International Journal on Advanced Science Engineering Information Technology

Vol.12 (2022) No. 5
ISSN: 2088-5334

Heart Disease Prediction Using Genetic Algorithm with Machine Learning Classifiers

Basant Abdel Menem Metwally [a,*], Nagham Elsayed Mekky [a], Ibrahim Mahmoud Elhenawy [b]

[a] *Faculty of Computer and Information Science, Mansoura University, Mansoura 35516, Egypt*
[b] *Faculty of Computer and Information Sciences, Zagazig University, Zagazig 44519, Egypt*
*Corresponding author: *eng.basant_fci88@hotmail.com*

Abstract —**Based on the world health organization (WHO), heart disease is the major reason for the loss of life everywhere on earth. Heart attacks are the leading reason of loss of life among heart diseases. This disease is called a silent disease. The person does not feel any pain until the last level of sickness and may arrive at death if not saved at the right time. The datasets for this disease are becoming available, so it is a particularly good branch of study. Predicting a heart attack for a medical practitioner is difficult since it requires increased expertise. However, over the last few decades, resolving complicated, extremely non-linear classification and prediction problems have been using machine learning algorithms (ML). Hence, it is feasible to establish a prediction model that would see the existence or nonexistence of heart disease based on many heart-related symptoms (features). The essential contribution of this research is to introduce various prediction models for heart disease using a genetic algorithm (GA) to find optimal features combined with classical machine learning techniques. The optimized prediction model uses a genetic algorithm that performs better than classical models. The execution of the algorithms is tested using Cleveland and Framingham datasets. The prediction models' performance is standardized using three evaluation criteria: accuracy, precision, and recall. The proposed system showed superior performance compared with other related systems. It reached an accuracy of 100% for the Cleveland dataset and 91.8% for the Framingham dataset.**

Keywords — Heart attacks; genetic algorithm; machine learning; K-Nearest Neighbor classifier; random forest.

Manuscript received 11 Aug. 2021; revised 11 Jul. 2022; accepted 12 Sep. 2022. Date of publication 31 Oct. 2022.

I. INTRODUCTION

Heart disease is a grave illness, and determining this disease is a challenging mission in the first period. The circulatory system is impacted by many diseases that apply to the term heart disease, which consists of the heart and blood vessels. Dealing with the circumstance ordinarily named "Heart Attack" and the attributes that conduct to like circumstance has been very important. Some classifications of heart diseases are Cardiomyopathy and cardiovascular disease. The expression cardiovascular disease covers outspread domain circumstances that influence the human body that is loaded with blood distributed throughout the blood arteries, the heart, and other organs. Cardiovascular disease (CVD) causes serious illnesses, functional decline, and fatalities. Coronary heart disease (CHD) happens because narrowing the coronary arteries causes drooping of blood and oxygen nutrition in the heart, chest discomfort, angina pectoris, and heart attacks surround these myocardial infarctions in CHD. The heart attack results in a blood clot, generally because of An unexpected coronary artery blockage. The pains in Chest happen because the blood taken by the heart muscles is inappropriate. Cardiovascular disease can take many distinct forms, including excessive blood pressure, coronary artery disease, valvular heart disease, stroke, and rheumatic fever/rheumatic heart disease [1].

Many people are unrealizable when heart disease happens to them due to many factors like lipid level, which scale the number of triglycerides and the level of cholesterol in the blood. Actuality, up to 25% of humans with heart illness have no syndrome or signs despite minus blood flow to the heart. This case is called silent heart disease. Many complications like heart failure happen when the ability to pump enough blood by the heart is not well that the body wants to work fine. The last stage of the disease is until heart disease patients felt sick, and the damages have become irretrievable for saving the patients [2].

In the present age, the number is on an ascent of persons suffering from heart disease. Heart disease results in a huge number of people losing their lives every year worldwide. However, proper diagnosis in the first period of disease followed by appropriate treatment can save many lives of patients. Unfortunately, obtaining a precise diagnosis of cardiac disorders has never been a simple task. The identification of cardiac problems might become complicated if the proper diagnosis is delayed due to several factors. For instance, several human organs other than the heart are associated with the clinical symptoms, the functioning, and the sick looks of heart disorders, and these conditions frequently manifest as diverse syndromes. Similar symptoms appeared simultaneously in different types of heart diseases. Therefore, it is imperative to develop medical diagnostic decision support systems that can assist clinicians with the diagnostic procedure [3].

The patient might end up with strong ramifications in a short period when the early symptoms of heart disease are ignored. Lack of movement lifestyle and excrescent stress in today's world have increased the worth of the situation [4]. It can be holed under domination if the disease is discovered early. However, playing sports every day and renouncing deleterious habits at the earliest is always recommended. Smoking exhaustion and unhealthy diets increase the opportunities for stroke and heart diseases. Eating at least five aids of fruits and vegetables a day is a nice habit. The intake of salt to one teaspoon per day is recommended for heart disease patients [5].

Numerous research has attempted to predict the development of heart illness using a variety of machine learning techniques, such as classification trees, Naive Bayes (NB), neural networks, support vector machines (SVM), and the K-Nearest Neighbor (KNN) algorithm. A few of the publications also used feature selection methods, including wavelet transformation, principal component analysis, and information gain module, to pinpoint crucial characteristics for a classifier's effective performance in predicting heart disease. Dulhare et al.[6], determined efficient heart disease prediction by applying Naïve Bayes (NB) classifier merged with particle swarm optimization (PSO) feature selection algorithm. VA Long Beach dataset that was from the University of California, Irvine (UCI) was applied to train and test the model processes. The performance of this model showed that the Naïve Bayes with the PSO selection algorithm was enhanced to 87.91%. The basic NB execution enhanced heart disease prediction accuracy by NB combined with PSO model.

Ayatollahi et al. [7] attended an Artificial neural network (ANN), and SVM classification methods were compared based on the positive predictive value (PPV) of cardiovascular diseases. The dataset was obtained from associated hospitals with AJA University of Medical Sciences in Iran. The same characteristics at the University of California, Irvine (UCI) machine learning Cleveland heart disease data policy repository were discovered when getting the dataset. The model's execution revealed that the SVM algorithm outperformed the ANN model, which proposed higher sensitivity, accuracy, and better execution with 92.23% sensitivity.

Lakshmanarao et al. [8] used three machine learning classifiers to foretell heart disease. They used sampling techniques for dealing with unbalanced datasets. To predict the overall threat, various machine learning methods were employed. The dataset for Framingham heart disease was used, which was publicly free on Kaggle. This model achieved an accuracy of 99% in heart disease detection.

Alotaibi et al. [9] introduced comparison research on the classification and prediction of cardiac disease using machine learning techniques. In the Rapid-Miner, the methods Naive Bayes (NB), decision tree (DT), RF, SVM, and logistic regression (LR) were applied. Cleveland heart disease datasets were used. Cross-validation methodology The 10-fold method was used to train the model. The results of this model showed that the DT algorithm and SVM provided predictions of heart illness with the highest accuracy, with accuracies of 93.19 and 92.30 percent, respectively.

IrfanJavid [10], created ML and Deep learning (DL) standardized performance techniques that integrated several ML and DL techniques to afford the highest result and strong technique for diagnosing any potential of owning heart illness. UCI Cleveland heart dataset repository was applied for this experiment. This Ensemble approach was achieved better when the Hard Voting ensemble method was organized with 85.71% accuracy.

Soumonos et al. [11] introduced a model for the prediction of cardiovascular disease with ECG analysis and symptom-based detection. The project proposed two modules: ECG report analysis and prediction of atrial fibrillation with a convolutional neural network. Second Predicting risk of heart disease by a multiclass artificial neural network. Datasets obtained from UCI and Physionet data repositories were applied for implementation training and testing of the modules. The system was a promised accuracy of 97% on the unseen patients" data.

Anna et al. [12] introduced the proposition of a dimensionality decrease function and applied a feature selection technique to find attributes of heart illness. The UCI Machine Learning dataset for heart illness was used, and the database includes 74 qualities and a target verified as true using six ML approaches. The Cleveland, Hungarian, and Cleveland-Hungarian (CH) datasets' accuracy for chi-square and principal component analysis (CHI-PCA) with random forests (RF) was 98.7%, 99.0%, and 99.4%, respectively. The experi-mental output showed that the chi-square combined with PCA gets the best executions in most clas-sifiers. The bad outputs got from applying PCA, which needs more dimensionality to enhance the outputs.

Sangya et al. [13] used performance metrics for comparing different machine learning classifiers. The classifiers used were NB, DT, RF, SVM, KNN, and logistic regression (LR). Each classifier acted better in some attitudes and worse in others. Here dataset was used in Cleveland, which had 303 records of patients along with 14 features. This was done as part of the preprocessing of these datasets: removing all the noise and losing data. And then analyze the preprocessed dataset. Six alternative machine learning algorithms were used in this experiment based on several performance measures. The results of these comparisons revealed that SVM had the best accuracy, coming in at 89.34%.

Rahul et al. [14] examined and compared all the classifiers based on the dataset from UCI for heart disease prediction. It used the most used performance evaluation criteria, including precision, recall, and RMSE (Root Mean Square Error). Naive Bayes (NB), decision trees (DT), RF, SVM, KNN, artificial neural networks (ANN), deep neural networks (DNN), multilayer perceptrons (MLP), and logistic regression are the classifiers that were employed (LR). The best values observed that SVM, Logistic Regression, and ANN had close result accuracy. The outcomes showed that the random forest classifier delivered the most accurate outcomes. With the best precision and recall among all other classifiers, accuracy comes out to be 95.60 percent.

Harshit et al. [15] predicted a patient with cardiovascular disease by eliciting the patient's medical information. It used a dataset from the UCI repository with patients' medical information and features. Each row in this dataset, which had 303 rows and 14 columns, represented a single record. The given model was constructed using methods like KNN, random forest classifier, and logistic regression. This system enhanced medical care and reduced the cost. It was found that the accuracy of KNN is best among the three classifiers, with 88.52%.

Awais et al. [16] presented a procedure named CardioHelp to diagnose heart illness in the early stage to prevent it and know the reasons for it. A machine learning algorithm named Convolutional Neural Networks was used in this procedure. The introduced procedure used a state-of-the-art dataset for this paper. The introduced procedure concerned temporal data modeling by using CNN for HF prediction in its early period. The output performance showed that the accuracy of the introduced framework was 97%. Sumaya et al. [17] determined efficient heart disease prediction and offered suitable medicine very fast by applying Naïve Bayes (NB), LR, DT, RF, Gradient Boosting Classifier, and Linear Support Vector Classifier. After prepressing, 1614 rows of an Austrian medical data collection with 25 attributes were employed. The performance of this model proved that the finest result with 91% accuracy was the Logistic Regression.

Nagaraj et al. [18] applied various classifiers like Naïve Bayes and SVM for heart disease prediction. Cleveland Dataset was gotten from a UCI was used. The comparison between those classifiers showed that SVM with radial kernel was the best accuracy of other classifiers, and this study showed that females were the most likely impacted by heart disease.

Shorewala [19] presented a model of heart disease prediction. This model used a Cleveland dataset consisting of 14 attributes and 303 observations. This model worked on feature standards, and selection features using PCA, and seven principal components were applied for training the machine learning algorithms. As a result of this model, LR and SVM attained an accuracy of 87% and 85%, respectively, as opposed to KNN's accuracy of 69%.

Latha et al. [20] proceed with a comparison for improving the prediction of heart diagnosis by investigating ensemble techniques applied to the Cleveland dataset, consisting of 303 instances. A poor classifier's accuracy increased significantly (7.26%) when ensemble approaches were utilized, and the accuracy of a nine-feature feature selection increased to 85.48% when the NB, BN, RF, and MLP procedures for majority voting were employed.

Mohan et al.[21], built a heart disease prediction model using a hybrid of a linear model and random forest (HRFLM). The dataset used was Cleveland, with 297 cases and 13 attributes. It achieved a high level of accuracy of 88.7% and got the best error rates. This research works to enhance the performance of heart disease prediction to help doctors get better diagnosis for heart disease in early stage, which will save more lives. The introduced model was applied in Cleveland and Framingham datasets and was used genetic algorithm combined with classical machine learning techniques to determine the important feature and get high performance to help medical practitioner for taking decision.

II. Materials and Method

Medical diagnosis can be greatly aided by machine learning techniques since they can provide a knowledge-rich environment. This paper employs five traditional machine learning techniques: KNN, LR, DT, SVM, and RF. They are used sklearn library of python. All algorithms are coded and executed on Intel Corei7 having an 4500U CPU and 16 GB ram processor up to 1.80 GHz.

A. Classical Machine Learning

1) Support Vector Classifier: A support vector classifier is an algorithm generally applied to solve classification problems. It is a discriminatory classifier. A dismissing hyperplane knows it. For a presented categorized training dataset, the Support Vector Classifier produces an optimal hyperplane as the output [22]. This hyperplane classifies new examples. Thus, it constructs a model that appoints the new examples to one classer and another. This model is actually an impersonation of the examples as points in space charted like examples of different classifications are divided by an obvious space. A hyperplane separates the points. Based on which portion of the gap the new instances close, they are projected to belong to a specific classification and are mapped into the same space as the existing examples [14].

2) Decision Tree Classifier: An illustration of every possible decision-making outcome in the form of a decision tree is one that is based on particular circumstances. The decision tree algorithms streamline a chain of test questions and orders. A decision tree's root and internal nodes have feature test conditions to split records with various characteristics. The terminal nodes are labeled 'Yes' or 'No' [23].

3) Random Forest Classifier: Using the Random Forest classifier, one can learn to get a solution for classification, regression, and many other types of problems. Random Forest builds multiple decision trees during training. This classifier's output, The mode of the classes of the individual trees in the classification scenario. The outcome of regression is each tree's mean prediction, which is the result. Random Forest Classifier removes an important drawback of decision trees that overfit to their training sets [14].

4) K-Nearest Neighbor (KNN) Classifier: A straightforward supervised learning technique called the k-nearest neighbor classifier can be applied to solve classification and regression problems. This categorizer depends on the assumption that identical things exist in close propinquity. The accuracy of the results produced by the algorithm is significantly influenced by the factor "k," hence in this work, the optimal value of K has been obtained. This algorithm is versatile but has a drawback: it becomes significantly slower as the number of independent variables increases [24].

5) Logistic Regression: A predictive analytic algorithm is a logistic regression. It is used to solve problems involving binary categorization. It predicts the outcome when the dependent variable is dichotomous, with only two possible outcomes. It is employed to describe data and to clarify the connection between several independent nominal, ordinal, interval, or ratio-level variables and a single binary dependent variable. An algorithm is being used to analyze a data set comprising a dependent variable and one or more independent variables [25].

B. Genetic algorithm (GA)

John Holland founded the rules of genetic algorithms in 1975. A genetic algorithm is one of the famous optimization tools in the evolutionary algorithm's family. A genetic algorithm relies on generating a set of the nominated solutions-chromosomes - and then picking the best solution. Fig. 1 shows the GA framework [26], the method starts with randomly generating chromosomes called population, and then iterative steps are implemented to create a new set of solutions. The iterative steps are mainly: selecting two chromosomes to be parents, mating them to have two offspring, and finally performing a mutation if the probability allows. For each step, there are some methods to implement. For example, binary tournament selection is a powerful method of selecting two individuals at random in preparation for the mating stage. In addition to this method, it is possible to use the selection of the roulette wheel, rank, etc. After the selection steps, the genetic algorithm moves toward another phase, a phase of mating, to generate two offspring by mixing the information of parents in some way, this can be carried out using traditional methods such as 1X, 2X or UX, but the nature of some problems may impose certain requirements and therefore need specific types of mating like PMX, OX, CX, etc.

The mutation process is applied to the individuals produced by the previous step under low probability to maintain the stability of the population. Replacing the values of two genes is a well-known mutation method. The complement approach can be used if the chromosomes are encoded as binary values. Complements mean that the value of selected gen is flipped from 0 to 1 or becomes 1 if it was 0. One way to increase the genetic algorithm's performance is to preserve the n best individual in the old population and use them as a substituent for the n worst individuals in the new population. This cycle continues until the stop criterion is met by reaching a specific number of generations, a goal is satisfied, or no change in performance is observed [27]. Classical machine learning techniques need to be improved to get more accurate predictions. This paper proposes using genetic algorithm as an optimization technique to get more accurate results.

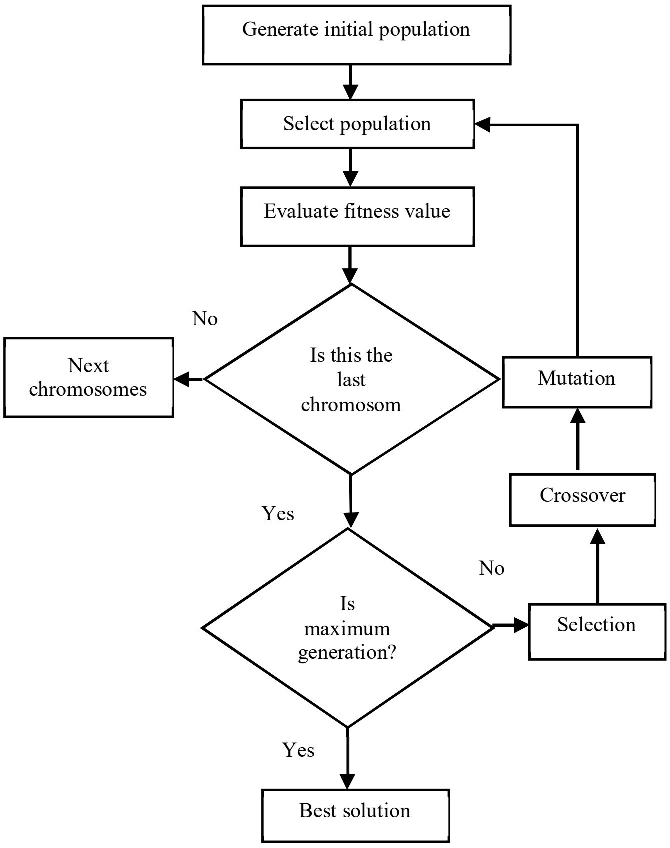


Fig. 1 Genetic Algorithm Framework [26].

C. Data Collection and characterization

This article's datasets were culled from "Kaggle" website. Officially, Heart Disease Dataset is the name of the collection. The Cleveland dataset [28] has 303 distinct cases in all, divided into two subsets based on the presence or absence of heart disease. It consisted of 76 attributes, but mostly all spreading researchers used a subset of 14 from the 76 attributes for the prediction [29]. The Framingham dataset [30] has a total of 4240 unique instances of 16 attributes. Cleveland and Framingham datasets have illustrated some attributes in Table I.

TABLE I
CLEVELAND AND FRAMINGHAM DATASET FEATURES

NO.	Cleveland Dataset Features	Framingham Dataset Features
1	age	male
2	sex	age
3	cp	Education
4	Trestbps	CurrentSmoker
5	Chol	CigsPerDay
6	Fbs	BPMeds
7	Restecg	PrevelantStrok
8	Thalach	PrevelantHyp
9	Exang	Diabetes
10	OldPeak	TotChol
11	Slope	SysBP
12	Ca	DiaBP
13	Thal	BMI
14	Target	Heart rate
15	-----	Glucose
16	-------	Ten YearsCHD

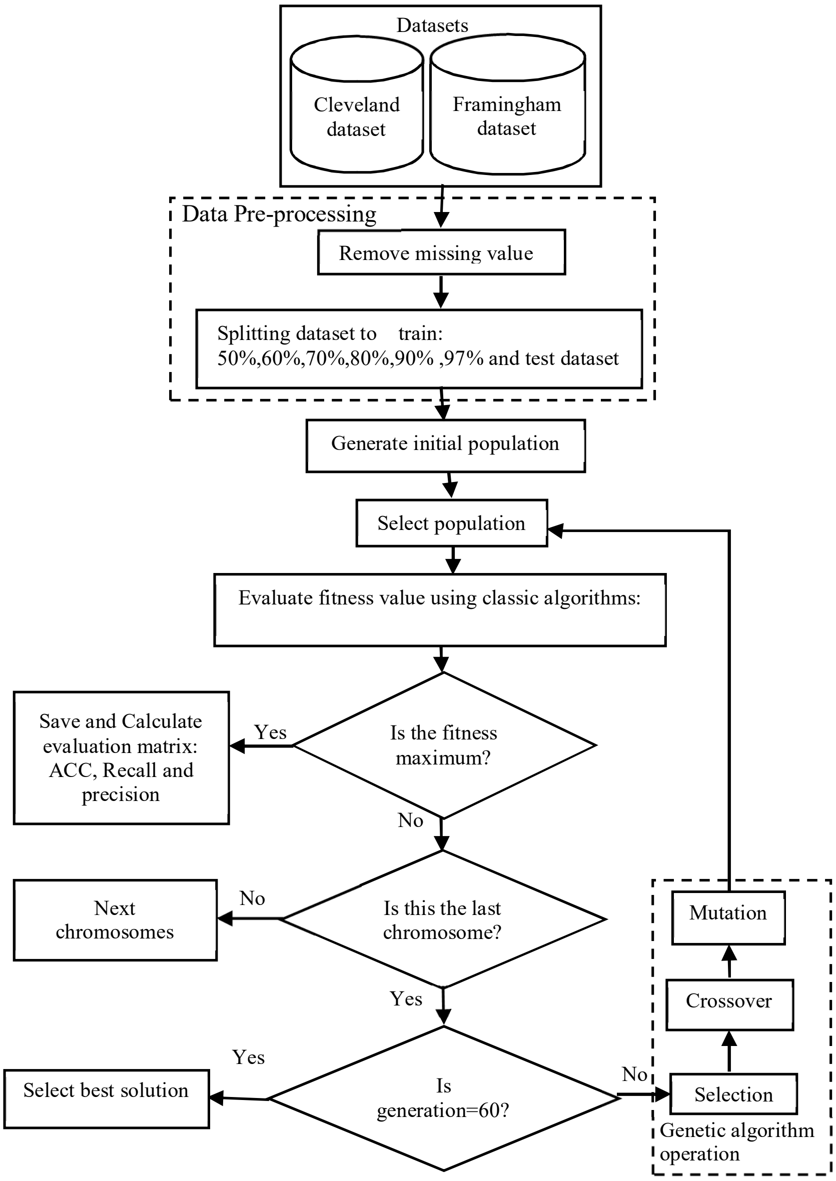

Datasets
Cleveland dataset
Framingham dataset
Data Pre-processing
Remove missing value
Splitting dataset to train: 50%,60%,70%,80%,90% ,97% and test dataset
Generate initial population
Select population
Evaluate fitness value using classic algorithms:
Is the fitness maximum?
Yes
Save and Calculate evaluation matrix: ACC, Recall and precision
No
Is this the last chromosome?
No
Next chromosomes
Yes
Is generation=60?
Yes
Select best solution
No
Selection
Crossover
Mutation
Genetic algorithm operation

Fig. 2 Proposed framework

D. Proposed framework algorithm

The steps of the framework as shown in Fig. 2 and Fig. 3.

Step1: Data preprocessing.

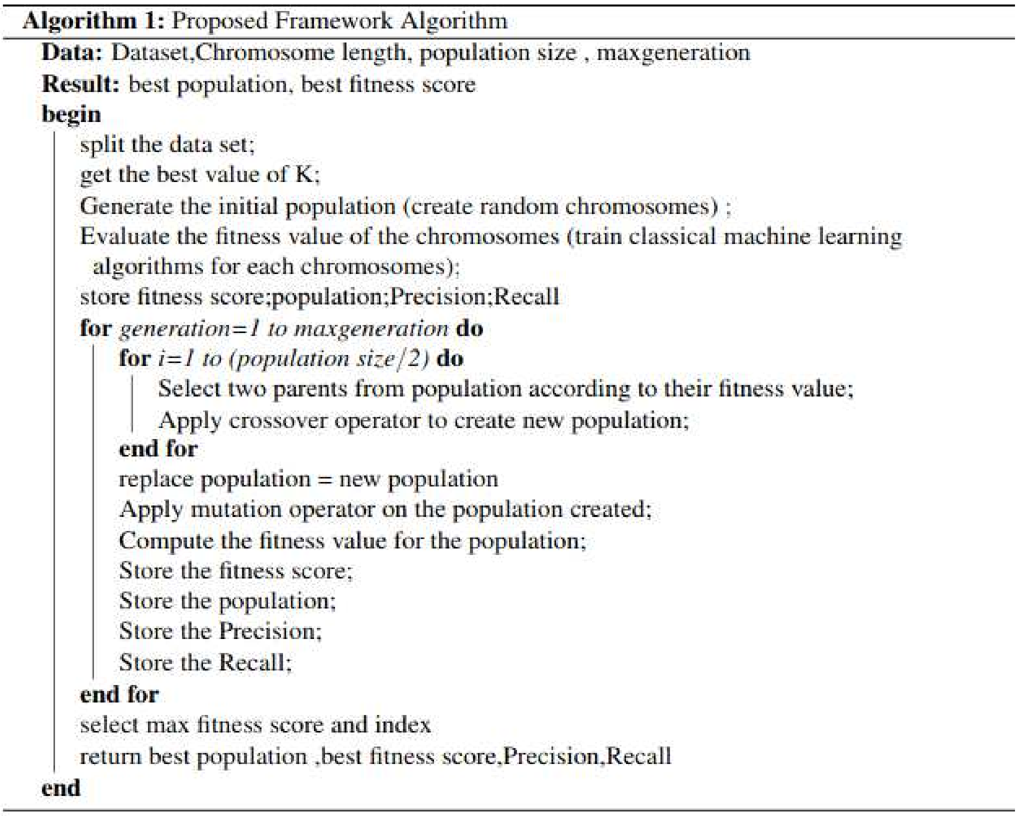
Step2: the chromosome population was initialized randomly. (Each chromosome length is equal to the total number of features), with notice that the first two genes in chromosome sex and age are always true.

Step3: For each chromosome in the initial population, apply the classic algorithms to find the classification accuracy as a fitness of the chromosome and store accuracy, recall, precision, and its population.

Step4: Repeat the steps following till the terminating condition (maximum number of generations) is reached.

- Select two chromosomes according to their fitness.
- Apply uniform crossover operation on the selected individuals.
- Apply complement mutation by selecting any bit randomly and flipping its value.
- Calculate the fitness of the new population were applying the classic algorithms.

Step5: Select the chromosome resulting in the highest classification accuracy of the algorithm classifier to be the solution presented by the system.



```
Algorithm 1: Proposed Framework Algorithm
  Data: Dataset,Chromosome length, population size , maxgeneration
  Result: best population, best fitness score
  begin
      split the data set;
      get the best value of K;
      Generate the initial population (create random chromosomes) ;
      Evaluate the fitness value of the chromosomes (train classical machine learning
        algorithms for each chromosomes);
      store fitness score;population;Precision;Recall
      for generation=1 to maxgeneration do
          for i=1 to (population size/2) do
              Select two parents from population according to their fitness value;
              Apply crossover operator to create new population;
          end for
          replace population = new population
          Apply mutation operator on the population created;
          Compute the fitness value for the population;
          Store the fitness score;
          Store the population;
          Store the Precision;
          Store the Recall;
      end for
      select max fitness score and index
      return best population ,best fitness score,Precision,Recall
  end
```

Fig. 3 Pseudo Code for Proposed Framework

E. Data Preprocessing

In medicinal informatics, if data does not have losing, repeating, and irrelevant data, diagnosing diseases is faster and easier. The two heart disease datasets have been preprocessed by removing noisy and missing data. Each dataset has two subsets: a training set and a testing set. Two sets were created from the entire dataset. Both were to be used, one as a training set and the other as a testing set. The first training set was drawn from half of the dataset, while the testing set was drawn from the other half. The dataset's second training set made up 60% of it, while the testing set made up the remaining 40%. The dataset's third training set made up 70% of it, with the remaining 30% being the testing set. The fourth training set comprised 80% of the dataset, with the remaining 20% being the testing set. The fifth training set utilized 90% of the dataset, and the testing set utilized the remaining 10%. The testing set made up the final 3% of the dataset, whereas the sixth training set made up 97% of it. The testing set was utilized to assess the algorithm's effectiveness and decide its accuracy score. The training set was used to train the algorithm.

The two datasets were loaded using the panda's library of python. This was followed by dividing a dataset into test and training sets. The training set consisted of different cases to get more improvement in accuracy: 50%, 60%, 70%, 80%, 90%, and 97% of the entire dataset, and the rest part constituted the testing set. The algorithms were taught using the practice data set. The testing set was used to evaluate the trained model's output correctness for a variety of input parameters. The five methods applied to the dataset are LR, SVM, DT, RF, and KNN.

This paper proposes applying a genetic algorithm in order to find optimal attributes for each machine learning algorithm, as shown in the pseudo-code of the proposed framework. Figuring out K's ideal value in KNN and the dataset is applying will greatly impact the value of best K. The optimum value of K for KNN is highly data dependent. The open answer is retaining a portion of system performance testing data. Then, using all of the data in the test set, choose k = 1, apply the modeling training, and quantify the accuracy of the prediction. Repeat this step, increasing the k, and get which k is the best for Cleveland trained datasets 70%, 80%, 90%, 97% was 10, 11,13,17 respectively, and the best k for Framingham datasets 70%, 80%, 90%, 97% was 10, 11, 9, 12 respectively.

III. Results and Discussion

The model is created in the Jupyter notebook, which is also used to classify the dataset's cardiac diseases. Here, multiple classification algorithms are used to predict cardiac disease's presence and absence. Two datasets are used in this investigation. The prediction systems are implemented in various evaluation metrics: accuracy, recall, and precision. It should be noted that all the outputs applied to size the execution of the system where the evaluation metrics are calculated by applying the following equations:

$$\text{Accuracy (ACC)} = (TP + TN) / (TP + FN + FP + TN) \quad (1)$$

$$\text{Precision(P)} = TP / (TP + FP) \quad (2)$$

$$\text{Recall(R)} = TP / (TP + FN) \quad (3)$$

where "true positive" (TP)" refers to those who have cardiac disease and were properly diagnosed, False positive

(FP) refers to patients who were incorrectly diagnosed with cardiac disease while not having a condition. True negative (TN) denotes the absence of a cardiac condition in patients and were rightly diag-nosed. False negative (FN) refers to patients with heart illness and fault diagnosis. Accuracy (ACC) evaluated the correctly categorized instances[31]. Recall (R) indicates the percentage of returned fields that were returned correctly and categorized as negative or unrelated to the query. Precision (P) indicates the percentage of returned fields that were got back correctly and categorized as positive as most associated records to the query [12].

This section is dedicated to clarifying the parameters of GA used in this work, but first, performance when using classic algorithms classifier with all the dataset is presented. The results of the classical classifier are summarized in Table II for accuracy, precision and recall of Cleveland and Framingham datasets. Even when 97% of all data is devoted to training, the results are unsatisfactory. Then describes the results obtained by classic algorithms and the proposed system. To make the comparison between classic algorithms and GA-Algorithms in all datasets honest, to be able to see the impact of genetic algorithms.

The proposed system results are recorded in Tables III and IV for Cleveland and Framingham datasets. These results showed that the best performance is (100%) in all algorithms when semi-full training (97% of data as the training set) is used. This draws attention to the fact that the credit for the accuracy achieved by the new system is not due to the use of the large size of the training set. The reason for perfect performance is the coalition of the three factors, their values determined by GA.

TABLE II
RESULT OF CLASSICAL ALGORITHMS OF CLEVELAND AND FRAMINGHAM DATASETS

Datasets	**ML**	**70/30**			**80/20**			**90/10**			**97/3**		
		ACC	**P**	**R**	**ACC**	**P**	**R**	**ACC**	**P**	**R**	**ACC**	**P**	**R**
Cleveland Dataset	KNN	69.23	0.72	0.7	75.40	0.74	0.8	77.41	0.91	0.6	90	1	0.8
	LR	80.21	0.82	0.8	86.88	0.87	0.8	80.64	0.82	0.8	90	0.83	1
	SVM	56.04	0.55	1	54.09	0.53	1	54.83	0.54	1	50	0.5	1
	DT	72.52	0.76	0.7	81.96	0.88	0.7	80.64	0.86	0.6	70	1	0.4
	RF	81.31	0.82	0.8	85.24	0.84	0.8	77.41	0.77	0.8	90	0.83	1
Framingham Dataset	KNN	84.15	0.62	0.02	83.87	0.64	0.07	84.42	0.72	0.12	91.81	1	0.35
	LR	84.60	0.73	0.06	84.42	0.9	0.07	85.24	1	0.12	90.9	1	0.28
	SVM	83.97	0	0	83.33	0	0	83.06	0	0	87.27	0	0
	DT	80.60	0.29	0.15	80.46	0.31	0.14	83.06	0.5	0.2	84.54	0.33	0.21
	RF	84.06	1	0	83.60	1	0.01	83.33	1	0.01	87.27	0	0

TABLE III
RESULT OF MACHINE LEARNING TECHNIQUES COMBINED WITH GENETIC ALGORITHM FOR CLEVELAND AND FOR FRAMINGHAM DATASETS

Datasets	**ML**	**70/30**			**80/20**			**90/10**			**97/3**		
		ACC	**P**	**R**	**ACC**	**P**	**R**	**ACC**	**P**	**R**	**ACC**	**P**	**R**
Cleveland Dataset	KNN	82.41	0.85	0.82	78.68	0.75	0.87	80.64	0.86	0.76	100	1	1
	LR	89.01	0.9	0.9	90.16	0.9	0.9	90.32	0.88	0.94	100	1	1
	SVM	85.71	0.87	0.86	85.24	0.84	0.87	87.09	0.93	0.82	100	1	1
	DT	86.81	0.95	0.8	88.52	0.96	0.81	93.54	1	0.88	100	1	1
	RF	89.01	0.91	0.88	95.08	1	0.9	93.5	1	0.88	100	1	1
Framingham dataset	KNN	84.42	0.85	0.03	84.42	0.7	0.11	84.97	0.76	0.16	91.81	1	0.35
	LR	85.24	1	0.07	84.97	1	0.09	85.51	1	0.14	90.9	1	0.28
	SVM	84.15	1	0.01	83.87	1	0.03	83.87	1	0.04	88.18	1	0.07
	DT	84.6	0.55	0.19	84.01	0.66	0.08	85.79	0.69	0.29	91.81	1	0.35
	RF	84.79	0.9	0.05	84.42	1	0.06	84.69	0.8	0.12	90.90	1	0.28

TABLE IV
OPTIMAL SELECTED ATTRIBUTES FOR CLEVELAND AND FRAMINGHAM DATASET

Datasets	**Machine Learning Techniques**	**No. Of Attribute**	**Selected Attribute**
Cleveland Dataset	KNN	6	(1,2,3,6,7,11)
	LR	9	(1,2,4,6,7,8,10,12,13)
	SVM	6	(1,2,3,6,7,11)
	DT	8	(1,2,5,6,7,8,9,13)
	RF	9	(1,2,3,4,5,6,8,11,13)
Framingham Dataset	KNN	13	(1,2,3,5,6,7,8,10,11,12,13,14,15)
	LR	9	(1,2,5,6,8,9,10,11,15)
	SVM	9	(1,2,3,4,6,7,8,10,11)
	DT	8	(1,2,3,4,5,6,9,14)
	RF	9	(1,2,4,5,7,9,10,14,15)

IV. Conclusion

It is really difficult for even a doctor to determine any heart disease on some raw data, and machine learning techniques are opted by many healthcare sectors to determine it. Different prediction models were generated in this study, and searches were conducted to find the most accurate algorithms for heart disease prediction. Five classifiers—random forest, logistic regression, decision tree, support vector machine, and KNN—were used to predict patients with cardiac illnesses. The performance of the models is tested using Cleveland and Framingham dataset. According to observation, Cleveland is the best dataset, as it has the fewest losing values and offers all 14 attributes as predictors.

The study's other finding is that the performance of the improved prediction models utilizing genetic algorithms is superior to that of standard prediction models, which achieved a classification accuracy of 100% for Cleveland and 91.8% for Framingham datasets. Having attributes with both discrete and continuous values for multivariate data analysis. The introduced genetic model's main goal is to increase the heart disease prediction model's accuracy and eliminate any patient misdiagnosis. However, there is still an area for enhancement.

In the future, identifying and including more features will be extended in the research, and it will use more classification techniques like deep learning etc. Future convolutional neural network-based wearable devices will be accessible on the market, and the proposed work will access trained and tested datasets.

References

[1] M. Swathy and K. Saruladha, "A comparative study of classification and prediction of Cardio-Vascular Diseases (CVD) using Machine Learning and Deep Learning techniques," *ICT Express*, vol. 8, no. 1, pp. 109–116, 2022.

[2] "Cardiovascular diseases (CVDs)." https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) (accessed Aug. 30, 2021).

[3] K. V. V. Reddy, I. Elamvazuthi, A. A. Aziz, S. Paramasivam, H. N. Chua, and S. Pranavanand, "Heart Disease Risk Prediction Using Machine Learning Classifiers with Attribute Evaluators," *Appl. Sci.*, vol. 11, no. 18, p. 8352, 2021.

[4] W. Zunaidi, R. R. Saedudin, Z. A. Shah, S. Kasim, C. Sen Seah, and M. Abdurohman, "Performances analysis of heart disease dataset using different data mining classifications," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 8, no. 6, pp. 2677–2682, 2018.

[5] "Heart-healthy diet: 8 steps to prevent heart disease - Mayo Clinic." https://www.mayoclinic.org/diseases-conditions/heart-disease/in-depth/heart-healthy-diet/art-20047702 (accessed Dec. 27, 2021).

[6] U. N. Dulhare, "Prediction system for heart disease using Naive Bayes and particle swarm optimization," *Biomed. Res.*, vol. 29, no. 12, pp. 2646–2649, 2018.

[7] H. Ayatollahi, L. Gholamhosseini, and M. Salehi, "Predicting coronary artery disease: A comparison between two data mining algorithms," *BMC Public Health*, vol. 19, no. 1, pp. 1–9, 2019, doi: 10.1186/s12889-019-6721-5.

[8] A. Lakshmanarao, Y. Swathi, and P. S. S. Sundareswar, "Machine learning techniques for heart disease prediction," *Int. J. Sci. Technol. Res.*, vol. 8, no. 11, 2019.

[9] F. S. Alotaibi, "Implementation of machine learning model to predict heart failure disease," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 6, pp. 261–268, 2019, doi: 10.14569/ijacsa.2019.0100637.

[10] I. Javid, A. K. Z. Alsaedi, and R. Ghazali, "Enhanced accuracy of heart disease prediction using machine learning and recurrent neural networks ensemble majority voting method," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 3, pp. 540–551, 2020, doi: 10.14569/ijacsa.2020.0110369.

[11] S. Mukherjee and A. Sharma, "Intelligent heart disease prediction using neural network," *Int. J. Recent Technol. Eng.*, vol. 7, no. 5, pp. 402–405, 2019.

[12] A. K. Gárate-Escamila, A. Hajjam El Hassani, and E. Andrès, "Classification models for heart disease prediction using feature selection and PCA," *Informatics Med. Unlocked*, vol. 19, 2020, doi: 10.1016/j.imu.2020.100330.

[13] S. Ware, S. Rakesh, and B. Choudhary, "Heart Attack Prediction by using Machine Learning Techniques," *Int. J. Recent Technol. Eng.*, vol. 8, no. 5, pp. 1577–1580, 2020, doi: 10.35940/ijrte.d9439.018520.

[14] R. Katarya and S. K. Meena, "Machine Learning Techniques for Heart Disease Prediction: A Comparative Study and Analysis," *Health Technol. (Berl).*, vol. 11, no. 1, pp. 87–97, 2021, doi: 10.1007/s12553-020-00505-7.

[15] H. Jindal, S. Agrawal, R. Khera, R. Jain, and P. Nagrath, "Heart disease prediction using machine learning algorithms," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1022, no. 1, 2021, doi: 10.1088/1757-899X/1022/1/012072.

[16] A. Mehmood *et al.*, "Prediction of Heart Disease Using Deep Convolutional Neural Networks," *Arab. J. Sci. Eng.*, vol. 46, no. 4, pp. 3409–3422, 2021, doi: 10.1007/s13369-020-05105-1.

[17] S. Habib, M. B. Moin, S. Aziz, K. Banik, and H. Arif, "Heart failure risk prediction and medicine recommendation using exploratory data analysis," in *2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)*, 2019, pp. 1–6.

[18] N. M. Lutimath, C. Chethan, and B. S. Pol, "Prediction of heart disease using machine learning," *Int. J. Recent Technol. Eng*, vol. 8, pp. 474–477, 2019.

[19] V. Shorewala, "Early detection of coronary heart disease using ensemble techniques," *Informatics Med. Unlocked*, vol. 26, p. 100655, 2021.

[20] C. B. C. Latha and S. C. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," *Informatics Med. Unlocked*, vol. 16, p. 100203, 2019.

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

[22] Z. Rustam, D. A. Utami, R. Hidayat, J. Pandelaki, and W. A. Nugroho, "Hybrid preprocessing method for support vector machine for classification of imbalanced cerebral infarction datasets," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 9, no. 2, pp. 685–691, 2019.

[23] S. Krishnan and S. Geetha, "Prediction of Heart Disease Using Machine Learning Algorithms.," in *2019 1st international conference on innovations in information and communication technology (ICIICT)*, 2019, pp. 1–5.

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

[25] M. D. A. Hossen *et al.*, "Supervised Machine Learning-Based Cardiovascular Disease Analysis and Prediction," *Math. Probl. Eng.*, vol. 2021, p. 1792201, 2021, doi: 10.1155/2021/1792201.

[26] A. K. Shukla, P. Singh, and M. Vardhan, "A new hybrid feature subset selection framework based on binary genetic algorithm and information theory," *Int. J. Comput. Intell. Appl.*, vol. 18, no. 03, p. 1950020, 2019.

[27] N. K. Ayoob, "Improving The Accuracy Of KNN Classifier For Heart Attack Using Genetic Algorithm," *J. kerbala Univ.*, no. المؤتمر العلمي
2016 ,الرابع لكلية العلوم.

[28] "Heart Disease UCI | Kaggle." https://www.kaggle.com/ronitf/heart-disease-uci?select=heart.csv (accessed Aug. 30, 2021).

[29] C. Bou Rjeily, G. Badr, A. Hajjarm El Hassani, and E. Andres, "Medical data mining for heart diseases and the future of sequential mining in medical field," in *Machine Learning Paradigms*, Springer, 2019, pp. 71–99.

[30] "Framingham Dataset | Kaggle." https://www.kaggle.com/ajayvarma91/framingham-dataset (accessed Aug. 30, 2021).

[31] T. Nagamani, S. Logeswari, and B. Gomathy, "Heart disease prediction using data mining with mapreduce algorithm," *Int. J. Innov. Technol. Explor. Eng. ISSN*, pp. 2278–3075, 2019.