Health and Technology (2021) 11:63–73
https://doi.org/10.1007/s12553-020-00508-4

ORIGINAL PAPER

# Early prediction model for coronary heart disease using genetic algorithms, hyper-parameter optimization and machine learning techniques

**Priya R. L**[1] · **S. Vinila Jinny**[1] · **Yash Vijay Mate**[2]

Received: 22 July 2020 / Accepted: 5 November 2020 / Published online: 13 November 2020


**Abstract**
Coronary Heart Disease (CHD) is one of the major causes of morbidity and mortality worldwide. According to the World Health Organization (WHO) survey, Cardiac arrest accounts for more deaths annually than any other cause. But the silver lining over here is that heart related diseases are highly preventable, if simple lifestyle modifications are carried out. However, it is a challenging factor to identify high risk heart patients at times due to other comorbidity factors such as diabetes, high blood pressure, high cholesterol and so on. Hence it is needed to develop an efficient early prediction model which can detect high risk patients and their life could be saved. The proposed system helps to identify the best set of features for diagnosis using traditional machine learning algorithms along with modern Gradient Boosting approaches. Genetic algorithm for feature selection to optimize performance by reducing the number of parameters by 20% whilst keeping the accuracy of the model intact is implemented in the proposed system. In addition, hyper parameter optimization techniques are executed to further improve the predictive model's performance.

**Keywords** Genetic algorithm · Evolutionary algorithms · Hyperparameter tuning · Machine learning · Coronary heart disease · Feature selection · Ensemble techniques · Boosting · SMOTE · Optimization · Binary classification · Random forest · Optimized pipeline · TPOT · AutoML · Extreme gradient boosting · Cardiac arrest · Heart attack · Early detection · AI in healthcare

## 1 Introduction

Based on World Health Statistics—2020 report [1] nearly 71% of worldwide mortality happens due to Non Communicable diseases (NCD). There are four major NCD diseases reported for a major rise in mortality rate. Among these, cardiovascular disease (nearly 17.9 million) and chronic respiratory diseases (nearly 3.8 million) continued to be the main causes for the rise in mortality rate. Also, the study measures the age distribution for dying due to cardiovascular and chronic respiratory disease is between 30 and 70 years.

Among all heart diseases, coronary heart disease (CHD), also known as Ischaemic Heart disease is one of the most fatal and challenging factors to prevent any healthy patients. According to the latest study [3], premature mortality in India increased to 59% due to cardiovascular disease and became one of the leading causes of mortality rate. It is hard to identify high-risk patients in cardiovascular disease (CVD) due to the contribution of several other risk factors such as diabetes, high blood pressure, etc. In addition, few other risk factors [4] such as unhealthy living conditions and high levels of stress also contribute to a higher extent of risk for cardiovascular disease.

As cardiac arrest is one of the most significant problems in the healthcare domain, it is essential to define a novel approach in prediction models. In order to reduce the mortality rate of sudden cardiac arrest, it is required to prevent such disease at an earlier stage. In this paper, we have applied a

✉ Priya R. L
priya.rl@ves.ac.in

S. Vinila Jinny
vinijini@gmail.com

Yash Vijay Mate
2017.yash.mate@ves.ac.in

[1] Department of Computer Science and Engineering, Noorul Islam Centre for Higher Education, Kumaracoil, India

[2] Department of Computer Engineering, Vivekanand Education Society's Institute of Technology, Mumbai, India

few machine learning techniques to develop screening tools, classification approaches, and compared with other traditional statistical approaches.

## 2 Background

Cardiac Arrest (CA) leads to spontaneous abnormality and ranges from asymptomatic to symptomatic with recurring chest pain or discomfort. It causes when coronary arteries are narrowed and thus limit the blood and oxygen flow to reach heart muscle [5]. The study [2, 3] says that the major risk factors included are tobacco use, unhealthy lifestyle, abnormal blood lipids, high blood pressure, high diabetes, and so forth. There is a belief.

in medical experts and scientists if CHD is predicted at an earlier stage will reduce mortality and morbidity rate in the country.

Certain statistical analysis methods are used in the prediction outcome of cardiac arrest patients. As discovered in a few studies, measures of Neuron-specific enolase(NSE) [12] and blood lactate levels [9] contribute majorly towards the prediction outcome of cardiac arrest patients. Jonathan Elmer et al. explored in his research using longitudinal model by k-means clustering algorithm and time-invariant patient characteristics data by Bayesian Regression algorithm.

[17] proposed Ten Year Coronary Heart Disease Prediction System using the Framingham's Dataset. The system uses Machine Learning Algorithms like Random Forests, Linear Regression, Support Vector Machines with Linear as well as Radial Basis Kernel Function and Naive Bayes. The maximum accuracy is achieved by Random Forests, which is 84.81%. The system does not take into consideration the class imbalance problem, nor does it use optimization algorithms for feature selection and hyper-parameter tuning.

[13] Proposed pre-arrest prediction tool for In-Hospital Cardiac Arrest patients based on the Good Outcome Following Attempted Resuscitation (GO-FAR) score. [16] applied an integrated approach of machine learning model, Multichannel Hidden Markov by considering a patient's physiological condition along with static risk scores in order to predict high risk cardiac patients by achieving an average sensitivity of 78%. In certain studies, researchers used Wald statistics for risk score calculation to develop predictive cardiac arrest outcome models. The obtained results clearly indicate higher accuracy is achieved especially using ensemble classifier models than traditional machine learning models.

[7] discussed in his study for Out-of Hospital patients risk factors prediction using data mining methods such as Regression analysis, apriori analysis and Classification and Regression tree(CART). The paper [14] used a random forest method with optimization technique to detect CA patients and achieved Area under receiver operating characteristics curve (AUC) values of 95%. S. [10] observed that ensemble techniques provide greater potential than conventional machine learning techniques to design predictive outcome models. Recent study [31] shows early warning to cardiac arrest patients can be given using wearable technology. Based on real-time parameters the system could achieve accuracy of nearly 97% using random forest classifiers.

## 3 Conceptual design

The proposed system is designed to predict coronary heart disease using various machine learning techniques and compares conventional classification algorithms with Modern Gradient Boosting algorithms. The abstract representation of the prediction model is depicted in Fig. 1 as shown below.

As per the Fig. 1, the kaggle dataset on cardiovascular study is fed as raw input to the data cleaning process. In data cleaning, a three steps procedure such as handling missing values, data augmentation and data normalization was processed to remove all unwanted data from the chosen dataset. Then the cleaned dataset was applied to build the training model in three different ways. One option is to apply genetic algorithms as feature selection technique, the second option is to build the model without applying any feature selection technique and the final option is TPOT classifier, used for hyperparameter tuning. Once the model is trained well, it was tested for unseen data and its performance measures are evaluated and compared to choose the best classifier.

## 4 Methodology applied

### 4.1 Advanced feature selection using genetic algorithms

Feature selection is the process through which the attributes having a significant impact on the predictor variable are taken into account while eliminating the irrelevant ones. Feature Selection provides better generalization and lessens the probability of overfitting. It results in a decrease in training time and producing models that are easier to interpret. Genetic Algorithms (GA) [18, 19] are an adaptive search approach that provides robust results than traditional feature selection approaches. GA works by initially exploring unknown search spaces and accumulating the information gained to transcend into subsequent search spaces that have a higher probability of converging to a global optimum.

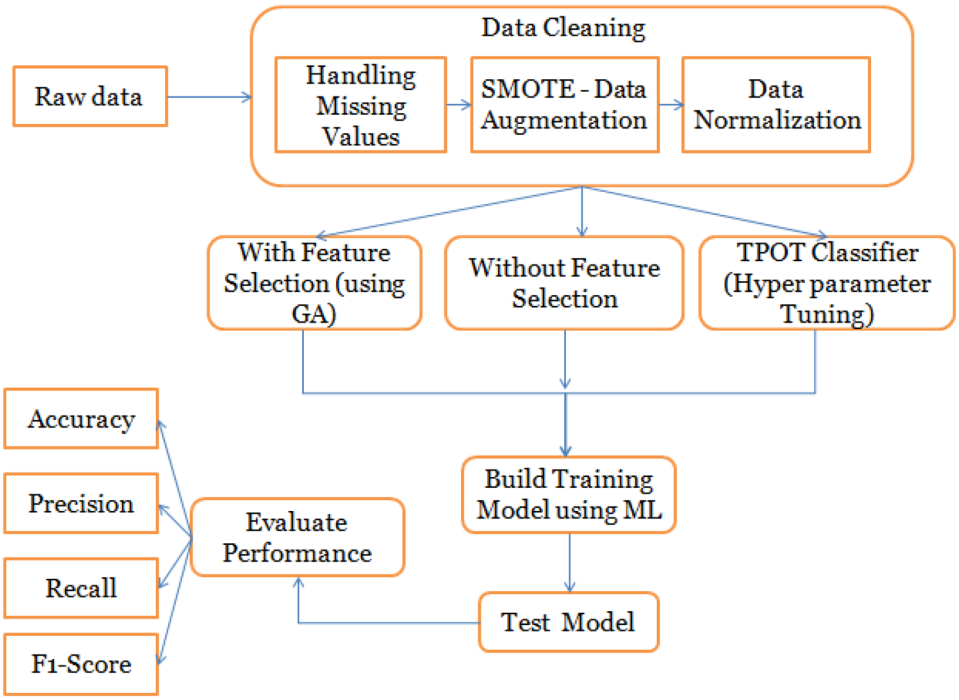


**Fig. 1** Conceptual Design of Early Prediction Model for Cardiac Arrest

#### 4.1.1 Creating an initial population

The initial population is a set of all valid candidate solutions. These candidates are chosen randomly. Since the feature selection problem boils down to either selecting a feature or not selecting it, the candidates are represented by Binary Chromosomes, having values of either zero or one. Because there are a total of 14 features in the dataset, a candidate is represented by a Binary String of length fourteen. The initial population size is set to 50 to predict the CA patients.

#### 4.1.2 Fitness function

The fitness value of an individual is the objective function that is desired to be maximized. It is calculated for every individual for the initial population and its subsequent generations that are created through selection, cross-over, and mutation operations. Since the fitness value of an individual is independent of others, it's calculation is done concurrently for all individuals in the set of population. The problem at hand is to select the most relevant features that have an impact on the target variable and the objective function which is desired to be optimized is the accuracy of the classifiers. Calculation of Fitness Function requires an intuitive understanding of the calculation of the desired objective function.

The chosen dataset is segregated into independent predictors(X) and a target function (y). To introduce optimality in the procedure, instead of just splitting the data into training and test sets, K-Fold Cross Validation methodology [20] is utilized. Data is divided into K equal parts, and the evaluation is conducted K times. Each time, one part is used for testing, and the rest (K-1) parts are available for training the model. In the proposed model, the value of K = 5. A method is designed which takes input as a list of Binary Values of Chromosomes of length equal to the number of features (14) from the given dataset. The representation denotes '1′ for an attribute being selected and '0′ if it is not. The evaluation of the selected set of features from the entire set of attributes is conducted and all those columns are dropped from the original dataset whose corresponding index value is zero.The modified dataset is created by dropping the irrelevant columns. It is then passed to the classifier where K-Fold Cross Validation is performed and performance of the model is evaluated over every data partition. All the accuracies are averaged to return the Mean Accuracy of the Classifier.

The objective function [21] is used to remove the possibility of the number of features selected being zero. A minor penalty factor of 0.001 is introduced to discourage the inclusion of a large number of features. It also acts as a tie- breaker between two candidate solutions that have the same accuracy. In such cases, the solution having a lesser number of features is preferred.

## 4.2 Following are the steps involved in selecting the optimal set of features

### 4.2.1 Selection

The selection marks the beginning of the cycle in the flow of Genetic Algorithms. The selection mechanism picks up individuals from the current generation that would be used as the parents to reproduce a new generation of offspring. The selection is probability-based and gives a higher preference to the individuals having a greater fitness value. This ensures that the population in subsequent generations is more optimum in terms of fitness than the preceding generations. Tournament Selection [22] is a paradigm of Selection method which is particularly beneficial in Feature Selection. In the proposed algorithmic approach, Tournament Size of 2 is employed. Tournament Selection selects two individuals at random from the population and the individual with the greater fitness value is selected.

### 4.2.2 Cross over

Crossover is a vital reason for a striking resemblance between an off-spring and its parents. Without crossover, the genetic information of the parents would be directly cloned into the subsequent generation without the exchange of chromosomes. Feature Selection approach employed for maximizing accuracy demands the application of a special type of cross-over method known as Two-Point Cross Over.

### 4.2.3 Mutation

The mutation is the last genetic operation carried out after Selection and Crossover operations to produce a new generation. Mutation essentially introduces some randomness and variety to the population created. The methodology applied in the proposed system for mutation is called Flip Bit Mutation. It alters a Binary Chromosome in a gene thereby complementing its value.

### 4.2.4 Elitism

Selection, Crossover and Mutation operations ensure that the average fitness of the next generations is greater than the current generation. However, it is also plausible that due to the randomness introduced and the probabilistic nature of selection, the fittest individuals of the current population might not get selected. In the majority of cases, the loss encountered by not selecting the best individual is temporary. The individuals are replaced by fit or even fitter individuals. Even so, to improve the optimality in terms of the time taken by the algorithm to converge, the suggested solution uses an Elitism approach, where the Top N fittest individuals of the population are always picked. N spots in the next generation are occupied by the elite individuals and the rest are picked through selection, crossover and mutation operations. A tremendous reduction in time- complexity is observed as time is saved in re-discovering the optimum solutions lost in the genetic flow.

## 4.3 Applying classification algorithms on the processed data

An appropriate Classification Model is Initialized for the training phase. The research aims at implementing the following Classification Model—Decision Trees, Random Forest, Adaptive Boosting, Gaussian Naive Bayes, Logistic Regression, K—Nearest Neighbors, XGBClassifier, Gradient Boosting Classifier.

### 4.3.1 Decision trees

Decision Trees represent a tree-like hierarchical structure consisting of nodes and branches. The top-most node is called the root node. Each node represents an attribute and each branch represents a decision. The leaf node indicates an outcome. A partition is created based on the attribute value. Partitioning is carried out via a recursive manner known as Recursive Partitioning. Decision trees are capable of handling high- dimensional data with great accuracy. The time complexity depends on the number of variables and the number of records in the dataset. The strategic split in a decision tree is made by using Attribute Selection Measures like Gini Index, Information Gain, and Chi-Square value.

### 4.3.2 Random forest

A random forest can be thought of as a collection of decision trees. The robustness of a Random Forest is proportional to the number of trees it consists of. The algorithm selects some samples at random. It creates separate decision trees for each sample. The decision tree that has the best solution is chosen by means of voting. Relative feature importance could be derived which gives an idea of the features that contribute the most to the predictor variable. The problem of overfitting vanishes as it removes the biases by averaging out all the outcomes.

### 4.3.3 Gaussian naive bayes

It is fundamentally based on the principles of Bayes Theorem. Naive Bayes assumes that all attributes are independent of each other. There is absolutely no correlation between them. A shortcoming arises when the attributes are interdependent, but it still considers them to be independent. The algorithm is fast and reliable on large datasets. The assumption of independence simplifies mathematical computation and hence this algorithm is called naive.

#### 4.3.4 Logistic regression

It is one of the simplest Binary Classification problems where the target variable is dichotomous in nature. It calculates the probability of occurrence of an event based on sigmoid function. It is a special case of Linear Regression, where the outcome of the regression model is mapped to a probability distribution using a sigmoid function. Maximum Likelihood Estimation estimates the set of parameters that contribute the most in predicting the target variable. The target variable follows Bernoulli distribution. The sigmoid function produces an S-shaped curve that maps the target value between 0 and 1. If the value is above 0.5, then we map it to 1 and to 0 in cases where the value is less than or equal to 0.5.

#### 4.3.5 K-nearest neighbors (KNN)

KNN is a widely used non-parametric lazy learning algorithm. Non-parametric because it makes no assumption about the input data distribution. The training phase of KNN is fastest as compared to any other ML algorithm which is characterized by its lazy nature. The ML model for KNN is built in the testing phase itself.K in KNN indicates the number of nearest neighbors. In the KNN algorithm implemented, K is set to 3. For predicting the class of a new data point, the KNN algorithm finds its K nearest neighbors using distance metrics like Euclidean Distance or Manhattan Distance. Voting is carried out and the class that has gotten the majority votes is selected as the prediction.

### 4.4 Boosting algorithms

Boosting algorithms [23] are based on the principle that a combination of weak classifiers gives rise to a stronger classifier having greater accuracy than its base classifiers that it's originally composed of. This combination strategy is known as the ensemble method. Ensemble methods like AdaBoost, XGBoost, Gradient Boosting achieve greater accuracy as compared to non- ensemble methods. Ensemble methods combine the power of Bootstrapping, Boosting, and Stacking to produce powerful classifiers.

## 5 Identification of optimized machine learning pipeline using TPOT classifier

Tree-Based Pipeline Optimization Tool (TPOT) [30] is an Automated Machine Learning Tool that optimizes algorithmic pipelines using Genetic Algorithms. TPOT explores a variety of Machine Learning Pipeline configurations and finds the most optimum set of hyper-parameters. Genetic Programming is utilized by TPOT to find the best hyperparameters and corresponding model ensembles. Genetic Algorithms are implemented for searching the ensemble model from a set of population and then calculating it's fitness by evaluation metrics. Parts of the pipeline and other parameters are modified randomly to ultimately discover the most efficient solution.

TPOT [30, 25] considers many Machine Learning Algorithms like Bernoulli Naive Bayes, Random Forest Classifier, Extra Tree Classifier, Support Vector Machines

**Table 1** Dataset Description

| Attribute Name | Description | Data Type |
|---|---|---|
| male | Gender | NOMINAL |
| age | Age of the patient | CONTINUOUS |
| education | Level of Education<br>1 = Some High School;<br>2 = High School or GED; 3 = Some College or Vocational School; 4 = College | NOMINAL |
| currentSmoker | Whether the patient smokes | NOMINAL |
| cigsPerDay | Number of cigarettes per day | CONTINUOUS |
| BPMeds | Whether the patient was on BP Medications | NOMINAL |
| prevalentStroke | Whether the patient previously suffered from a stroke | NOMINAL |
| prevalentHyp | Whether or not patient was Hyper Tensive | NOMINAL |
| diabetes | Whether or not patient was diabetic | NOMINAL |
| totChol | Total Cholesterol Level | CONTINUOUS |
| SysBP | Systolic Blood Pressure | CONTINUOUS |
| DiaBP | Diastolic Blood Pressure | CONTINUOUS |
| BMI | Body Mass Index | CONTINUOUS |
| HeartRate | Heart Rate | CONTINUOUS |
| Glucose | Blood Glucose Levels | CONTINUOUS |
| Coronary Heart Disease | 10-Year Risk of Developing Coronary Heart Disease (CHD) | NOMINAL (TARGET) |

**Table 2** Count of missing Values for Each Attribute

| | |
|---|---|
| male | 0 |
| age | 0 |
| education | 105 |
| currentSmoker | 0 |
| cigsPerDay | 29 |
| BPMeds | 53 |
| prevalentStroke | 0 |
| prevalentHyp | 0 |
| diabetes | 0 |
| totChol | 5Q |
| sysBP | 0 |
| diaBP | 0 |
| BMI | 19 |
| heartRate | : |
| glucose | 383 |
| TenYeartHD | 0 |
| dtype: int64 | |

(SVM), Gradient Boosting Classifier, and a variety of others along with the multiple ways to stack these algorithms and varying their corresponding hyperparameters. TPOT takes into account data preprocessing steps like Imputation, Feature Selection, Principal Component Analysis (PCA), etc. to improve its performance.

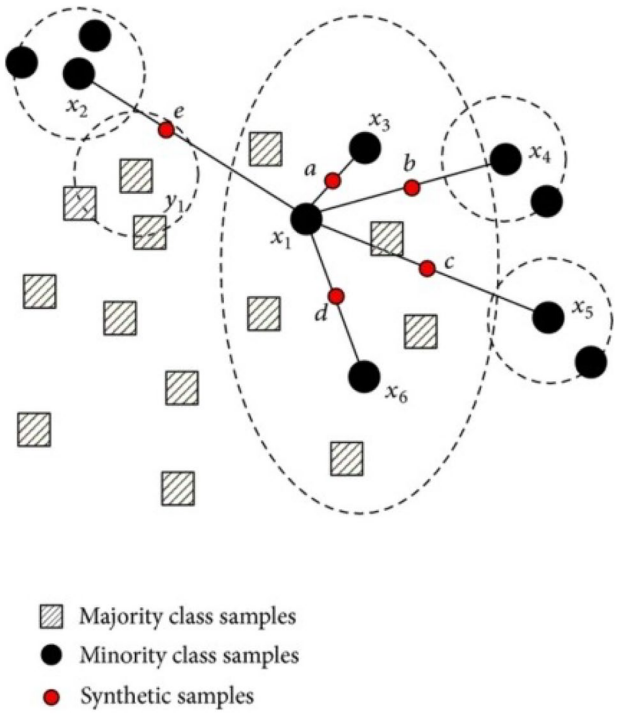


**Fig. 2** Visual Representation of SMOTE [27]

## 6 Implementation Details

### 6.1 Dataset collection

The dataset used for the proposed system is acquired from Framingham's Heart Study Dataset. It consists of attributes that describe Demographic, Behavioral, and Educational information about the patient along with the data on previous medical conditions and the current medical condition of the patient. The dataset comprises 3179 patients without CHD and 572 patients with CHD. The list of attributes, its description and data type are specified in Table 1.

## 7 Data cleaning

Data Cleaning is essential to remove the missing values in the dataset to make it compatible for building Machine Learning Models. The class imbalance problem is addressed and data normalization techniques are used to preprocess the data.The number of records after preprocessing is close to 6202 rows. For model building and evaluation, the dataset is split into 80% data for training and the remaining for testing. A random_seed parameter is taken into consideration while calling the 'train_test_split' function in scikit learn to get reproducible results.

### 7.1 Missing values

Tabular Data illustrated below (as in Table 2) highlights the number of missing values from each attribute. Glucose, Number of Cigarettes Per Day, Total Cholesterols, BP Meds are some of the columns that involve the majority of the missing data in the entire dataset. ML algorithms essentially restrict NaN values in the data and to overcome any further complications, the records with a missing value in any column are simply dropped out Table 2.

A major drawback that could potentially hamper the performance of an ML algorithm is an imbalanced dataset. When the number of records belonging to a particular class is much greater as compared to its counterpart, then the problem of overfitting arises, wherein the algorithm is biased towards the majority class and fails to account for the features in the minority class. After dropping the NaN values, the Framingham's dataset consists of 3101 examples of the patient not developing the risk of CHD and merely 557 examples of the patient developing CHD. Although predicting the risks of CHD (minority) is paramount, ignoring this could result in poor performance on the minority class.

**Synthetic Minority Oversampling Technique** (SMOTE) [26] is a Data Augmentation Technique that synthesizes new examples from the minority class and adds unique information to the model instead of just duplicating the records. The visual representation of SMOTE is depicted in Fig. 2.

```
Generation 1 - Current best internal CV score: 0.8939733375564435
Generation 2 - Current best internal CV score: 0.8955856316798233
Generation 3 - Current best internal CV score: 0.8955856316798233
Generation 4 - Current best internal CV score: 0.8998178783744274
Generation 5 - Current best internal CV score: 0.9000207094825067
Generation 6 - Current best internal CV score: 0.9044545690803366
Generation 7 - Current best internal CV score: 0.9044545690803366
Generation 8 - Current best internal CV score: 0.9044545690803366
Generation 9 - Current best internal CV score: 0.908687018809083
```

**Fig. 3** Generation-Wise CV Score for First Run

SMOTE [27] uses KNN methodology for Upsampling Minority Class instances. SMOTE selects a minority class instance from the feature space. Let's assume this point to be A. The algorithm then selects its K Nearest Neighbors. In the proposed technique K = 5. A sample is selected at random from the K points in the feature space. Let's call this point as B. A synthetic instance is chosen from a line connecting the points A and B. The records thus generated are synthetic instances of the convex combinations of the chosen instances A and B. The examples generated are plausibly close to the feature space of the existing examples of minority classes. After applying SMOTE in the chosen dataset is now balanced with each class having an equal number of examples i.e. 3101 records for each class.

## 8 Data normalization

Differences in the range of values of variables can lead to high generalization errors and make the model highly unstable. Normalization [29] involves rescaling the data within the range of 0 to 1. MinMax Scaler is an effective way for Normalization which preserves the original data distribution. It doesn't alter the original information present in the dataset. For an attribute, MinMax Scaler subtracts the minimum value and divides it by the range. The formula to calculate the MinMax scaler is given as below:

## 9 Training model

The problem for the prediction of Coronary Heart Disease involves a target variable that is categorical in nature. The variable that is to be predicted involves two classes represented by zeros and ones. Therefore, the predictive problem can be classified as a Binary Classification Problem. Following machine learning Classification algorithms are implemented in the course of this research. The performances of these classification models are evaluated by metrics like accuracy, precision, recall, and F1-Score, discussed in the next section. Further improvement in the model is achieved through Feature Selection using Genetic Algorithms and identification of the most optimal pipeline and hyperparameters is carried out using a Tree-Based Pipeline Optimization Tool.

### 9.1 Tree-based pipeline optimization tool (TPOT) classifier

The Fig. 4 shows a few parameters that the TPOT classifier taken into consideration are listed. Let us consider the Generations as

```
Generation 1 - Current best internal CV score: 0.8927614267615243
Generation 2 - Current best internal CV score: 0.8927614267615243
Generation 3 - Current best internal CV score: 0.8927614267615243
Generation 4 - Current best internal CV score: 0.901028367930351
Generation 5 - Current best internal CV score: 0.901028367930351
Generation 6 - Current best internal CV score: 0.9020346051392002
Generation 7 - Current best internal CV score: 0.9086884400480784
Generation 8 - Current best internal CV score: 0.9086884400480784
Generation 9 - Current best internal CV score: 0.9086884400480784
Generation 10 - Current best internal CV score: 0.9086884400480784
Generation 11 - Current best internal CV score: 0.9111053584770815
Generation 12 - Current best internal CV score: 0.9119109979534159
Generation 13 - Current best internal CV score: 0.9119109979534159
Generation 14 - Current best internal CV score: 0.9133216791735699
Generation 15 - Current best internal CV score: 0.9133216791735699
Generation 16 - Current best internal CV score: 0.9133216791735699
Generation 17 - Current best internal CV score: 0.9143305558262677
Generation 18 - Current best internal CV score: 0.9143305558262677
Generation 19 - Current best internal CV score: 0.9143305558262677
```

**Fig. 4** Generation-Wise CV Score for Second Run

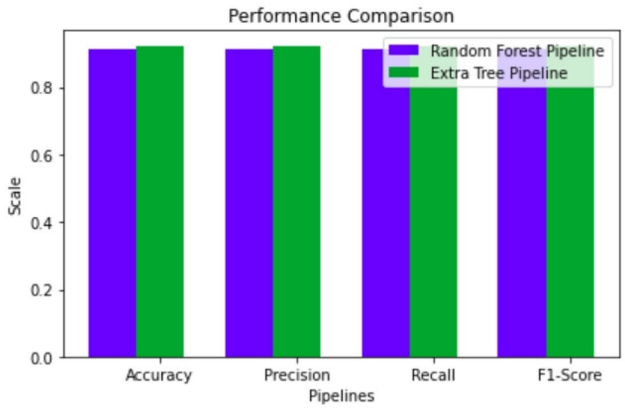


**Fig. 5** Performance Comparison of the Generated Pipelines

Number of iterations for optimization of pipeline, population as Number of individuals to retain in a given population and Offspring as Number of offsprings to produce in the subsequent generation. In default, OFFSPRING_SIZE is considered as equal as POPULATION_SIZE. To evaluate the number of pipelines, the following formula is used.

NUMBER OF PIPELINES EVALUATED = POPULATION_SIZE + GENERATIONS x OFFSPRING_SIZE.

The Fig. 3 depicts the CV score obtained in each generation during the first run. In the proposed system, during the first run the number of generations attained is 10 and population size is 50. Hence the number of pipelines evaluated is given as:

Number of Pipelines evaluated = 50 + 10*50 = 550

Identification of Random Forest Pipeline with following parameters:

Random Forest Classifier (Polynomial Features (Standard Scaler (input_matrix), degree = 2, include_bias = False, interaction_only = False), bootstrap = False, criterion = entropy, max_features = 0.15000000000000002, min_samples_leaf = 1, min_samples_split = 14, n_estimators = 100).

The experiment result depicts the best pipeline after the first run is the Random Forest classifier by achieving 91% accuracy, precision of 92% with Recall as 91% and F1 score as 91%. For further optimization, we set the number of generations as 20 and Population size as 100. Hence the Number of pipelines evaluated as 2100. The Fig. 4 describes the CV score obtained in each generation during the second run.

Identification of Extra Tree Classifier Pipeline with following parameters:

ExtraTreesClassifier (PolynomialFeatures(RobustScaler(MinMaxScaler(input_matrix)), degree = 2, include_bias = False, interaction_only = False), bootstrap = False, criterion = entropy, max_features = 0.25, min_samples_leaf = 2, min_samples_split = 5, n_estimators = 100).

The experiment result depicts the best pipeline after the second run is the ExtraTrees classifier by achieving 92% accuracy, precision of 92% with Recall as 92% and F1 score as 92%. The comparison of these two pipelines performance measures are depicted in Fig. 5.

The performance comparison suggests that the ExtraTreeClassifier Pipeline is marginally better than RandomForestClassifier Pipeline. At Least one percent improvement in accuracy as compared to the previously optimized pipeline is observed when optimized for a population size of 100 for 20 generations.

**Table 3** Performance Comparison of Classification Algorithms Without Feature Selection

| Algorithms Used | Accurac y | Precisio n | Recal l | F1-Score |
|---|---|---|---|---|
| Decision Tree | 0.838 | 0.838 | 0.837 | 0.837 |
| Random Forest | 0.916 | 0.916 | 0.916 | 0.916 |
| AdaBoost | 0.869 | 0.873 | 0.870 | 0.869 |
| Gaussian Naive Bayes | 0.596 | 0.641 | 0.604 | 0.570 |
| Logistic Regression | 0.684 | 0.683 | 0.683 | 0.683 |
| KNN | 0.869 | 0.871 | 0.828 | 0.868 |
| XGBClassifier | 0.879 | 0.884 | 0.880 | 0.878 |
| Gradient Boosting Classifier | 0.883 | 0.889 | 0.885 | 0.883 |

## 10 Evaluation metrics

Once the prediction model is trained well on the training dataset, it is vital to evaluate the performance of the model on unknown data. The unknown data is called testing data. The evaluation metrics are the performance indicators of a model that help in determining if the model is worthy to be utilized for real-word use-cases. The evaluation metrics used are Accuracy, Precision, Recall, and F1-Score and these metrics are calculated using the below mentioned formulae:

**Table 4** Mapping Indexes to Features

| X | |
|---|---|
| 0 | Male |
| 1 | Age |
| 2 | Education |
| 3 | Current Smoker |
| 4 | CigsPerDay |
| 5 | BPMeds |
| 6 | PrevalentStroke |
| 7 | PrevalentHyp |
| 8 | Diabetes |
| 9 | totChol |
| 10 | sysBP |
| 11 | diaBP |
| 12 | BMI |
| 13 | Heart Rate |
| 14 | Glucose |

**Table 5** Performance Comparison of Classification Algorithms after Feature Selection Using Genetic Algorithm

| Algorithms Used | Accurac y | Precision | Recall | F1-Score | No. of Feature s | Feature Indexes |
|---|---|---|---|---|---|---|
| Decision Tree | 0.887 | 0.907 | 0.891 | 0.886 | 3 | 0,1,6 |
| Random Forest | 0.907 | 0.908 | 0.908 | 0.907 | 12 | 0,1,2,3,4,7,8,9,10,11,13,14 |
| AdaBoost | 0.855 | 0.868 | 0.0.860 | 0.855 | 6 | 0,2,3,5,7,1 |
| Gaussian Naive Bayes | 0.627 | 0.658 | 0.636 | 0.636 | 6 | 1,5,7,8,10,12 |
| Logistic Regression | 0.704 | 0.704 | 0.704 | 0.704 | 7 | 0,1,3,5,7,8,10 |
| KNN | 0.845 | 0.851 | 0.845 | 0.845 | 4 | 0,1,3,6 |
| XGBClassif ier | 0.852 | 0.866 | 0.857 | 0.852 | 3 | 0,1,2 |
| Gradient Boosting Classifier | 0.868 | 0.886 | 0.873 | 0.867 | 7 | 0,2,3,4,5,6,7 |

Accuracy = (TP + TN)/ (TP + TN + FP + FN).

Precision = TP/(TP + FP) Recall = TP / (TP + FN).

F1-Score = (2 * Precision * Recall) / (Precision + Recall)

The Table 3 describes the performance comparison of various classification algorithms without applying the feature selection techniques. It clearly indicates that among all, random forest obtained with the highest accuracy of 91.6% followed by Gradient Boosting Classifier and XBG Classifier obtained accuracy of nearly 88% compared with other traditional machine learning algorithms.

## 10.1 Comparison of performance measures with feature selection techniques

We also implemented feature selection using GA and its indexes are mapped to 15 features as shown in Table 4. The Table 5 indicates the comparison of performance measures after applying the feature selection techniques. It clearly indicates that among all, random forest classifier obtained with the highest accuracy of nearly 91% by considering the maximum number of features for classification model as 12. Then followed by Decision Tree classifier achieved accuracy of nearly 88%, but the precision obtained by decision tree is almost same as Random Forest classifier by considering only 3 features. Gradient Boosting Classifier and XBG Classifier obtained accuracy of nearly 86% compared with other traditional machine learning algorithms.

The performance evaluation metrics such as accuracy, precision, recall and F1-Score of various machine learning classifiers are compared with and without applying feature selection techniques such as Genetic Algorithm are represented in the Fig. 6, 7, 8, 9 respectively.

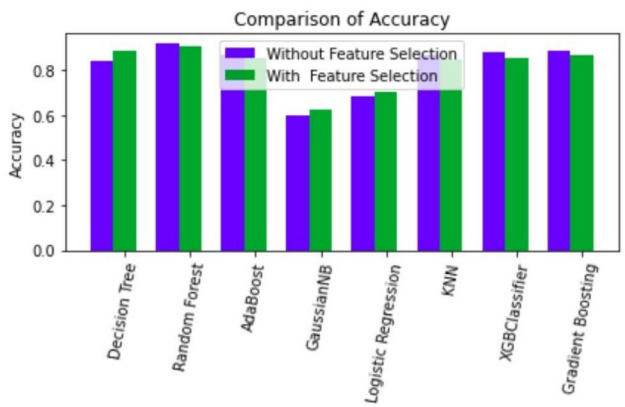


**Fig. 6** Comparison of Accuracy with and without feature selection

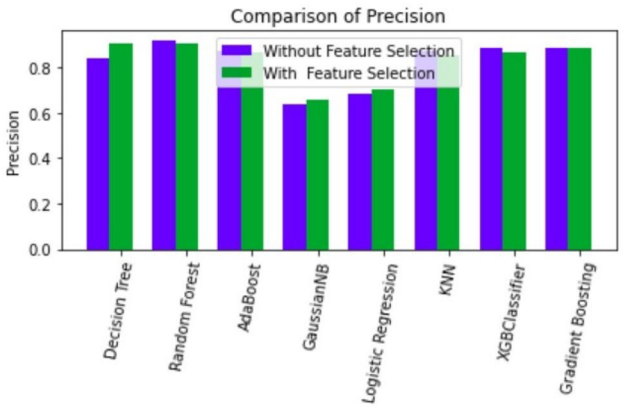


**Fig. 7** Comparison of Precision with and without feature selection

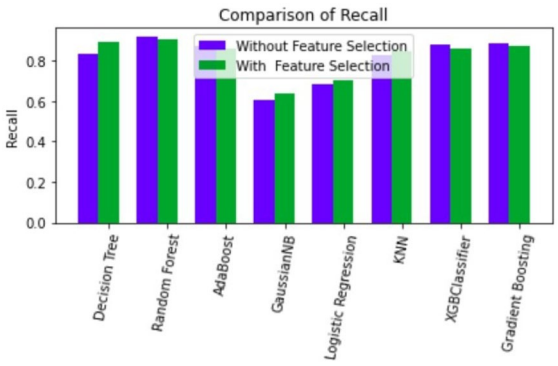


**Fig. 8** Comparison of Recall with and without feature selection

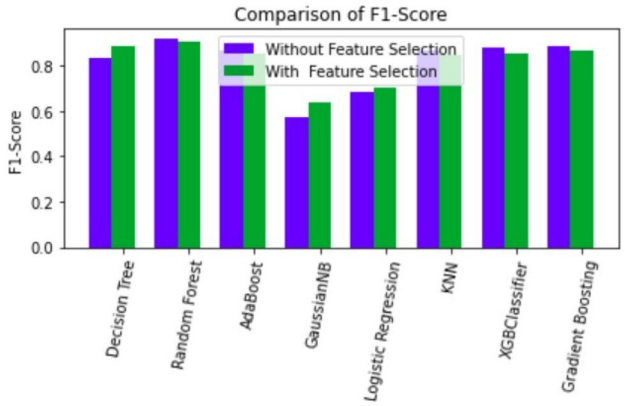


**Fig. 9** Comparison of F1-Score with and without feature selection

## 11 Discussion

The use of Machine Learning techniques to predict one of the major chronic morbidity with high mortality rate of Coronary Heart Diseases provides an accurate estimate as compared to the traditional statistical and mathematical modeling approaches. The target variable in Framingham's Dataset represents if the person could encounter the risk of CHDs in a ten-year time frame. In the tabular representation, it is represented by 0 s and 1 s. A zero indicates that a person is safe from the risk and one indicating a risk of contracting the disease. To get things in perspective, the problem is categorized as a Binary Classification Problem. For Classification, both traditional Classification Algorithms like Logistic Regression, Decision Trees, Random Forests, Gaussian Naive Bayes along with Modern Gradient Boosting Approaches like XGBClassifier, Adaptive Boosting (AdaBoost) and other Ensemble Methods are used. Ensemble Learning harnesses the power of weak classifiers to combine and form a robust classifier to achieve a tremendous improvement over the state-of-the-art approaches.

## 12 Conclusion

The Cardiac Muscle in the heart is one of the hardest working muscle groups in the entire human body. Beating over 72 times in a minute and more than 3 billion times in a lifetime, a salutary heart can sustain various biological functions. Prediction of 10-year risk of contracting Coronary Heart Diseases (CHDs) is therefore crucial for a prolonged life. The proposed system identifies the best set of hyper-parameters for Extra Tree Classifier to achieve an accuracy of over 92%. Traditional Random Forest Algorithm with the number of estimators parameter set to 100 has an accuracy of over 91%. Although these models consider all the attributes of the dataset for building the model, a refined Genetic Algorithm approach for selecting the set of features that have the greatest influence on the target variable is achieved through selection, crossover, and mutation and elitism operations. Feature selection reduces the number of parameters selected to twelve from fifteen for Random Forest while achieving an accuracy of over 90%.

**Funding** The above study was not funded by any organization.

### Compliance with Ethical Standards

**Conflict of Interest** There is no conflict of interest among the authors for the submitted manuscript.

## References

1. World health statistics 2020: monitoring health for the SDGs, sustainable development goals. Geneva: World Health Organization; 2020. Licence: CC BY- NC-SA 3.0 IGO.
2. Kow CS, Zaidi STR, Hasan SS. Cardiovascular Disease and Use of Renin-Angiotensin System Inhibitors in COVID-19. American Journal of Cardiovascular Drugs. 2020. https://doi.org/10.1007/s40256-020-00406-0.
3. Global Burden of Cardiovascular Disease. Cardiovascular Diseases in India - Current Epidemiology and Future Directions. Centre for Control of Chronic Conditions, Public Health Foundation of India, Gurgaon, India (D.P., P.J.); and All India Institute of Medical Sciences, New Delhi, India (A.R.) https://doi.org/10.1161/CIRCULATIONAHA.114.008729
4. Chandola T. Ethnic and class differences in health in relation to British South Asians: using the new National Statistics Socio-Economic Classification. Soc Sci Med. 2001;52:1285–96.
5. Patel K, Hipskind JE. Cardiac Arrest. [Updated 2020 Jan 21]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2020 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK534866/.
6. World Health Organization. Prevention of cardiovascular disease : guidelines for assessment and management of total cardiovascular risk, ISBN 978 92 4 154717 8. 2007.
7. Chang C-C, Kao J-H, Hsu C-Y, Liaw H-T, Wang T-C. Data Mining Technology Combined with Out-of- Hospital Cardiac Arrest. Symptom Association and Prediction Model Probing. 2018. https://doi.org/10.1007/978-981-10-7398-4_31.

8. Elmer J, Jones BL, Nagin DS. Comparison of parametric and nonparametric methods for outcome prediction using longitudinal data after cardiac arrest. Resuscitation. 2020;148:152–60. https://doi.org/10.1016/j.resuscitation.2020.01.020.
9. Uyar H, Yesil E, Karadeniz M, et al. The Effect of High Lactate Level on Mortality in Acute Heart Failure Patients With Reduced Ejection Fraction Without Cardiogenic Shock. Cardiovasc Toxicol. 2020;20:361–9. https://doi.org/10.1007/s12012-020-09563-9.
10. Layeghian Javan S, Sepehri MM, Layeghian Javan M, Khatibi T. An intelligent warning model for early prediction of cardiac arrest in sepsis patients. Comput Methods Programs Biomed. 2019;178:47–58. https://doi.org/10.1016/j.cmpb.2019.06.010.
11. Usha M, Debabrata S, Dilip C. IoT-Based Cardiac Arrest Prediction Through Heart Variability Analysis. 2020. https://doi.org/10.1007/978-981-15-1483-8_30.
12. Luescher T, Mueller J, Isenschmid C, et al. Neuron-specific enolase (NSE) improves clinical risk scores for prediction of neurological outcome and death in cardiac arrest patients: Results from a prospective trial. Resuscitation. 2019;142:50–60. https://doi.org/10.1016/j.resuscitation.2019.07.003.
13. Piscator E, Göransson K, Forsberg S, et al. Prearrest prediction of favourable neurological survival following in-hospital cardiac arrest: The Prediction of outcome for In-Hospital Cardiac Arrest (PIHCA) score. Resuscitation. 2019;143:92–9. https://doi.org/10.1016/j.resuscitation.2019.08.010.
14. Seki T, Tamura T, Suzuki M, SOS-KANTO 2012 Study Group. Outcome prediction of out-of-hospital cardiac arrest with presumed cardiac aetiology using an advanced machine learning technique. *Resuscitation*. 2019;141:128–135. https://doi.org/10.1016/j.resuscitation.2019.06.006
15. Nachiket T, Tushar L, Damodar R, Venkatanaresh K. Prediction of cardiac arrest recurrence using ensemble classifiers. Sadhana - Academy Proceedings in Engineering Sciences. 2017;42:1–7. https://doi.org/10.1007/s12046-017-0683-z.
16. Akrivos E, Vasilios P, Maglaveras N, Ioanna C. Prediction of Cardiac Arrest in Intensive Care Patients Through Machine Learning. 2018. https://doi.org/10.1007/978-981-10-7419-6_5.
17. Rubini PE, Deeksha GS, Varshaa Shree B, Deepa N, Srivastava A. A Cardiovascular Disease Prediction using Machine Learning Algorithms. International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249–8958, Volume-8 Issue-5, June 2019.
18. Lingaraj H. A Study on Genetic Algorithm and its Applications. International Journal of Computer Sciences and Engineering. 2016;4:139–43.
19. Chaikla N, Qi Y. Genetic algorithms in feature selection. IEEE SMC'99 Conference Proceedings. 1999 IEEE International Conference on Systems, Man, and Cybernetics (Cat. No.99CH37028), Tokyo, Japan, 1999, pp. 538–540 vol.5, https://doi.org/10.1109/ICSMC.1999.815609.
20. Paek A, Agashe H, Contreras-Vidal J. Decoding repetitive finger movements with brain activity acquired via non- invasive electroencephalography. Frontiers in neuroengineering. 2014;7:3. https://doi.org/10.3389/fneng.2014.00003.
21. Chehouri A, Younes R, Perron J, Ilinca A. A Constraint- Handling Technique for Genetic Algorithms using a Violation Factor. Adam Chehouri et al. / Journal of Computer Sciences. 2016;12(7):350.362. https://doi.org/10.3844/jcssp.2016.350.362.
22. Genetic Algorithms, tutorialspoint by Tutorials Point (I) Pvt. Ltd (2016) Available from : https://www.tutorialspoint.com/genetic_algorithms.
23. Bühlmann P. Bagging. Boosting and Ensemble Methods: Handbook of Computational Statistics; 2012. https://doi.org/10.1007/978-3-642-21551-3_33.
24. Fu W, Olson R, Nathan, Jena G, PGijsbers, Augspurger T, Carnevale R. (2020, June 1). EpistasisLab/tpot: v0.11.5 (Version v0.11.5). Zenodo. https://doi.org/10.5281/zenodo.3872281.
25. Olson RS, Bartley N, Urbanowicz RJ, Moore JH. Evaluation of a Tree-based Pipeline Optimization Tool for Automating Data Science. Proceedings of GECCO 2016. 2015;485-492.
26. Shidha MV, Mahalekshmi T. An Empirical Study on the Effect of Resampling Techniques in Imbalanced Datasets for Improving Consistency of Classifiers. International Journal of Applied Engineering Research ISSN 0973–4562 Volume 14, Number 7 (2019) pp. 1516–1525 © Research India Publications. https://www.ripublication.com.
27. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, AI Access Foundation, June 2002, ISSN : 1076–9757, volume - 16, pg.nos. 321–357, https://doi.org/10.1613/jair.953.
28. He H, Ma Y. Imbalanced Learning - Foundations, Algorithms and Applications. Wiley-IEEE Press; 1 edition (July 1, 2013), ISBN: 978–1118074626.
29. Kjell J, Max K. Applied Predictive Modeling. Springer 5th Edition 2016 ISBN: 978-1461468486 Page 30. https://doi.org/10.1007/978-1-4614-6849-3.
30. Le TT, Fu W, Moore JH. Scaling tree-based automated machine learning to biomedical big data with a feature set selector. Bioinformatics. 2020;36(1):250–256.
31. Mukherjee R, Ghorai SK, Gupta B, et al. Development of a Wearable Remote Cardiac Health Monitoring with Alerting System. Instrum Exp Tech 2020;63:273–283. https://doi.org/10.1134/S002044122002013X.