervention, Open Heart 2016; vol. 3, no. 1, e000343.

[4] Dua, Dheeru and Graff, Casey, 2019, UCI Machine Learning Repository, University of California, Irvine, School of Information and Computer Sciences [Online] Available: http://archive.ics.uci.edu/ml

[5] Centers for Disease Control and Prevention, [Online] Available: https://chronicdata.cdc.gov/Heart-Disease-Stroke-Prevention/Heart-Disease-Mortality-Data-Among-US-Adults-35-by/i2vk-mgdh

[6] Davide Chicco, and Giuseppe Jurman. "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone." BMC Medical Informatics and Decision Making vol. 20, no. 1, pp. 16, Feb, 2020

[7] Liu, Yan AND Wang et al. “Variations of CITED2 Are Associated with Congenital Heart Disease (CHD) in Chinese Population”, PLOS ONE, Vol. 9, May, 2014

[8] P. Leijdekkers and V. Gay, "A Self-Test to Detect a Heart Attack Using a Mobile Phone and Wearable Sensors," 2008 21st IEEE International Symposium on Computer-Based Medical Systems, Jyvaskyla, 2008, pp. 93-98

[9] Ravish, D. et al. “Heart function monitoring, prediction and prevention of Heart Attacks: Using Artificial Neural Networks.” 2014 International Conference on Contemporary Computing and Informatics (IC3I) (2014): 1-6.

[10] Kathleen H. Miao, Julia H. Miao and George J. Miao, “Diagnosing Coronary Heart Disease using Ensemble Machine Learning” International Journal of Advanced Computer Science and Applications (IJACSA), 7(10), 2016.

[11] Robert Detrano, “Cleveland Heart Disease Dataset”,Cleveland Clinic Foundation,UCI Machine Learning Repository [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Heart+Disease

[12] 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 Information Systems, Vol. 8, pg 1-21, June, 2015.

[13] Hossam Meshref, "Cardiovascular Disease Diagnosis: A Machine Learning Interpretation Approach" International Journal of Advanced Computer Science and Applications (IJACSA), 10(12), 2019.

[14] M. A. Jabbar and S. Samreen, "Heart disease prediction system based on hidden naïve bayes classifier," 2016 International Conference on Circuits, Controls, Communications and Computing (I4C), Bangalore, 2016, pp. 1-5

[15] Statlog (Heart) Data Set, UCI repository, [Online] Available: http://archive.ics.uci.edu/ml/datasets/statlog+(heart)

[16] Peter John," Study and development of FSS for disease prediction", IJSRP, Vol 2, Issue 10, (2012)

[17] C. Thirumalai, A. Duba and R. Reddy, "Decision making system using machine learning and Pearson for heart attack," 2017 International conference of Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, 2017, pp. 206-210

[18] L. Ali et al., "An Optimized Stacked Support Vector Machines Based Expert System for the Effective Prediction of Heart Failure," in IEEE Access, vol. 7, pp. 54007-54014, 2019

[19] Yang, Hui & Garibaldi, Jonathan. (2015). A Hybrid Model for Automatic Identification of Risk Factors for Heart Disease. Journal of biomedical informatics, pp. 171-182, September, 2015

[20] S. Rajamhoana, C. A. Devi, K. Umamaheswari, R. Kiruba, K. Karunya, and R. Deepika, "Analysis of Neural Networks Based Heart Disease Prediction System," 2018 11th International Conference on Human System Interaction (HSI), Vol. 9 Gdansk, pp. 233-239, March, 2020

[21] Y. Li et al., "Combining Convolutional Neural Network and Distance Distribution Matrix for Identification of Congestive Heart Failure," in IEEE Access, vol. 6, pp. 39734-39744, 2018

[22] S. Kaura, A. Chandel and N. K. Pal, "Heart Disease-Sinus arrhythmia prediction system by neural network using ECG analysis," 2019 2nd International Conference on Power Energy, Environment and Intelligent Control (PEEIC), Greater Noida, India, 2019, pp. 466-471,

[23] E. O. Olaniyi, O. K. Oyedotun, and K. Adnan, "Heart diseases diagnosis using neural networks arbitration," International Journal of Intelligent Systems and Applications, vol. 7, no. 12, p. 72, November, 2015.

[24] A. Singh and R. Kumar, "Heart Disease Prediction Using Machine Learning Algorithms," 2020 International Conference on Electrical and Electronics Engineering (ICE3), Gorakhpur, India, 2020, pp. 452-457

[25] Tibshirani, Robert. "Regression Shrinkage and Selection via the Lasso." Journal of the Royal Statistical Society. Series B (Methodological), vol. 58, no. 1, pp. 267–288. 1996

[26] Chin-Wei Hsu, Chih-Chung Chang and Chih-Jen Lin ," A practical guide to support vector classification", Technical Report, National Taiwan University. May, 2016

[27] Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324

[28] X. Wu and V. Kumar, Top 10 Algorithms in Data Mining, Springer, Berlin, Germany, vol. 1, pp 151-159 2007.

[29] Scikit-learn: Multi-Layer Perceptron [Online] Available: https://scikit-learn.org/stable/modules/neural_networks_supervised.html

[30] Friedman, N., Geiger, D. & Goldszmidt, M. Bayesian Network Classifiers. Machine Learning 29, 131–163 (1997).