Prediction of Heart Disease using Ensemble Learning and Particle Swarm Optimization and Genetic Algorithm

Patmanjari Nautiyal(00401042019)
Nupur Goel(0401042019)
Manya(04801042019)
Chhavi Bhadana(05001042019)
Prerna Khera(06301042019)
Priya Chaturvedi(06801042019)

April 2021

Student Name	Enrollment no.	Contributions
Patmanjari Nautiyal	00401042019	Wrote introduction and Performed Ensemble Methods: Random Forest and AdaBoost and applied feature selection on Random Forest,Decision Tree and AdaBoost.
Nupur Goel	04001042019	Read the abstract, introduction and data section of papers [6] ,[5],[2],[12] and wrote their summaries. Performed SVM. Applied Genetic Algorithm on SVM, KNN and LR.
Manya	04801042019	Read the abstract, introduction and data section of papers. [8], [7] , [9], [3] and wrote their summaries. Performed ANN and applied feature selection on ANN and KNN.
Chhavi Bhadana	05001042019	Wrote the abstract and has given a brief description of base paper in related work section. Performed Decision Tree, Naive Bayes and Logistic Regression.
Prerna Khera	06301042019	Read the abstract, introduction and data section of papers [1], [4], [10], [11] and wrote their summaries. Performed KNN. Wrote about Genetic Algorithm and Applied Genetic Algorithm on Random Forest, AdaBoost and Decision Tree. Represented Tables 12 and 13 in graphical form and wrote the conclusion.
Priya Chaturvedi	06801042019	Visualize the data by analyzing the dataset attributes , label encoding and plotting graphs or charts. Applied PSO on various Machine Learning Algorithms like Bagged Tree, AdaBoost, KNN and Decision Tree.

Abstract

With the world becoming a smaller and paced space. The pace has led us to a number of health problems. With the advancement of technology the lazy man's lifestyle has become a host to heart problems. Every year the number of people affected by heart diseases increases by a concerning number. Early and accurate prediction has become a requirement. Heart diseases do not have clear early signs and health professionals being humans can do as much as their human potential dictates. In addition taking tests for heart diseases is not a common and economical process. Hence prediction using Artificial Intelligence and machine learning can be a serviceable tool. Models with high accuracy have been built but heart diseases being a crucial health problem need a 100 percent accuracy. AI and ML technologies have great abilities and that can help health professionals as well as patients to identify numerous diseases in early stages. Assisting reduction of mortality rate, number of surgeries performed and other complications. To find a tool with the highest possible accuracy in predicting cardiovascular disease at early stages, we used multiple feature selection methods like PSO and Genetic Algorithm and found that

PSO (Particle swarm Optimization) paired with Bagged Tree ensemble returned 90% accuracy and PSO with LR gave the highest accuracy of 100%.

1 Introduction

Cardiovascular disease or commonly known as heart disease is a class of diseases that includes conditions that affect the heart. Types of heart diseases are coronary artery disease, cardiomyopathy, congenital heart disease, pulmonary embolism, stroke, and peripheral artery disease. The most common symptoms of heart disease may include chest pain, chest discomfort, and shortness of breath. Other symptoms may include pain in the neck, jaw, upper abdomen, or back, fatigue, nausea, weakness in the legs and hand (narrowed blood vessels). Sometimes heart disease can go undiagnosed until a person experiences a heart attack, arrhythmia, or heart failure. High blood pressure, high cholesterol, and diabetes are leading medical conditions that increase the risk of heart diseases. Lifestyle choices like obesity, smoking, and excessive use of alcohol can also affect the heart. Family history of heart problems and age can also lead to serious heart problems. In 2019 nearly 18 million people died due to heart disease which contributes to 32 percent of all global deaths. China has the highest number of deaths due to heart conditions followed by India, Russia, and the United States. Death rates due to heart disease were lowest in Japan, France, and Peru. More than 50 million cases of heart diseases are registered every year. According to surveys conducted, the annual death rate due to heart disease has increased by more than 50 percent from 1990 to 2017. Studies have shown that people aged between 45-60 years are highly prone to a heart attacks. Also, deaths caused by cardiovascular diseases are higher in men than in the female. We can protect ourselves from heart problems by improving our lifestyle and being active. It is necessary to detect heart disease at an early stage to get proper medical guidance and treatment. Therefore various methods are used by the healthcare system to prevent such medical conditions. These methods make data easily accessible and save both time and effort. One such is the data mining concept which plays a vital role in the health care industry. Data mining helps in the proper usage of data and analytics by the health system. It also identifies patterns and predicts various diseases such as liver disease, heart disease, etc. Due to the early detection of high-risk diseases by using data mining, patients receive better and more affordable health care treatment that eventually increases their life span. It also manages customer relationships and detects waste, insurance fraud, and abuse. Data mining has introduced various methods and techniques to predict heart diseases in the medical industry. One such system is a radial base function with a support vector machine. But this system could not provide satisfactory classification results. There was another system that used a genetic support vector machine to predict heart valve conditions. Another system uses multiple data mining technologies which predict heart disease. All these methods are highly complex and difficult to understand. Also, different

factors can lead to heart problems in both males and females. In data mining, the classification is trained and tested to predict a patient's medical condition. It also helps in predicting if a person is suffering from heart disease. Due to a significant increase in cases of heart disease, there is an urgent requirement for a higher level of classification accuracy that a traditional classifier cannot provide. In such a case ensemble classifier is used. Ensemble classifier combines the prediction of different components and then gives the output. An ensemble classifier gives maximum accuracy and is better than an individual classifier. Feature subset selection removes irrelevant data, which helps in improving the accuracy. Particle swarm optimization and Genetic Algorithm are feature selection techniques. They remove the extra features from an enormous amount of data, which helps to increase the accuracy of the classifier. In some cases, PSO gives better accuracy as compared to GA while in some other Machine Learning models GA performs better. The rest of the paper is organised as follow: Section 2 discusses the prior work as well as our work in the field. Section 3 deals with data set description and visualise the data. The main contributions of this paper are:
1. Using different ensemble classifier methods to get the accuracy that is useful for the prediction of heart disease.
2. Removing irrelevant data using the feature selection method to increase the accuracy
3. Applying different types of feature selection methods like Particle Swarm Optimization, and Genetic Algorithm on the dataset to extract relevant features that affect heart disease.
4. Selecting the best performance-based algorithm which gives the highest accuracy for the prediction of heart disease.

2 Related Work

Base paper(s): We selected the work of Indu Yekkala, Sunanda Dixit and M.A.jabbar [**?**] as our **base paper** and worked on **dataset** used by them. They used ensemble learning and particle swarm optimization(PSO) to predict Coronary heart disease. Their paper analysed various ensemble methods accompanied by PSO as a subset feature selection. The results displayed an accuracy of 100 percent in one of the methods.

Important paper(s): M.A.Jabbar, B.L.Deekshatulu, and Priti Chandra [9] proposed a classification model which uses the random forest as their classifier with chi square and genetic algorithm as their feature selection measures.Chi square and genetic algorithm are used to select attributes that will contribute more to diagnosis. The model showed an accuracy of 83.70 Per for the heart stat-log dataset...

Jabbar [6] proposed a decision support system in which KNN is used as classifier and PSO as feature subset selection measure to predict heart disease.

Author's main aim is to increase the performance of KNN classifier. Their method reached to an accuracy of 100 Per with hybrid feature selection. Sensitivity and TP rate of this model was 100 Per.

Emakhu et al. [5] implemented ensemble classifiers for the prediction of heart disease. Data mining classification techniques include AdaBoost, Random Forest, Bagging and Voting Ensemble that is applied to ensemble classifier. Random Forest Ensemble classifier is used to enhance accuracy of the model. This method achieved an accuracy of 87.04 Per.

H. Benjamin [4] introduced various ensemble learning methods and illustrated that higher performance and better accuracy in the prediction of heart disease could be achieved using these methods. He used Support Vector Machine, Naive Bayes and K-Nearest Neighbours classifiers for constructing the ensemble. He performed Bagging, Stacking and AdaBoost algorithms and evaluated their performances across various folds of cross-validation. He also indicated that the folds of improved the the values of precision, recall and f-measure. It became evident from his results that the AdaBoost algorithm outperformed the others.

Other papers: Rajwant Kaur [8] used a dataset the contains 50 people data collected by American Heart Associations.She tested data using the Support Vector Machine (SVM) classifier and Genetic Algorithm for predicting heart diseases based on the risk factor. As of result SVM and Genetic algorithm gives higher accuracy as compared to Neural network and Genetic algorithm.Accuracy of system for SVM+Genetic Algorithm = 0.95152 whereas for Neural Network = 0.904444.

Kathleen H. Miao .[7] used methods like Ensemble learning classification and prediction models have been used to detect coronary heart disease. The developed Ensemble classification and prediction models are based on an adaptive Boosting algorithm. For the result, an average sensitivity of 86.61 % (the presence of disease) and an average specificity of 83.76 % (the absence of disease) were obtained. The accuracy of testing result for different datasets were 80.14 % for CCF, 89.12 % for HIC, 77.78 % for LBMC, and 96.72 % for SUH.

Jabbar [3] used alternating decision tree method with Principal component analysis (PCA) for diagnosing heart disease. The data was collected from the corporate hospital in Hyderabad for about 96 patients with 10 features. The accuracy for the approached model was increased by 3.13 % for the full Hyd training set. Also, the accuracy for the model was increased by 2.11% as compared to J48 for the heart disease full Hyd training set.

Jabbar et al. [2] used this research paper in Andhra Pradesh to predict risk score of heart disease. Author investigated to predict heart disease by using feature subset selection and classification. Associative classification is used to

amend the classification accuracy.

Latha.B et al. [12] proposed ensemble classifier for classification and also used particle swarm optimization as feature subset selection to predict heart disease. An ensemble classifier like AdaBoost and PSO is used to improve learning accuracy of this model. Accuracy of this model is 84.88 % by using AdaBoost and PSO and the error rate is 4%.

Madhumita Pal et al. [10] proposed a heart disease prediction system using Random Forest algorithm. The authors used a dataset consisting of 303 instances and 14 attributes. They expressed the outcomes of the dataset with a classification accuracy of 86.9 %, sensitivity value of 90.6%, and specificity value of 82.7%. They expressed results in a confusion matrix and plotted an ROC curve using Random Forest Algorithm and predicted a diagnosis rate of 93.3 Per.

Mohammed et al. [1] Proposed that various heart diseases can be predicted with the help of data mining and machine learning algorithms. For this, they used a classification algorithm under the supervision of Random Forest and Naive Bayes for classification and prediction purposes. They used 384 instances with 14 attributes of Cleveland, Hungarian and Switzerland datasets. By performing filter method and feature selection wrap method, they obtained that Naive Bayes has more accuracy over Random Forest.

Dhyan Chandra Yadav et al. [11] Proposed a system of prediction of heart disease by using feature selection methods and applying them on tree-based classification algorithms. By using Pearson Correlation, Recursive Features Elimination and Lasso Regularization and applying them on M5P, Random Tree and Reduced Error Pruning with Random Forest Ensemble method they calculated the values of classification accuracy, precision, sensitivity and ROC. They concluded that Pearson Correlation and Lasso Regularization along with Random Forest gave 99 Per accuracy.

Table 1: Comparative analysis between prior research works who have used **same dataset** as the basepaper.

Author(s)	**Brief Description**	**Results**
Jabbar [6] [a]	Author proposed a decision support system in which KNN is used as classifier and PSO as feature subset selection measure to predict heart disease. Author's main aim is to increase the performance of KNN classifier.	Accuracy with hybrid feature selection was 100 Per. Sensitivity and TP rate of this model was 100 %.
Emakhu et al. [5] [b]	Author implemented ensemble classifiers for the prediction of heart disease. Data mining classification techniques include AdaBoost, Random Forest, Bagging and Voting Ensemble that is applied to ensemble classifier. Random Forest Ensemble classifier is used to enhance accuracy of the model.	Accuracy of this model for feature selection using PSO + Bagged Tree gave 100 %value, PSO + Random Forest gave 90.37%, PS O + AdaBoost gave 88.89 %.
M.A.Jabbar et al.[9] [c]	Proposed a classification model which uses the random forest as their classifier with chi square and genetic algorithm as their feature selection measures.Chi square and genetic algorithm are used to select attributes that will contribute more to diagnosis.	The model showed an accuracy of 83.70 % for the heart stat-log dataset
H. Benjamin [4]	Author introduced various ensemble learning methods and illustrated that higher performance and better accuracy in the prediction of heart disease could be achieved using these methods. He used Support Vector Machine, Naive Bayes and K-Nearest Neighbours classifiers for constructing the ensemble. He performed Bagging, Stacking and AdaBoost algorithms and evaluated their performances across various folds of cross-validation.	The model indicated that the folds of cross-validation improved the values of precision, recall and f-measure. It was evident from his results that the AdaBoost algorithm outperformed the others.

[a] https://www.researchgate.net/publication/317357702
[b] https://www.researchgate.net/publication/350620141
[c] https://ieeexplore.ieee.org/abstract/document/7057816

Table 2: Prior research works related to our problem but using different datasets.

Author(s)	DataSet	Brief Description
Jabbar et al.[2]	14 datasets from UCI data were taken and experiments were conducted on them by using 10 cross fold validation	Author used this research paper in Andhra Pradesh to predict risk score of heart disease. Author investigated to predict heart disease by using feature subset selection and classification. Associative classification is used to amend the classification accuracy.
Latha.B et al.[12]	Dataset was taken from kaggle.	Author proposed ensemble classifier for classification and also used particle swarm optimization as feature subset selection to predict heart disease. An ensemble classifier like AdaBoost and PSO is used to improve learning accuracy of this model. Accuracy of this model is 84.88 % by using AdaBoost and PSO.
Rajwant Kaur et al.[8]	The author used a dataset the contains 50 people data collected by American Heart Associations.	She tested data using the Support Vector Machine (SVM) classifier and Genetic Algorithm for predicting heart diseases based on the risk factor. As of result SVM and Genetic algorithm gives higher accuracy as compared to Neural network and Genetic algorithm. Sensitivity of system for SVM+Genetic algorithm = 0.957895 whereas for Neural Network = 0.954545. Accuracy of system for SVM+Genetic Algorithm = 0.95152 whereas for Neural Network = 0.904444.
Kathleen H. Miao al.[7]	Heart Disease Data Set link[a] of this dataset	Author used methods like Ensemble learning classification and prediction models have been used to detect coronary heart disease. The developed Ensemble classification and prediction models are based on an adaptive Boosting algorithm. For the result, an average sensitivity of 86.61 perr(the presence of disease) and an average specificity of 83.76per(the absence of disease) were obtained. The accuracy of testing result for different datasets were 80.14 per for CCF, 89.12per for HIC, 77.78 per for LBMC, and 96.72 per for SUH.
Jabbar et al.[3]	The data was collected from the corporate hospital in Hyderabad(Hyd) for about 96 patients with 10 features.	Author used alternating decision tree method with Principal component analysis (PCA) for diagnosing heart disease. The accuracy for the approached model was increased by 3.13% for the full Hyd training set. Also, the accuracy for the model was increased by 2.11% as compared to J48 for the heart disease full Hyd training set.
Madhumita Pal et al.[10]	The datasets [b] are collected and gathered from the Machine Learning Repository (UCI). It contains 303 instances with 14 attributes.	Author proposed a heart disease prediction system using Random Forest algorithm. She used a dataset consisting of 303 instances and 14 attributes. She expressed the outcomes of the dataset with a classification accuracy of 86.9 %, sensitivity value of 90.6%, and specificity value of 82.7 %. She expressed results in a confusion matrix and plotted an ROC curve using Random Forest Algorithm and predicted a diagnosis rate of 93.3 %.

[a] https://archive.ics.uci.edu/ml/datasets/Heart+Disease
[b] http://archive.ics.uci.edu/ml/datasets/Heart+Disease

Table 3: In continuation with Table 2

Mohammed et Al. [1]	The datasets [1] are collected and gathered from the Machine Learning Repository (UCI). It contains 394 datasets copies with 14 attributes.	Mohammed et Al. Proposed that various heart diseases can be predicted with the help of data mining and machine learning algorithms. For this, they used a classification algorithm under the supervision of Random Forest and Naive Bayes for classification and prediction purposes. They used 384 instances with 14 attributes of Cleveland, Hungarian and Switzerland datasets. By performing filter method and feature selection wrap method, they obtained that Naive Bayes has more accuracy over Random Forest.
Dhyan Chandra Yadav et al. [11]	The dataset [2] was taken from the Machine Learning Repository (UCI). It contains 1025 instances with 14 attributes.	Proposed a system of prediction of heart disease by using feature selection methods and applying them on tree-based classification algorithms. By using Pearson Correlation, Recursive Features Elimination and Lasso Regularization and applying them on M5P, Random Tree and Reduced Error Pruning with Random Forest Ensemble method they calculated the values of classification accuracy, precision, sensitivity and ROC. They concluded that Pearson Correlation and Lasso Regularization along with Random Forest gave 99% accuracy.

3 Dataset Description

Dataset has been obtained from UCI Machine Learning which is a treasure of dataset where a person can search or download the data set according to size/dimensions or Machine learning function .This data set is known as Statlog(heart) Dataset which is used to predict absence or presence of heart disease.

Table 4 is describing the dataset information through counts of some key entities involved in the dataset.

Table 4: Dataset description

Details	count/type
Number of instances/rows	270
Number of attributes/columns	13
Data set feature	multivariate
Attribute feature	categorical,real
Missing values	no

Table 5 describes the attributes of the dataset. Also in text here, explain those attributes briefly. As evident, there are so and so many attributes.

Table 5: Attribute Description

Attribute Name	**Attribute brief description**	**Attribute Data type**	**Domain values**
Age	Age of patient	real	29-77
Sex	Gender	binary	0 - female 1 - male
Chest	Type of chest pain	nominal	1- typical angina 2- atypical angina 3- nonanginal pain 4 - asymtomatic
Resting blood pressure(mmHg)	Normal blood pressure	real	94 - 200
Serum cholestoral(mg/dl)	It is the sum of high density and low density (HDL+LDL) cholestoral in blood.	real	126-564
Fasting blood sugar<100mg/l	A test used to retrieve the patient's blood sugar after a overnight fasting.	binary	1 - yes 0 - no
Resting electrocardiographic results(ECG)	A process used to emulate the root of different heart conditions.	nominal	0 - normal 1- having St-t unusual 2 - left ventricular hypertropy
Maximum Heart Rate achieved	Measurement of a person's heart rate during a physical activity.	real	<141 - low 111-194 - moderate 152>- high
Exercise induced angina	It triggers when you exerted yourself with physical activity like exercise.	binary	1- yes 0 - no
Old peak	ST depression causes by exercise	real	<2 - low 1.5-4.2 - risk 2.5>- high
Slope	Displacement of ST segment with respect to induced exercise in heart rate	nominal	1 - upward 2- flat 3- downward
Number of major vessels	Major vessels colored by flouroscopy(blood flow by coronary artery)	nominal	0-3
Thal	Anemia can be induced by Thalassemia	nominal	3 - normal 6 - fixed defect 7 - reversible defect
Class	Heart Disease	binary	0 - absent 1 - present

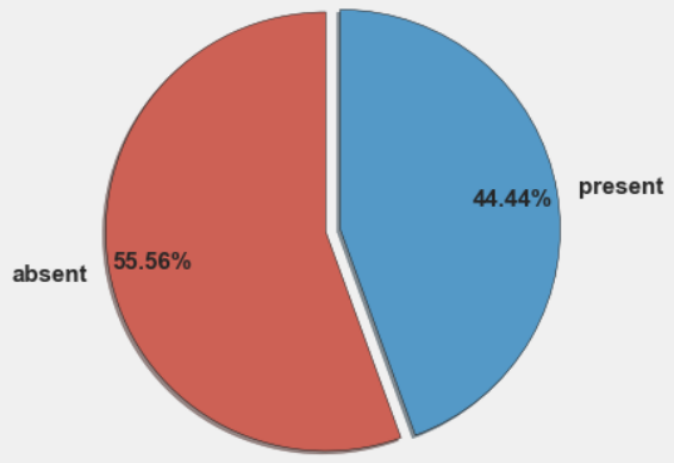

Figure 1: This is a pie chart depicting percentage contribution of present(1) and absent(0) data member of class attribute where 44.44% of present and 55.56% of absent data member.

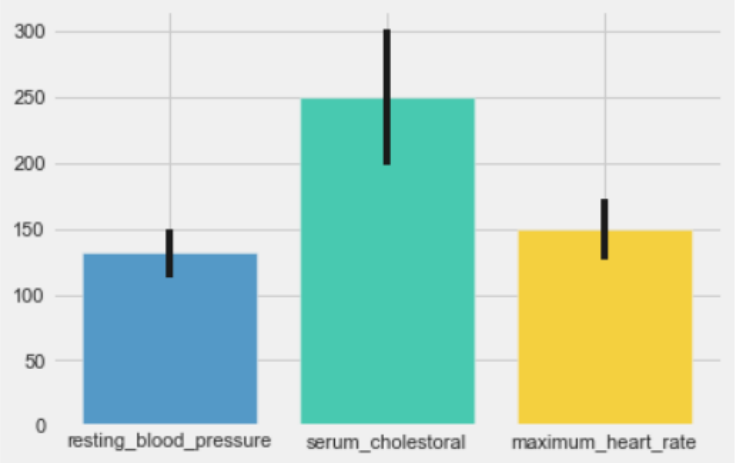

Figure 2: Bar charts helps us to visualize the magnitudes of a quantitative variable so , here we have taken 3 attributes and the black vertical line shows the variability in the data set larger the black line more will be the variability i.e values are varying too much from the mean like here serum cholestoral have the highest variability.

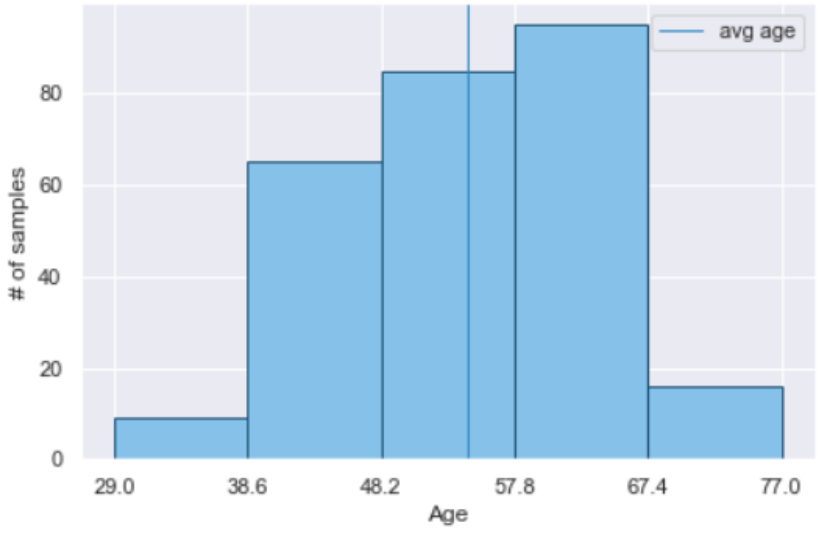

Figure 3: Histograms are univariate i.e one requires only one continuous variable to plot a histogram here we have used age attribute from the dataset and the long vertical blue line shows the average age .

4 Proposed Methodology

4.1 Data Pre-processing

The main focus of this research paper is to intensify the ensemble classifier for classification with Particle swarm optimization as a feature subset selection for detection of heart disease.At first Load the data set and then data-cleaning and pre processing which includes renaming features to appropriate name , converting features to categorical variables, feature encoding ,filling missing values has to be done .After this step we will also detect and remove the outlier using z score.

$\mathbf{z = (x\text{-}\mu)/\sigma}$ where, z=standard score x=observed score μ=mean σ=standard deviation

There are 16 data points which are outliers and all will be removed. Now before splitting data set into train and test we first encode categorical features as dummy features and segregate feature and class variable.Then we will normalize all the numeric feature in the range of 0 to 1 and cross validation will also be done and final step will be model building or evaluation.

4.2 Proposed Workflow/Architecture

Different models will be evaluated on most important evaluation metrics are sensitivity, specificity, Precision, F1-measure,PPV,NPV. The confusion matrix

is used to measure the accuracy, sensitivity, and specificity.

Table 6: Confusion Matrix

	TRUE(1)	FALSE(0)
TRUE(1)	TP	FN%
FALSE(0)	FP	TN%

Where
TP – Positive tuples that are correctly labelled by the classifier.
TN – Negative tuples that are correctly labelled by the classifier.
FN – Positive tuples that are incorrectly labelled by the classifier.
FP –Negative tuples that are incorrectly labelled by the classifier
To evaluate the performance of ensemble methods we have used following classification measures
• Sensitivity = TP/(TP+FN)
• Specificity = TN/(FP+TN)
• Positive predictive value (PPV) = TP/(TP+FP)
• Nega+tive predictive value (NPV) = TN/(FN+TN)
• Accuracy = (TP+TN)/(TP+TN+FP+FN)

We apply feature selection methods like PSO and Genetic Algorithm on data set and Remove least ranked attributes and keep predominant features.

Apply ensemble classifiers on primary features (Bagging, Random Forest, and AdaBoost) At last measure the performance of proposed method.

5 Experiment Design

Random forest: it is a type of ensemble classifier which is used for predicting better accuracy. It is used to help in medical application. Random Forest algorithm constructs N number of decision trees by using random attributes as their input. In Random Forest, the bias remains the same but the number of tree can vary. With increase in number of trees the variance of model decreases. The prediction can be made by average voting or majority voting.

Adaptive Boosting: It is another type of ensemble classifier method. It converts weak learner into strong learner. It works by choosing a base classifier such as decision tree. We can improve it by taking misclassified attributes from the dataset in an iterative method. Equal weights are assigned to the attributes and a base classifier is choosen.

ANN: It is one of the models used for experimenting in this project. The dataset on which we trained our model was Stat-log heart diseases dataset. We trained the dataset using multilayer perceptron classifier (MLP). MLP consist of an input, hidden and an output layer. For this dataset we had only one hidden layer and max iteration = 300. After training the dataset we got 83.33% as the mean accuracy of the given test data and labels.

KNN: K-Nearest Neighbors belongs to the supervised learning domain and finds intense application in pattern recognition, data mining and intrusion detection. KNN works on a principle assuming every data point falling in near to each other is falling in the same class. KNN algorithms decide a number k which is the nearest Neighbor to that data point that is to be classified. The optimal value of K using cross-validation. It gives high accuracy.

SVM: Support Vector Machines are popular Supervised Learning Algorithm which is used for both classification and regression problems. But primarily, it is used in classification problems. This is popular because it can handle multiple continuous and categorical variables. It's main goal is to create best line or decision boundary which separate n-dimensional space into classes and this decision boundary is known as hyperplane. It selects extreme points that hepls in creating the hyperplane.

LR: Logistic Regression is a popular algorithm based on the sigmoid function. It is linear algorithm with a non-linear transform on output. It does assume a linear relationship between the input variables with the output. NB: Gaussian Naive Bayes is a variant of Naive Bayes that follows Gaussian normal distribution and supports continuous data. DT: It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches

Table 7: Ensemble methods Approaches

	Ensemble method	Accuracy	Precision	Sensitivity	Specificity	F1 Score	Positive Predictive Value (PPV)	Negative Predictive Value (NPV)
1.	Bagged Tree	82%	91%	91%	93%	94%	92%	92%
2.	Random Forest	77.78%	73.07%	79.17%	76.67%	76%	87%	86%
3.	AdaBoost	74%	70.83%	70.83%	76.67%	70.83%	75%	88%

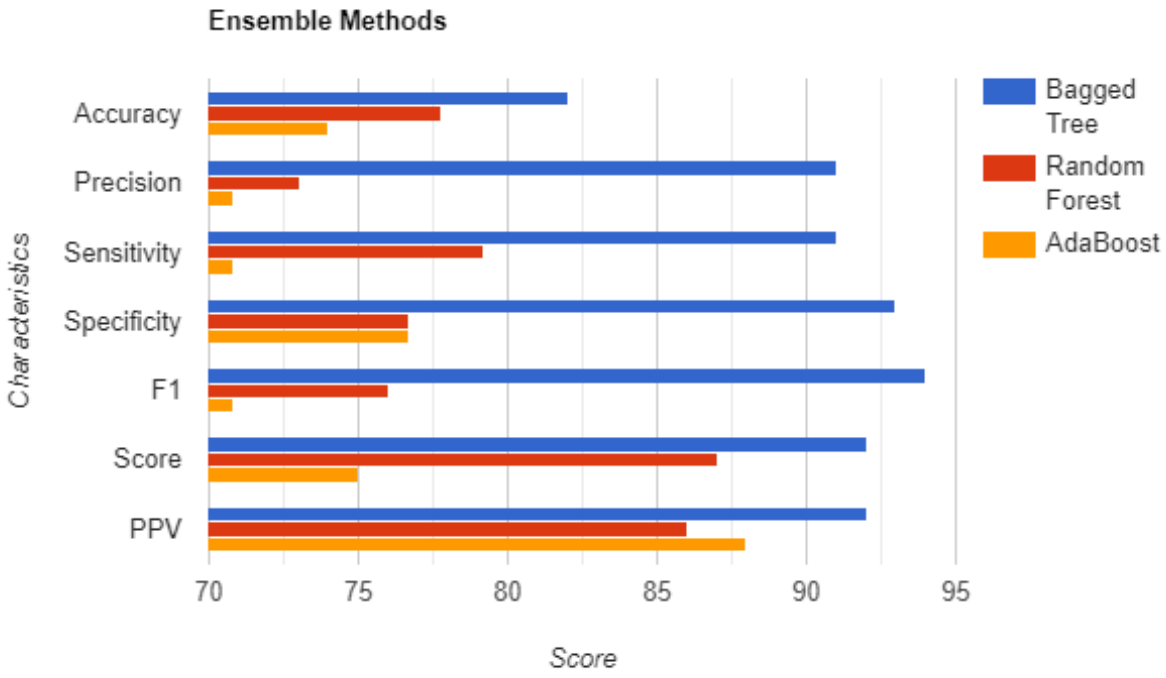

Figure 4: Comparison for different measures

Table 8: Comparison with various models

S.no	Techniques	Accuracy
1.	Decision Tree	75.30%
2.	Naïve-Bayes	77.78%
3.	KNN	83.33%
4.	SVM	85.19%
5.	ANN	77.7%
6.	LR	80.25%

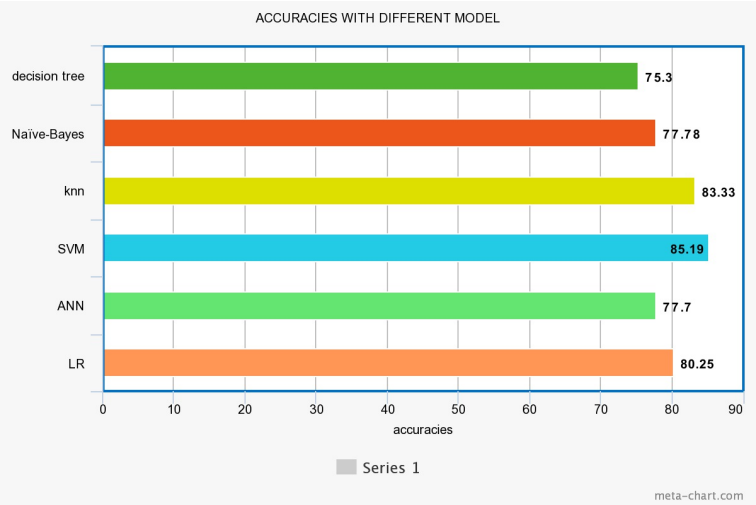

Figure 5: Accuracies with different Classifiers

Table 9: Feature selection with Different models

S. No	Ensemble Method	Accuracy	Precision	Recall	F1 Score	Sensitivity	Specificity
1.	Random Forest	75%	79%	75%	77%	76%	75%
2.	AdaBoost	79%	75%	77%	83%	75%	77%
3.	ANN	88.88%	89.28%	88.87%	89.28%	89.28%	88.46%
4.	KNN	72.22%	84%	73.71%	73.68%	65.62%	81.81%
5.	Decision Tree	79%	78%	78%	82%	86%	70%

6 Result/Observations

6.1 Feature engineering

In order to enhance the performance of these models various approaches are used which also require re-engineering or re-processing the dataset. Data gathered from the real world can be tremendously disorganized. Therefore, feature engineering is needed to modify the data into something that is more significant. And new features can be produced from raw data. Features can be weakly significant or strongly significant. Weakly significant features are those which are carrying the information of other features might fully or partially might carry. Strong features are those which carry information which other features do not whereas weak features can carry either partially or fully features of others. It's considered a good practice to have extra features than to miss out on a few important ones. The process of feature selection includes not only attaching

features but also deleting few not so known or unnecessary features.Therefore, feature selection techniques can be used in order to reduce or avoid overfitting.

6.2 Feature Selection

Feature Selection is the process of decreasing the number of disordered or disarranged features under inspection generating only a smaller set of predominant features. Higher the number of features, the more tough to visualize and train the machine learning models with the dataset. Many of the features might be tally and hence, their impact would be needless. This is where the feature selection algorithms come into the picture. Therefore, Particle Swarm Optimization (PSO) and Genetic Algorithm (GA) were the techniques used for feature selection .

6.3 Particle Swarm Optimization (PSO)

This method was originally proposed by Kennedy and Eberhart it is a algorithm based on the concept of swarm intelligence. The particles that constitute the PSO system fly around in a multidimensional search space; each particle regulate its position as stated by its own experience during its flight, and the experience of the neighboring particles, utilizing the best position face by itself and its neighbors.

Heart stalog data set contains 14 instances.PSO removed 6 least ranked features and selected 7 as predominant features which will enhance the accuracy of the ensemble classifier.

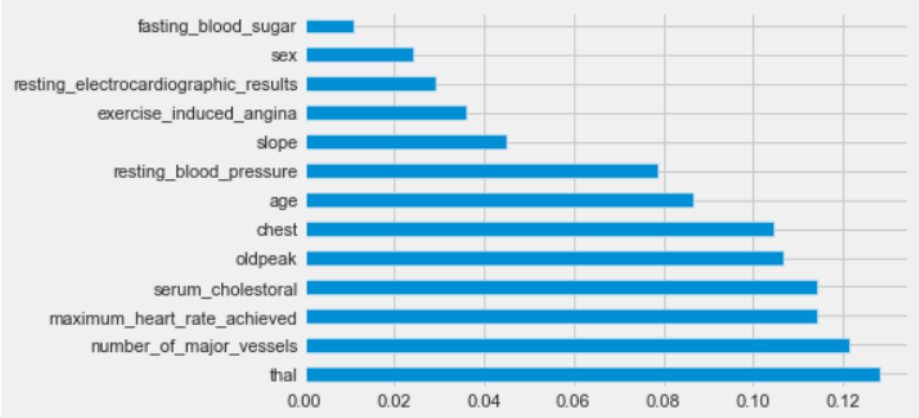

Figure 6: Comparison for different measures

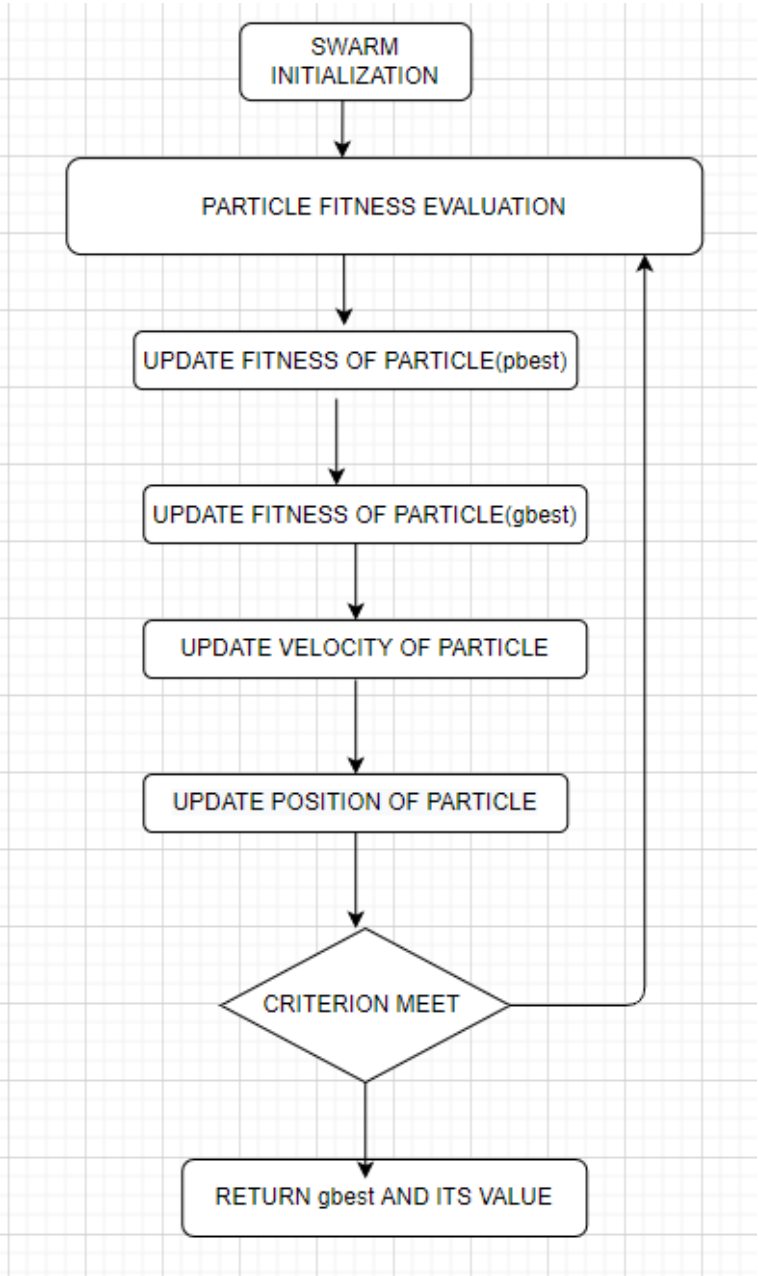

Figure 7: Flow chart of pso for feature selection

Table 10: Features selected by PSO

S.no	Selected Features
1.	Thal
2.	number_of_major_vessels
3.	maximum_heart_rate_achieved
4.	serum_cholestoral
5.	old peak
6.	chest
7.	age

Table 11: Accuracy of our proposed method for feature selection using PSO

S.no	Proposed Methodolgy	Accuracy
1.	PSO+ BAGGED TREE	90.6%
2.	PSO+RANDOM FOREST	83.3%
3.	PSO+ADABOOST	74.05%
4.	PSO+KNN	55.5%
5.	PSO + LR	100%
6.	PSO + DT	91.24%
7.	PSO + NB	90.06%

Table 11 records the accuracy of the proposed method using feature subselection method PSO. By using PSO feature selection measure on Bagged Tree, accuracy obtained is 90.6%. With feature selection method PSO on Random Forrest algorithm, accuracy is 83.3%. With feature selection method PSO on AdaBoost accuracy is 74%.

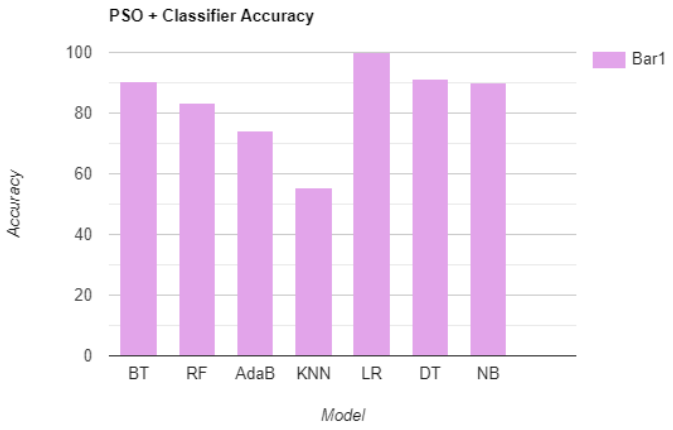

Figure 8: Comparison of Accuracy with Particle Swarm Optimization

6.4 Genetic Algorithm

The Genetic Algorithm is a model of biological evolution based on Charles Darwin's theory of natural selection. It was introduced by John Holland in the year 1975. It is a randomized algorithm that is developed to mimic the mechanics of natural selection and natural genetics. Holland used crossover and recombination, mutation, and selection to enhance the algorithm's problem-solving strategy. It supports the "survival of the fittest" concept for solving a problem and generates high-quality solutions for optimization problems.

The genetic algorithm works by initializing a population of chromosomes and determining its fitness. Then we perform the selection operation through which the parents of the next generation are selected. Crossover between the selected parents creates a new population. The new population is mutated to maintain its diversity. Ultimately, the fitness for this new generation is estimated.

Table 12: Comparison Of Ensemble Methods

S. No	Ensemble Method	Accuracy	Precision	Recall	F1 Score	Sensitivity	Specificity
1.	Random Forest	77.78%	73.07%	79.17%	76%	79.17%	76.67%
2.	AdaBoost	74.07%	70.83%	70.83%	70.83%	70.83%	76.67%
3.	Genetic Algorithm + Random Forest	79.63%	78%	75%	77%	75%	83.33%
4.	Genetic Algorithm + AdaBoost	79.63%	78%	75%	77%	75%	83.33%

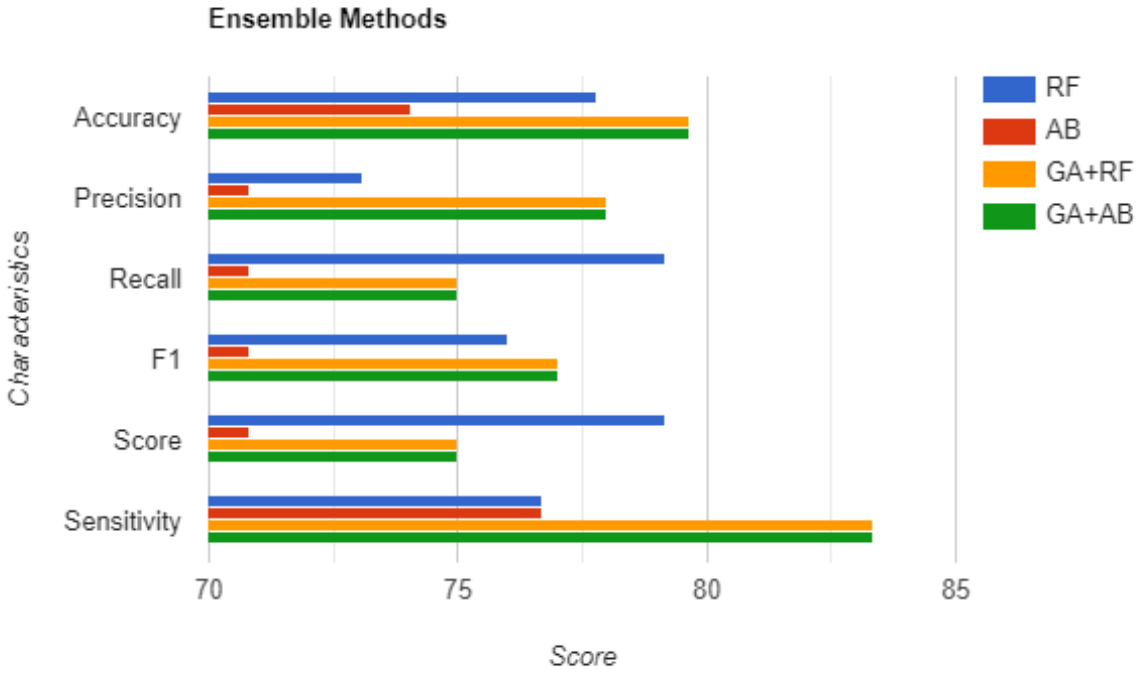

Figure 9: Comparison of Ensemble Methods before and after GA

Table 13: Comparison Of Various Machine Learning Models

S.no	ML Techniques	Accuracy	Precision	Recall	F1 Score	Sensitivity	Specificity
1.	DT	64.81%	60%	62.5%	61.22%	62.55%	66.67%
2.	SVM	85.19%	80.77%	87.5%	84%	87.5%	83.33%
3.	KNN	83.33%	80%	83.33%	81.63%	83.33%	83.33%
4.	LR	85.19%	80.77%	87.5%	84%	87.5%	83.33%
5.	GA + DT	75.93%	87%	54%	67%	54%	80%
6.	GA + SVM	88.89%	87.5%	87.5%	87.5%	87.5%	90%
7.	GA + KNN	87.04%	84%	87.5%	85.71%	87.5%	86.67%
8.	GA + LR	92.59%	91.67%	91.67%	91.67%	91.67%	93.33%

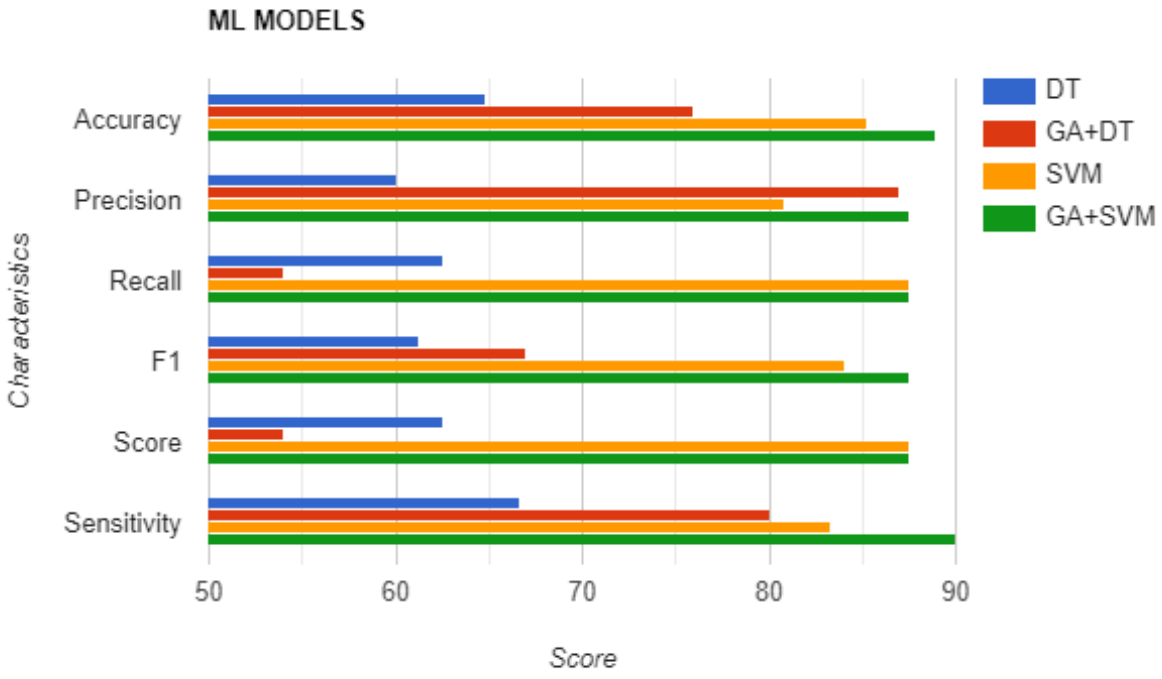

Figure 10: Comparison of ML Models before and after GA

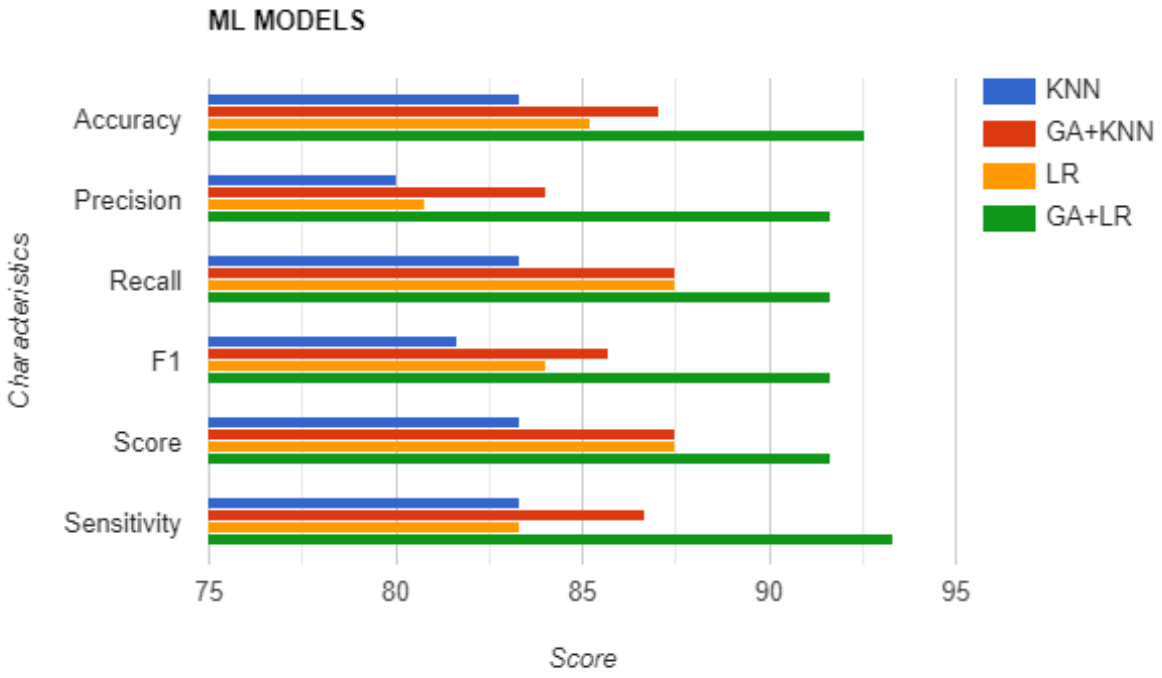

Figure 11: Comparison of ML Models before and after GA

6.5 CONCLUSION

In this paper, we have applied various machine learning algorithms on heart stat-log data set from UCI Repository and we have used those results for the prediction of heart diseases. The proposed approach uses PSO and Genetic Algorithm (GA) as a feature selection method to reduce the less important features. Then applied ensemble methods as a classifier to decrease misclassified rate and to enhance the classification production. The experimental results by applying 2 different feature selection algorithms, i.e. PSO and GA, on the same data set we got different results. Some machine learning models like DT, KNN, LR and AdaBoost classifier performed better with GA while some other models like RF gave better results with PSO. It has been shown that PSO + BT gives a high accuracy of 90.6% while PSO + LR gives the highest accuracy of 100%. Therefore, these models will help medical specialists to accurately predict and early diagnosis of heart diseases using a subset of features. In future we would want to develop an intelligent system that can determine the probability of a person developing a heart disease. And once such a system is developed focus can be shifted towards developing models that specify the line of treatment for

a diagnosed patient based on past records.

References

[1] Mohammed Alrifaie, Zakir Ahmed, Asaad Hameed, Modhi Mutar, Hang Tuah, and Jaya Tunggal. Using machine learning technologies to classify and predict heart disease. *International Journal of Advanced Computer Science and Applications*, 12:2021, 03 2021.

[2] Priti Chandra, BL Deekshatulu, et al. Prediction of risk score for heart disease using associative classification and hybrid feature subset selection. In *2012 12th International Conference on Intelligent Systems Design and Applications (ISDA)*, pages 628–634. IEEE, 2012.

[3] M.A.JABBAR B.L Deekshatulu Priti Chndra. Alternating decision trees for early diagnosis of heart disease. In *Proceedings of International Conference on Circuits, Communication, Control and Computing*, pages 322–328, 2014.

[4] H Benjamin Fredrick David. Impact of ensemble learning algorithms towards accurate heart disease prediction.

[5] Joshua Emakhu, Sujeet Shrestha, and Suzan Arslanturk. Prediction system for heart disease based on ensemble classifiers.

[6] MA Jabbar. Prediction of heart disease using k-nearest neighbor and particle swarm optimization. *Biomedical Research (0970-938X)*, 28(9), 2017.

[7] Julia H. Miao Kathleen H. Miao and George J.Miao. Diagnosing coronary heart disease using ensemble machine learning. In *International Journal of Advanced Computer Science and Applications (IJACSA)*, pages 30–39, 2016.

[8] Rajwant Kaur and Sukhpreet Kaur. Prediction of heart disease based on risk factors using genetic svm classifier. pages 205–208, 2015.

[9] B.L.Deekshatulu M.A.Jabbar and Priti Chandra. Intelligent heart disease prediction system using random forest and evolutionary approach. In *Journal of Network and Innovative Computing ISSN 2160-2174 Volume 4*, pages 175–184, 2016.

[10] Madhumita Pal and Smita Parija. Prediction of heart diseases using random forest. In *Journal of Physics: Conference Series*, volume 1817, page 012009. IOP Publishing, 2021.

[11] DHYAN CHANDRA Yadav and SAURABH Pal. Prediction of heart disease using feature selection and random forest ensemble method. *International Journal of Pharmaceutical Research*, 12(4):56–66, 2020.

[12] Indu Yekkala, Sunanda Dixit, and M. A. Jabbar. Prediction of heart disease using ensemble learning and particle swarm optimization. In *2017 International Conference On Smart Technologies For Smart Nation (SmartTechCon)*, pages 691–698, 2017.