Multimedia Tools and Applications (2022) 81:18155–18179
https://doi.org/10.1007/s11042-022-12425-x

A hybrid GA and PSO optimized approach for heart-disease prediction based on random forest

Mohamed G. El-Shafiey [1] · Ahmed Hagag [2] · El-Sayed A. El-Dahshan [1,3] · Manal A. Ismail [4]

Received: 10 February 2021 / Revised: 10 June 2021 / Accepted: 25 January 2022 /
Published online: 8 March 2022

Abstract

Nowadays, heart diseases are significantly contributing to deaths all over the world. Thus, heart-disease prediction has garnered considerable attention in the medical domain globally. Accordingly, machine-learning algorithms for the early prediction of heart diseases were developed in several studies to help physicians design medical procedures. In this study, a hybrid genetic algorithm (GA) and particle swarm optimization (PSO) optimized approach based on random forest (RF), called GAPSO-RF, is developed and used to select the optimal features that can increase the accuracy of heart-disease prediction. The proposed GAPSO-RF implements multivariate statistical analysis in the first step to select the most significant features used in the initial population. After that, a discriminate mutation strategy is implemented in GA. GAPSO-RF combines a modified GA for global search and a PSO for local search. Moreover, PSO achieved the concept of rehabbing individuals that had been refused in the selection process. The performance of the proposed GAPSO-RF approach is validated via evaluation metrics, namely, accuracy, specificity, sensitivity, and area under the receiver

✉ Ahmed Hagag
ahagag88@gmail.com

Mohamed G. El-Shafiey
mismailaly@eelu.edu.eg

El-Sayed A. El-Dahshan
e_eldahshan@yahoo.com

Manal A. Ismail
manal_shoman@yahoo.com

[1] Faculty of Computers and Information Technology, Egyptian E-Learning University, Dokki, Giza 12611, Egypt

[2] Department of Scientific Computing, Faculty of Computers and Artificial Intelligence, Benha University, Benha 13518, Egypt

[3] Department of Physics, Faculty of Science, Ain Shams University, Abbasia, Cairo 11566, Egypt

[4] Faculty of Engineering, Helwan University, Helwan, Cairo 11731, Egypt

operating characteristic (ROC) curve by using two datasets from the University of California, namely, Cleveland and Statlog. The experimental results confirm that the GAPSO-RF approach attained the high heart-disease-prediction accuracies of 95.6% and 91.4% on the Cleveland and Statlog datasets, respectively. Furthermore, the proposed approach outperformed other state-of-the-art prediction methods.

Keywords Cleveland dataset · Feature selection (FS) · Genetic algorithm (GA) · Particle swarm optimization (PSO) · Heart-disease prediction · Random forest (RF) · Statlog dataset

1 Introduction

The heart pumps blood to the entire human body. Coronary arteries are the blood-vessels that transport oxygenated blood to the heart [36]. The shrinking of coronary arteries is the primary cause of heart failure (HF). Heart diseases are one of the major reasons of human mortality, as reported by the World Health Organization. In 2013, heart diseases caused the highest number of deaths globally, at approximately 17.3 million. Similarly, in 2016, approximately 17.6 million deaths were attributed to heart diseases, amounting to a rise of 14.5% from 2006 [10]. Moreover, patients with HF suffer from other symptoms as well, including difficulty in breathing, weakness, and swollen feet [14]. Heart diseases may be managed or controlled if trained medical professionals detect them at their early stages, thereby enabling them to make the correct decision. Therefore, early detection of heart diseases is critical to improving HF symptoms and extending the lives of patients [8]. The medical history of a patient includes a substantial number of features. However, not all these features may be equally significant, and some may even be redundant. Additionally, using all the features at once deteriorates the performance of diagnosis. Most research-based on heart-disease prediction focused on two factors: selecting the best features while dismissing the irrelevant ones and choosing an appropriate classifier. Therefore, the prediction methods are aimed at selecting the optimal features and an appropriate classifier. Recently, machine-learning-based methods have improved the quality of our lives, especially in the medical domain [2, 5, 7, 20, 46, 48, 49, 53].

Many research papers have used machine learning to diagnose heart disease and predict whether a patient has heart disease [1, 25, 28–30, 51]. Recently, Amin et al. [5] presented a hybrid technique that comprised Naïve Bayes (NB), logistic regression, and feature selection (FS). Revett et al. [44] deployed the use of rough sets to determine the information content of each subset of the feature space. Furthermore, support vector machines (SVMs) were applied in some researches including [48, 49]. Saqlain et al. [46] implemented Fisher score for FS and SVM for classification. Saifudin et al. [45] applied bagging based on random forest (RF) to improve classification accuracy of heart disease. Subsequently, Gupta et al. [19] used Yule-Walker (YW) and Principal Component Analysis (PCA) for R-peak Detection in Electrocardiogram (ECG) signal, during the detection process regular and abnormal signals were considered. The results obtained using PCA with YW carried out the results using PCA without YW. Besides, FS was also implemented in other domains [27] to increase the classification accuracy by presenting a multi-layer hybrid technique to detect peer to peer botnets. A decision tree algorithm is applied for feature selection to extract the most relevant features and ignore the irrelevant features. They achieved high accuracy by using a decision tree algorithm and their experiments prove the benefits of using multi-layer instead of single layer. In addition, Reddy et al. [41] proposed an approach for diabetes diagnosis, the authors

used Locality Preserving Projection (LPP) algorithm for feature reduction and Firefly-BAT (FFBAT) optimization algorithm with artificial neural network (ANN) for diabetes disease classification. The results have proved that the proposed classification framework outperforms the existing method by achieving better accuracy. In conclusion, FS is the most crucial step in increasing the accuracy of heart-disease diagnosis. For example, a doctor might decide regarding a patient who suffers from HF based on classification implemented using the selected features. The previous researches gave more attention to improving and developing classification methods than selecting the best features. In addition, it needs to improve the accuracy rate.

The objectives of this work are: 1) Select the best features, 2) Improve the heard disease prediction accuracy, and 3) Improve the complexity time. Therefore, we introduce an efficient, hybrid genetic algorithm (GA) and particle swarm optimization (PSO) approach based on random forest (RF) for optimizing the FS process to select the crucial features that increase the accuracy of heart-disease diagnosis. The main contribution of this paper is to develop a hybrid approach, called GAPSO-RF, for heart-disease prediction. First, a discriminate mutation strategy based statistical analysis is applied to be used in the adaptive mutation operator for GA. Second, a modified genetic algorithm with PSO supported by the RF algorithm is used to select the best features. PSO is used to target the rejected individuals of each generation to fulfill the concept of rehabbing the rejected individuals, maximizing the utilization of all individuals in each generation. Finally, the proposed GAPSO-RF is validated via evaluation metrics, namely, accuracy, specificity, and area under the receiver operating characteristic (ROC) curve by using two heart-disease datasets from the University of California (UCI), Irvine, machine learning repository [13], namely, Cleveland and Statlog. Experimental findings suggest that the proposed GAPSO-RF achieves high prediction accuracies.

The rest of this paper is structured as follows. Section 2 illustrates the related work. The materials and proposed approach are discussed in Section 3, including the description of both the datasets, background concepts related to FS, and classification process. The experimental results are provided in Section 4, including a comparative analysis of our method with those in the literature. Finally, the conclusions are drawn in Section 5.

2 Related works

Recent researches have been focused on FS, prediction, and increasing the heart-disease-prediction accuracy. This section overviews the recently published related researches. Lately, Mohammad S. Amin et al. [5] developed a heart-disease-prediction model by using the identified best features and data-mining algorithms on the Cleveland dataset. Subsequently, Saqlain et al. [46] employed Fisher score and the Matthews correlation coefficient as an FS algorithm and SVM for binary classification to diagnose heart diseases on several datasets. Purnomo et al. [39] applied FS in the form of backward elimination on NB to increase classification accuracy on heart-disease from 84.29% to 89.45%. Besides, a fuzzy algorithm was used as another solution by Vivekanandan and Iyengar [51]. Priyatharshini and Chitrakala [38] developed a self-learning fuzzy rule-based system to predict heart disease, the authors achieved an overall accuracy 90.7%. Subsequently, Halder er al. [21] implemented computerized diagnosis system using Rough set classifier from multi-lead ECG signal for the classification of myocardial infarction (MI) disease. Dwivedi [15] applied different algorithms, namely, ANN, SVM, logistic regression, k-nearest neighbors (KNN), classification tree, NB,

Table 1 Summary of the related work

Study	Proposed	Techniques & Tools	Finding	Limitation
Mohammad S. Amin et al. [5]	Developed a heart-disease-prediction model by using the identified best features and data-mining algorithms.	Vote algorithm as a hybrid methodology of logistic regression and NB. (RapidMiner)	They achieved high accuracy in heart-disease prediction on the Cleveland dataset.	The feature selection algorithm is not use.
Saqlain et al. [46]	A cardiac disease diagnosis system is presented by proposing the feature selection algorithms.	Fisher score & Matthews correlation coefficient as an FS algorithm and SVM for binary classification.	The proposed technique achieved considerably better prediction results than the other comparative techniques.	The feature selection algorithm needs to improve.
Purnomo et al. [39]	Predicting heart-disease using different FS approached and NB.	NB + (Backward Elimination / Optimize Selection / Forward Selection).	(NB+forward selection) outperforms other feature selection approach.	The classifier algorithm needs to improve.
Vivekanandan and Iyengar [51]	Optimal feature selection using a modified differential evolution (DE) algorithm and its effectiveness for prediction of heart disease.	Modified differential evolution algorithm and integrated model of fuzzy analytic hierarchy process (AHP) & ANN.	The results show that the modified DE outperforms the best proven traditional DE for feature selection.	The classification accuracy needs to increase.
Priyatharshini and Chitrakala [38]	Build an efficient mining framework for coronary disease diagnosis.	A self-learning approach for a fuzzy rule-based system. (MATLAB)	The experimental results show that the self-learning approach has the efficacy to categorize the risk levels of individuals with suspected coronary disease.	The datasets are limited to validate the experiments.
Halder er al. [21]	A computerized diagnosis system implemented for the classification of myocardial infarction (MI) disease.	Rough set classifier from multi-lead ECG signal.	The proposed method not only finds the suitable rules to explore better knowledge but also the important factors affecting the decision making of MI by using rough set explorer system (RSES).	
Long et al. [30]	A heart-disease-diagnosis system.	A chaos-based firefly algorithm, rough set-based attribute reduction, and an interval type-2 fuzzy logic system for classification. (MATLAB+WEKA)	The proposed combination rough sets-based attribute reduction with interval type-2 fuzzy logic system overcomes others in terms of accuracy, convergence speed and processing time.	The authors need to utilize another hybrid feature selection algorithms and other classifiers to increase accuracy.
Dwivedi [15]	Different machine learning techniques was evaluated for prediction of heart-disease.	ANN, SVM, NB, logistic regression, KNN, and classification trees.	The analysis of heart disease prediction using different machine algorithms.	The feature selection algorithm is not use.
Krishnaiah et al. [29]	Heart-disease prediction system using data mining technique.	An exponential membership function with standard deviation & incorporating fuzzy set methods into the classical KNN decision rule. (WEKA)	The result shows the capable in nature to remove the redundancy of the data and the better accuracy of the system & the performance analysis shows that the fuzzy K-NN classifier	The experiment environment is not mentioned. And the classification accuracy needs to increase.

Table 1 (continued)

Study	Proposed	Techniques & Tools	Finding	Limitation
			is more accurate as compared with K-NN classifier.	
Buettner and Schunter [11]	Efficient machine learning based detection of heart-disease.	RF	RF outperforms other machine learning techniques using the same database.	The feature selection algorithm is not used.
Suresh and Ananda Raj [50]	An optimization approach using genetic algorithm for heart-disease prediction.	RF, NB, Decision Tree, SVM and GA. (WEKA)	NB classifier performance achieved higher accuracy compared with other three classification methods in different datasets.	The feature selection algorithm needs to improve.
Paul et al. [35]	A system for the diagnosis of heart-disease.	Correlation coefficient & GA+fuzzy rules in classification. (C++)	A better performance is obtained by the proposed approach when comparing with other existing methods.	The experiment environment is not mentioned. And the classification accuracy needs to increase.
Ismaeel et al. [22]	Proposed Extreme learning machine (ELM) technique for heart-disease diagnosis.	Three-layer neural network with sigmoid activation function.	The proposed ELM model improves the performance when compared with other models.	The experiment environment is not mentioned. Moreover, the feature selection algorithm needs to improve.
El-Bialy et al. [16]	Feature analysis of coronary artery heart disease.	Fast decision tree & C4.5 pruning tree algorithms. (WEKA)	The results show that the classification accuracy of the collected dataset is higher than the average of the classification accuracy.	The feature selection algorithm needs to improve.
Chitra and Seenivasag-am [12]	Developed an intelligent heart-disease-prediction system.	A feed-forward neural network (FFNN) & cascade-correlation neural network (CCNN).	The performance analysis proved that the time complexity is less in CCNN, whereas the design complexity is less in FFNN.	The experiment environment is not mentioned. Furthermore, the feature selection algorithm needs to improve.
Saxena et al. [47]	Efficient heart-disease prediction system using data mining.	Decision-trees and rules generated. (JAVA)	The proposed method improves the performance when compared with other methods.	
Luo and Wu [31]	Heart rate prediction model based on neural network.	Long short-term memory (LSTM) neu-ral network.	The result shows that ADAM optimizer is significantly better than SGD.	
Reddy et al. [42]	Efficient system for heart-disease prediction.	FS using locality preserving projection + a hybrid oppositional firefly with BAT and rule-based fuzzy logic. (MATLAB)	The proposed method minimized the complexity and improved the accuracy using locality preserving projection.	The feature selection algorithm needs to improve. Moreover, the classification accuracy still needs to increase.
Reddy et al. [43]	Implemented adaptive genetic algorithm with fuzzy logic to predict heart disease	Fuzzy classifiers.	The results outperform the other methods.	The authors did not mention the parameters for genetic algorithms. Moreover, the classification accuracy still needs to increase.

and achieved the highest accuracy in logistic regression. Recently, Krishnaiah et al. [29] proposed a fuzzy KNN approach by presenting an exponential membership function with standard deviation, and they calculated the mean of the attributes measured. Buettner and Schunter [11] performed classification using the RF algorithm, which they validated on the Cleveland dataset. Notably, their method was not an FS approach. However, several studies, including [35, 50], used GAs for performing FS. Ismaeel et al. [22] proposed an improved extreme learning machine algorithm and implemented it on the Cleveland dataset; their algorithm performed better than back-propagation neural networks. El-Bialy et al. [16] performed FS using fast decision tree and C4.5 pruning tree algorithms. Saxena et al. [47] used decision-trees for rule generation. Reddy et al. [43] implemented an adaptive genetic algorithm with fuzzy logic to predict heart disease based on a rough set for features selection. However, the previous studies on heart-disease prediction still lack optimizing FS and using an appropriate classifier to enhance the performance of the heart-disease classification. Table 1 provides a summary of the related methods included in this study.

However, the previous studies on heart-disease prediction still lack optimizing FS and using an appropriate classifier to enhance the performance of the heart-disease classification. Although several studies proposed different FS algorithms, they did not considerably focus on GA. While GA is best known for searching and finding the best subset features from the original features that enhance the classification. In addition, FS is the most important factor in improving the accuracy of heart-disease diagnosis to help doctors make the correct decision. Therefore, we aim to select the best features by using the GAPSO-RF approach. Which selects the best features using GA and PSO and at the same time enhance the performance of the heart-disease classification with RF algorithm.

3 Materials and the proposed approach

We aim to select the best features to increase the heart-disease-diagnosis accuracy. Thus, an GAPSO-RF based FS approach is proposed. Before the FS process, we implement discriminate mutation strategy based statistical analysis to be used in adaptive mutation operator in GA. After that, the features ranges are normalized by applying *min–max* normalization. During the FS process, the proposed GAPSO-RF utilizes GA to search for a set of optimal features by optimizing the hyper-parameters of the GA and the modified selection operator. The rejected individuals from selection are passed to the PSO for reformation. The population of PSO will be made of these rejected individuals who will connect to update their position and velocity to achieve the best possible result from the non-fit individuals. The best individuals of PSO will be injected into the new population of GA. The fitness function in both GA and PSO is optimized using an optimized RF classifier to increase the classification accuracy. The overall workflow of the proposed approach is illustrated in Fig. 1. Four tasks have to be performed for prediction: (1) statistical analysis and data pre-processing, (2) GAPSO-RF utilization, (3) RF-based classification, and (4) performance measurement. In the following subsections, we describe the datasets, and then each step of the proposed approach is discussed.

3.1 Datasets description

In the proposed approach, two datasets, namely, Cleveland and Statlog, from the UCI machine-learning repository are used [13]. Table 2 lists the features of both the datasets.

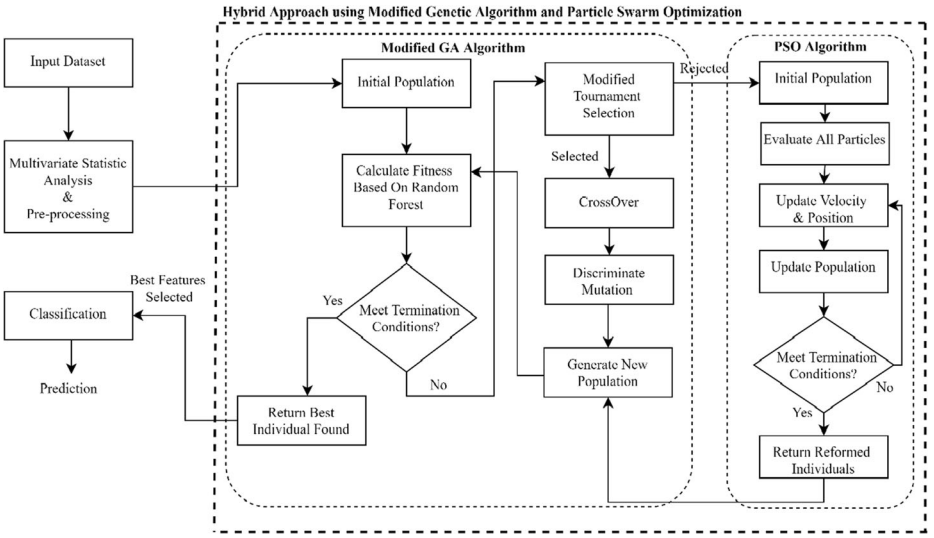


Fig. 1 Workflow of the proposed GAPSO-RF approach

Table 2 Cleveland and Statlog datasets contain 14 features each

Feature	Feature Type	Feature Description
Age	Numeric	Patient age in years
Sex	Nominal	Patient gender (0 means female, and 1 means male)
Cp (suffering in chest)	Nominal	comprises the following values: 1. Angina pectoris 2. atypical angina 3. non-anginal pain 4. no symptoms
Trestbps (Resting BP)	Numeric	Resting blood pressure (in mm/Hg on hospital entry)
Chol	Numeric	Cholesterol in mg/dl
Fbs	Nominal	blood glucose when the patient is fasting >120 mg/dl; 1 when true, and 0 when false
Restecg	Nominal	Rest electrocardiographic results assume one of the following three values: 0. normal 1. has ST-T wave anomalies (T-wave inversions and/or ST segment elevation or depression of >0.05 mV) 2. it displays potential or particular left ventricular hypertrophy according to ESTES standard
Thalach	Numeric	Reached maximum cardiac rate
Exang	Numeric	Workout led to angina (1 means yes, and 0 means no)
Oldpeak	Numeric	ST depression due to exercise relative to rest
Slope	Nominal	Slope of the ST segment during maximum workout • upsloping • flat • downsloping
Ca	Nominal	Key vessel number colored by fluoroscopy
Thal	Nominal	The heart status is defined by the following three values: • 3 means no defect • 6 means an irreversible defect • 7 means a reversible defect
Num	Nominal	It outlines two values for cardiac diagnosis: 0 means healthy (the patient has no heart disease), and 1 means unhealthy (the patient suffers from a heart disease)

The (*Num*) variable represents two values of the heart-disease diagnosis: 0 means healthy (the patient has no heart disease), and 1 means unhealthy (the patient has a heart disease). As shown in Fig. 2, in the Cleveland dataset, 165 records have the value of (1), and 138 have the value of (0). In addition, in the Statlog dataset, 120 records have the value of (1), and 150 have the value of (0).

3.2 Statistical analysis and pre-processing

3.2.1 Multivariate statistic analysis

The first step is to analyze the conditional mean and variance for each attribute conditioning on '*Num* = 1' and '*Num* = 0', and calculate the T^2 metric for each attribute, where T^2 metric is defined as follows:

$$T^2 = \left[\overline{X}_1 - \overline{X}_0\right]^2 \left[\left(\frac{1}{n_1} + \frac{1}{n_0}\right) S_p\right]^{-1} \tag{1}$$

where $\overline{X}_1$ and $\overline{X}_0$ are the mean for *Num* equals 1 and 0, respectively, and S_p is defined as follows:

$$S_p = \left(\frac{n_1 - 1}{n_1 + n_0 - 2}\right) S_1 + \left(\frac{n_0 - 1}{n_1 + n_0 - 2}\right) S_0 \tag{2}$$

where n_1 and n_0 are the sample numbers when *Num* equals 1 and 0, respectively. In addition, S_1 and S_0 are the standard deviations for *Num* equals 1 and 0, respectively. Tables 3 and 4 report the statistical analysis for all the attributes in the selected datasets when *Num* equals 1 and 0, respectively. Moreover, the T^2 metric is calculated in Tables 5 and 6 for Cleveland and Statlog datasets, respectively.

As depicted in Table 5, the attributes (*Age, Trestbps, Chol, Thalach,* and *Oldpeak*) of Cleveland dataset have the most considerable variance. It is because these attributes are continues, and a considerable variance is expected for continuous attributes. The attributes

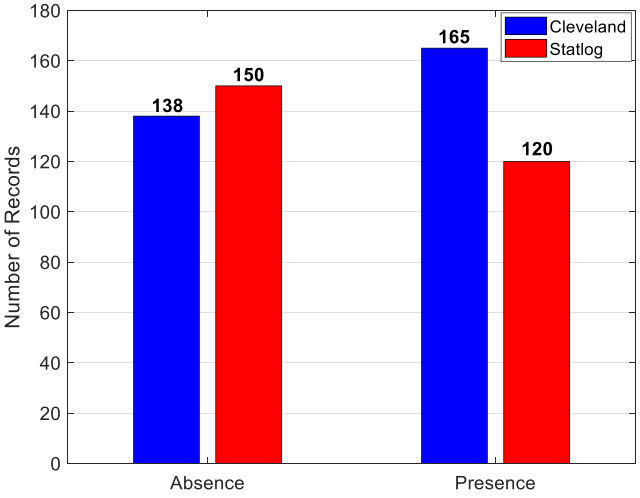


Fig. 2 Distributions for the Cleveland and Statlog datasets

Table 3 Statistics of the attributes in the Cleveland and Statlog dataset when *Num* = 1

Feature	Cleveland		Statlog	
	Mean	Standard Deviation	Mean	Standard Deviation
Age	52.496970	9.550651	56.591667	8.116273
Sex	0.563636	0.497444	0.833333	0.374241
Cp	1.375758	0.952222	3.616667	0.779823
Trestbps	129.303030	16.169613	134.441667	19.095424
Chol	242.230303	53.552872	256.466667	47.969166
Fbs	0.139394	0.347412	0.141667	0.350170
Restecg	0.593939	0.504818	1.225000	0.974140
Thalach	158.466667	19.174276	138.858333	23.130719
Exang	0.139394	0.347412	0.550000	0.499580
Oldpeak	0.583030	0.780683	1.584167	1.282067
Slope	1.593939	0.593635	1.816667	0.564843
Ca	0.363636	0.848894	1.150000	1.034286
Thal	2.121212	0.465752	5.833333	1.769648

(*Sex, Fbs,* and *Restecg*) are non-continuous, and have the slightest variances, so they provide much less information according to the entropy theory. The remaining attributes (*Cp, Exang, Slope, Ca,* and *Thal*) are non-continuous, and have the highest variances, providing the essential information for classification. Further correlation analysis is implemented for (*Cp, Exang, Slope, Ca,* and *Thal*) and shows some correlation between (exang) and the other four attributes as shown in Table 7. After excluding (*Exang*), the remaining four attributes (*Cp, Slope, Ca,* and *Thal*) are the most significant. These four attributes will have a lower mutation probability (e.g., 10^{-3}) throughout GA evolution.

Table 6 presents the T^2 metric for each attribute in the Statlog dataset. The attributes (*Thalach, Chol, Thal, Trestbps*, and *Age*) have the most considerable variance. The attributes (*Sex, Fbs,* and *Restecg*) are non-continuous, and have the slightest variances, so they provide much less information according to the entropy theory. The remaining attributes (*Cp, Exang, Oldpeak, Slope,* and *Ca*) are non-continuous, and have the highest variances, providing the essential information for classification. Further correlation analysis for (*Cp, Exang, Oldpeak,*

Table 4 Statistics of the attributes in the Cleveland and Statlog dataset when *Num* = 0

Feature	Cleveland		Statlog	
	Mean	Standard Deviation	Mean	Standard Deviation
Age	56.601449	7.962082	52.706667	9.509830
Sex	0.826087	0.380416	0.553333	0.498813
Cp	0.478261	0.905920	2.820000	0.927362
Trestbps	134.398551	18.729944	128.866667	16.457660
Chol	251.086957	49.454614	244.213333	54.019085
Fbs	0.159420	0.367401	0.153333	0.361516
Restecg	0.449275	0.541321	0.860000	0.990085
Thalach	139.101449	22.598782	158.333333	19.283357
Exang	0.550725	0.499232	0.153333	0.361516
Oldpeak	1.585507	1.300340	0.622667	0.800851
Slope	1.166667	0.561324	1.400000	0.590757
Ca	1.166667	1.043460	0.286667	0.648557
Thal	2.543478	0.684762	3.786667	1.556914

Table 5 The T^2 metric for all attributes in the Cleveland dataset

Feature	S_P	$\overline{X}_1 - \overline{X}_0$	T^2
Age	8.827614	−4.104480	143.414563
Sex	0.444178	−0.262451	11.653548
Cp	0.931148	0.897497	65.008119
Trestbps	17.334947	−5.095520	112.557643
Chol	51.687551	−8.856653	114.044344
Fbs	0.356510	−0.020026	0.084538
Restecg	0.521432	0.144664	3.016085
Thalach	20.732938	19.365217	1359.265531
Exang	0.416513	−0.411331	30.526314
Oldpeak	1.017205	−1.002477	74.243914
Slope	0.578929	0.427273	23.697672
Ca	0.937450	−0.803030	51.693507
Thal	0.565434	−0.422266	23.697943

Table 6 The T^2 metric for all attributes in the Statlog dataset

Feature	S_P	$\overline{X}_1 - \overline{X}_0$	T^2
Age	8.891049	3.885000	113.171686
Sex	0.443499	0.280000	11.785068
Cp	0.861850	0.796667	49.094218
Trestbps	17.628906	5.575000	117.536598
Chol	51.332741	12.253333	194.994691
Fbs	0.356478	−0.011667	0.025455
Restecg	0.983005	0.365000	9.035219
Thalach	20.991701	−19.475000	1204.525619
Exang	0.422820	0.396667	24.808727
Oldpeak	1.014525	0.961500	60.749741
Slope	0.579250	0.416667	19.981132
Ca	0.819832	0.863333	60.609500
Thal	1.651374	2.046667	169.105398

Slope, and *Ca*) shows some correlation between (*Exang*) and the other four attributes as shown in Table 8. After excluding (*Exang*), the remaining four attributes (*Cp, Oldpeak, Slope,* and *Ca*) are the most significant. These four attributes will have a lower mutation probability (e.g., 10^{-3}) throughout GA evolution.

Table 7 Correlation analysis in the Cleveland dataset

Features	*CP*	*Slope*	*Ca*	*Thal*	*Exang*
Cp	1	0.11971659	−0.18105303	−0.16173557	−0.39428027
Slope	0.11971659	1	−0.08015521	−0.10476379	−0.25774837
Ca	−0.18105303	−0.08015521	1	0.15183213	0.11573938
Thal	−0.16173557	−0.10476379	0.15183213	1	0.20675379
Exang	−0.39428027	−0.25774837	0.11573938	0.20675379	1

Table 8 Correlation analysis in the Statlog dataset

Features	*CP*	*Slope*	*Ca*	*Exang*	*Oldpeak*
Cp	1	0.13689972	0.22588953	0.35315984	0.16724401
Slope	0.13689972	1	0.10949768	0.25590835	0.60971157
Ca	0.22588953	0.10949768	1	0.15334736	0.25500546
Exang	0.35315984	0.25590835	0.15334736	1	0.2746722
Oldpeak	0.16724401	0.60971157	0.25500546	0.2746722	1

3.2.2 Discriminate mutation strategy in genetic algorithm

In the current GA algorithm, the individual dimension is 13 (as there are 13 attributes), so it requires a population size of $2^{13} = 8192$ individuals to cover all possible combinations. An improvement is always selecting the attributes with the most critical information to be endowed with less mutation probability (i.e., 10^{-3}). In comparison, the remaining attributes will have a higher mutation probability to explore more individuals with higher fitness. In the simulation, the sets (*Cp, Slope, Ca,* and *Thal*) and (*Cp, Oldpeak, Slope,* and *Ca*) are the most significant attributes for Celeveland and Statlog, respectively, which have less mutation probability, (i.e., 10^{-3}). The remaining attributes will have higher but equal mutation probabilities. The initialization of the population at the start of the GA should also be modified using this discriminate mutation strategy. In the initial population, the individual should always have the most significant attributes.

3.2.3 Data pre-processing

The data are normalized before performing FS. To that end, we normalize the dataset values. This process has been gaining importance because all features may have different data types, and it eliminates the numerical difficulties due to the different range of values during the computation process. In the proposed approach, we implemented *min–max* normalization, a technique that converts a value a to a in the range of $[max\_new - min\_new]$ as follows:

$$a = \frac{a - a_{min}}{a_{max} - a_{min}} \times [max\_new - min\_new] + min\_new; \tag{3}$$

where from min_new to max_new denotes the range of the transformed values. We implemented $min\_new = 0$ and $max\_new = 1$. Subsequently, these transformed values were used as input for the FS method.

3.3 Hybrid modified genetic algorithm and particle swarm optimization

FS, which is a method of selecting reduced number of appropriate features, enhances the classification by determining the best subset of features from the set of original features. It eliminates unnecessary features, thereby lowering the computational and memory costs. It involves selecting a subset of features t from the total features T based on a particular optimization criterion. GA combined with PSO was used as an FS approach to search for

optimal solutions. GA was introduced by John Holland in 1975. Although they can be used for solving both search and optimization problems, they are best known for solving the latter (Holland, 1992). A GA is a search heuristic that mimics the natural evolution process. It is regularly used to produce useful solutions for the problems related to search and optimization. Notably, GA belongs to the broader category of evolutionary algorithms (EAs). In several real-world optimization projects, EAs have proved to be the most efficient solution. Substantially, Holland supposed that the population size is limitless, that the fitness function correctly represents the convenience of a solution, and that the correlations between the genes are very small, which leads to problems [9]. Population size is limited, impacting the GA's sampling capacity and efficiency.

PSO were introduced by Eberhart and Kennedy [26] in 1995. It is a population-based optimization technique that is inspired by the behavior of fish schooling or bird flocking. PSO algorithm is one of several types of swarm intelligence algorithms. One of several PSO's main advantages is that it is computationally inexpensive due to its low system requirements [37]. Using a local search approach in combination with GA solve much of the hindrances that occur due to the finite population size. Hybridization has proven to be an efficient way to construct capable genetic algorithms. By adding new genes, a local search approach with GA helps neutralize much of the challenges that exist because of the limited population size as well as the genetic drift dilemma [6]. A GA uses the laws of genetics as its paradigm for implementing problem-solving on a population (P) of individuals. Each individual is characterized by a set of variables called genes. To build a chromosome, genes are combined into a string. Therefore, each solution is represented by a chromosome.

Chromosomes C_k, where $k = (1, \ldots, P)$, are encoded in a binary vector B_k of length n. Binary encoding is used to identify whether a feature is selected for input or not. The group of all the chromosomes is referred to as population. In the initial population, the individual should always have the four most significant attributes, (*Cp, Slope, Ca,* and *Thal*). After that, a GA accomplishes its task via four basic operations: modified selection, crossover, modified mutation, and fitness calculation. We believe that (non-fit individuals) can contain good genes that can direct the search process' cursor to locations in the search space where significant improvements can be found. Accordingly, in the selection operation, the rejected chromosomes (non-fit individuals) are passed to PSO for reformation as GA searches for good chromosomes, not good genes, and the best individuals (fittest individuals) are selected based on the value of the fitness function, which is calculated in both GA and PSO using the RF algorithm and with high efficiency to survive to next generation. Moreover, RF prevents over-fitting, which is one of the main challenges in heart-disease prediction. Therefore, in the proposed GAPSO-RF, the RF classifier is used with GA to select the best features. Algorithm 1 presents the hybrid approach using modified genetic algorithm and PSO. RFs comprise many individual decision-trees, which function as an ensemble. An algorithm that can construct many small decision-trees using a few features is considered computationally cheap. If we can create several small, weak decision-trees in parallel, then by averaging or taking the majority vote, we can combine the trees to form a single, strong learner. Practically, RFs are the most effective learning algorithms to date. The RF algorithm is illustrated in Algorithm 2. In the following subsections, we detail the GAPSO-RF processes, namely, selection, crossover, and mutation.

ALGORITHM 1: HYBRID APPROACH USING MODIFIED GENETIC ALGORITHM AND PSO

```
Input:  set parameters, produce P to random population solutions (individuals) m denotes the maximum
        number of generations; encode each individual in the binary vector
Output: the best individual determined P(n)
begin
  foreach individual j to P do
    calculate fitness (j);  /* using the RF algorithm */
    while iteration number < m
      /* Modified tournament selection*/
      Selected = SelectBest (j);
      If Selected then
          If Crossover then
            randomly choose two parents j_a and j_b;
            produce offspring j_c = crossover (j_a and j_b);
          else
            /* modified mutation based discriminate mutation strategy*/
            randomly choose one individual j;
            produce offspring j_c = mutate (j);
          end
          compute the evaluation fitness of individual j_c;
          replace the least fit individual by j_c;
      else
        PSOOutput = PSOReformation (rejected j)
      end
       If PSOOutput then
          new generation + PSOOutput
       else
         new generation via tournament selection;
       end
    end
  end
end
```

ALGORITHM 2: RF ALGORITHM

```
Input:  training set comprises s samples and p variables, and n denotes the number of nodes in the tree
Output: ensemble of trees, T_r(1^R)
begin
  foreach r to R do
    Build a sample from the original training set D with the replacement of size n;
    Feed the bootstrap r to a learning decision-tree T_r;
    foreach n to minimum node size do
      Feed the bootstrapped data to an RF tree T_r;
      randomly choose p' variables from p variables;
      select the best variables split among these p' variables;
      split a node into two child nodes;
    end
  end
  return an ensemble T_r(1^R);
  Classification: to determine which C_r(x) is the prediction class of the rth RF tree;
  Therefore, C_rf^B(x) majority vote {C_b(x)}_1^B;
end
```

3.3.1 Modified tournament-selection operator

In a GA cycle, the initial population is set to be 50 and the maximum iteration number to be 30 generations. Subsequently, we begin the tournament-selection process, which is critical to selecting the best individuals, which have been appraised for their fitness value, from the current generation for reproduction or to be survived in the successive generation and the rejected individuals are passed to the PSO for reformation. The population of PSO will be made of these rejected individuals who will connect with one another to update their position and velocity to achieve the best possible result from the non-fit individuals. The best individuals of PSO will be injected to the new population of GA.

The fitness function was applied in GA by implementing the RF classifier, evaluating the score of each solution, and observing how close the solution was to the one we needed. To calculate the fitness function, a chromosome must first be decoded in the binary representation. Tournament selection was applied with a size of 0.26. Although tournament selection is equivalent to rank selection with respect to the selection pressure, it is more effective in computation and more appropriate for parallel implementation [33].

The fitness of the individuals is calculated by 1) transforming the feature space of the dataset, and 2) applying k-fold cross-validation or holdout validation and obtaining a high accuracy score from the RF classifier. The selection probability of each individual is calculated as follows:

$$P_s(c) = \frac{v(c)}{\sum_{j=1}^{N} v(c)} \tag{4}$$

where $P_s(c)$ and $v(c)$ denote the probability of selection and the fitness value for the cth chromosome, respectively.

3.3.2 Crossover operator

In this process, two parent chromosomes are used to construct a new chromosome on the basis of crossover probability, which is 0.5 in the experiments. The constructed chromosome has a better string than those of its parent chromosomes. The following are the steps of the crossover:

1) A combination of two individual strings is chosen with the assistance of the reproduction operator.
2) A cross-site is randomly picked alongside the length of the string.
3) swapping the positions of values between the two strings.

3.3.3 Modified mutation operator

Upon the completion of the crossover process, the strings undergo mutation, which is the random change in the value of a gene. Mutation means that we flip a single bit from 0 to 1 or vice versa. The mutation operator is used to obtain a better solution by changing the current one. Mutation prevents the GA from being stuck in a local minimum. The mutation operator is

modified by implementing the discriminate mutation strategy based statistical analysis as illustrated in Section 3.2.2.

3.4 Random forest classification

In the proposed approach, the RF algorithm is used for binary classification. RF constructs many decision-trees during the training period and generates a class that has a mean prediction. The hyper-parameters of RF were tuned using a grid search. A wide range of parameter values were implemented in grid search as shown in Table 9. The best set of parameters extracted from the grid search was used to train random forest to get max classification accuracy.

We specify the number of random trees as 1000, maximal depth as 10 using confidence 0.5 in vote strategy and Gini impurity in criterion (split criterion); pruning and pre-pruning are applied; minimal leaf size is 2; the minimal size for splitting is 4. The Gini impurity is calculated as follows:

$$G = \sum_{j=1}^{C} p(j) * (1 - p(j)) \tag{5}$$

where C denotes the number of classes and $p(j)$ the probability of choosing a class j data point.

3.5 Performance measures

Four measures were implemented to assess the performance of the classification models: accuracy, recall, precision, receiver operating characteristic (ROC), and area under the ROC curve (AUC). Accuracy represents the rate of correctness of a classifier. Therefore, we take the sum of true positive (TP) records and true negative (TN) records and then divide by the total number of records which represents the sum of TN, TP, false negative (FN) and false positive (FP); thus, accuracy denotes the ratio of the number of correctly predicted records to the total number of records, as shown in Eq. (6). Recall represents the rate of values that measures positive records that the classifier correctly predicted. Moreover, it is called true positive rate (TPR) or sensitivity. Thus, recall is calculated as shown in Eq. (7). Precision is the ratio of TP records to the total of positive predicted records, as shown in Eq. (8). The ROC curve is a graph of TPR versus false positive rate (FPR), where TPR is on the y-axis and FPR on the x-axis. The AUC metric is used to calculate AUC, and it describes the separability measurement or degree. It informs how the model can identify among classes.

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

Table 9 Grid search values for the proposed RF

Parameter	Grid search value
No of trees	50, 100, 200, 500, 1000, 2000
max_depth	3, 5, 10, 15
min_samples_split	2, 5, 10, 20
min_samples_leaf	1, 5, 10, 12
max_features	auto, log2, sqrt

$$Recall = \frac{TP}{FN + TP} \tag{7}$$

$$Precision = \frac{TP}{FP + TP} \tag{8}$$

4 Experiments results and discussion

In this section, two public datasets, namely, Cleveland and Statlog, are used to evaluate the proposed approach, and then the classification performance of our approach is compared with those of other state-of-the-art methods. Moreover, the proposed approach will be compared with the methods that implement GA and with those that do not implement GA. In addition, we will discuss the complexity of the proposed approach.

4.1 Experimental setup

In this section, two types of experiments are implemented on the Cleveland and Statlog datasets to assess the efficacy of the proposed model. All the computations are performed on Google CoLab, which provides GPU Tesla k80 with 12 GB of GDDR5 VRAM, and Intel Xeon Processor with two 2.20-GHz cores and 13 GB RAM. Moreover, the Python programming software package scikit-learn is used for the experiments.

4.2 Results of the Cleveland dataset

The model was applied on the Cleveland heart-disease dataset, which had 13 features. All the 303 heart-disease records of the dataset were considered. To assess the classification performance, the results obtained from the experiments of the proposed model were compared with those of other state-of-the-art methods in terms of heart-disease prediction. In the first experiment, throughout cross-validation, the data records were divided into 10 folds; one-fold was used in testing, and the remaining nine folds were used in training. The data were divided into folds via stratified sampling, meaning that the class distribution (defined by the label attribute) in the subsets/folds was the same as that in the complete dataset. Finally, the result was obtained by averaging all the 10 iterations.

In the second experiment, we performed the train/test holdout validation. The data were split as 70% for training and 30% for testing. The model was trained on 212 records and tested on the remaining 91 as unseen data. The primary reason behind using this distribution is to satisfactorily compare our approach with those in other researches on the same dataset. We ran the same experimental procedure five times, following which the mean of the five results was calculated. Table 10 compares the results of these two experiments with those of recent researches. Evidently, the proposed approach achieves better classification results than those of most methods. The experimental results on the Cleveland dataset confirm that the proposed approach achieves the accuracy rates of 87.8% and 95.6% for 10-fold and holdout (TR =

Table 10 Benchmarking our approach with others in the literature on the Cleveland dataset

Study	Method			Accuracy
	Pre-Processing	Feature Selection	Classification	
El-Bialy et al. [16]	–	Manual	C4.5 algorithm and fast decision tree	78.54 (10-fold)
Paul et al. [35]	handling missing values	Correlation coefficient & GA	fuzzy rules	80 (holdout)
Saqlain et al. [46]	Data standardization	Fisher score and the Matthews correlation coefficient	SVM	81.19 (10-fold)
Shah et al. [48]	Normalization	parallel probabilistic principal component analysis	SVM	82.18 (10-fold)
Vivekanandan and Iyengar [51]	*min--max* normalization	modified differential evolution algorithm	Integrated model of fuzzy AHP & ANN	83 (holdout)
Suresh and Ananda Raj [50]	handling missing values	GA	NB	83.20 (k-fold)
Buettner et al. [11]	Grouping	–	RF	84.40 (10-fold)
Chitra and Seenivasagam [12]	*min--max* normalization	–	Feed Forward and Cascaded Correlation NN	85 (holdout)
Mathan et al. [32]	–	–	Gini index-based decision tree & Neural network	85.30 (10-fold)
Jha et al. [23]	–	–	Random subspace classifier	86.14 (3-fold)
Saxena et al. [47]	–	All Possible-MV algorithm to handle the missing values	Decision-trees and rules generated	86.70 (10-fold)
Subanya et al. [49]	–	Artificial bee colony algorithm	SVM	86.76 (holdout)
Amin et al. [5]	remove the missing values	Tests 8100 combinations of the features	Vote algorithm as a hybrid methodology of logistic regression and NB	86.87 (10-fold)
Gokulnath and Shantharajah [18]	*Z* score	GA	SVM	88.34 (holdout)
Yazid et al. [54]	–	–	Flower pollination neural network	89.60 (holdout)
Ali et al. [3]	–	–	Hybrid grid search algorithm using two optimized SVM models	92.20 (holdout)
Ali et al. [4]	–	χ^2 statistical model	Deep neural network	93.33 (holdout)
GAPSO-RF (ours)	*min--max* normalization	GA-PSO	RF	**87.80 (10-fold)**
GAPSO-RF (ours)	*min--max* normalization	GA-PSO	RF	**95.60 (holdout)**

70%) validations, respectively. In the 10-fold validation, the proposed approach increases the average accuracy by 6.61%, 5.62%, and 2.67% over Saqlain et al. [46], Shah et al. [48], and Mathan et al. [32], respectively. The results of the other experiment (holdout) demonstrate that

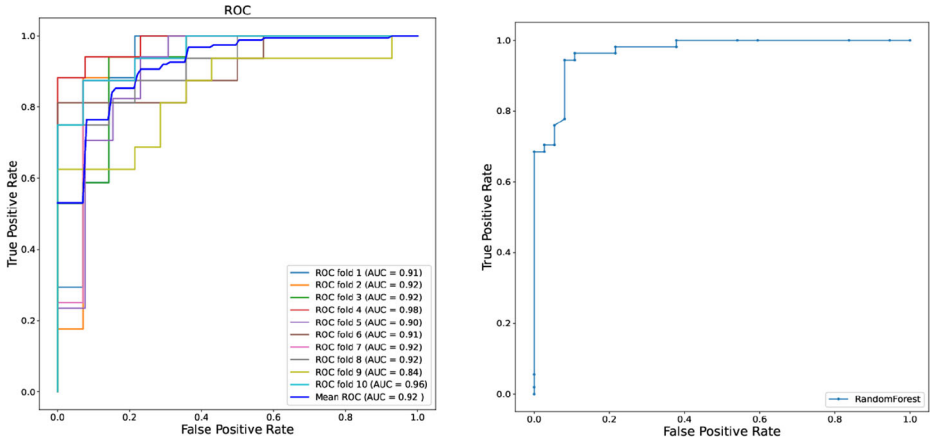
the proposed increases the average accuracy by 7.26%, 3.4%, and 2.27% over Gokulnath and Shantharajah [18], Ali et al. [3], and Ali et al. [4], respectively. Figure 3a and b depict the ROC analysis for the 10-fold and holdout (TR = 70%) validations, respectively. Table 11 compares results of using subset features selected of GA for Cleveland dataset with other models. The results show that RF overcomes the other models.

4.3 Results of the Statlog dataset

We compare our approach with several benchmark approaches on the Statlog heart-disease dataset, which contained 13 features. All the 270 heart-disease records of the dataset were considered. We performed the same experiments on the Statlog dataset as those performed on the first dataset, i.e., Cleveland. First, the data records were partitioned into 10 folds, and the results were obtained by calculating the mean of all the ten iterations. Second, the data were split as 70% for training and 30% for testing (i.e., holdout (TR = 70%). The model was trained on 189 records and tested on the remaining 81 records as unseen data. The primary reason behind using this distribution was to satisfactorily compare our approach with those in other researches on the same dataset. We performed the same experimental procedure five times and recorded the average of the five results.

Table 12 compares the results of the proposed approach with those of the recent state-of-the-art heart-disease-prediction methods. Evidently, our approach obtains the accuracy rates of 87.78% and 91.4% for the 10-fold and holdout (TR = 70%) validations, respectively, the best results achieved on the Statlog dataset thus far. In the 10-fold validation, the proposed approach increases the average accuracy by 11.18% and 3.78% over El-Bialy et al. [16] and Rado et al. [40], respectively. In the other experiment (holdout (TR = 70%)), the results proved that the proposed approach increases the average accuracy by 12.62% and 1.4% over Long et al. [30] and Karthikeyan and Kanimozhi [24], respectively.

Figure 4a and b depict the ROC analysis for the 10-fold and holdout (TR = 70%) validations, respectively. Table 13 compares results of using subset features selected of GA for statlog dataset with other models. The results show that RF overcomes the other models.



(a) (b)

Fig. 3 ROC curve of the Cleveland dataset for **a** 10-fold, and **b** holdout (TR = 70%)

Table 11 Compare the subset features selected from GA with other classifiers for Cleveland dataset

Model	Accuracy (%)
Logistic regression (SVM)	80.22
SVM	81.32
Decision trees	82.42
NB	85.71
Proposed RF	**95.60**

Although several studies have implemented FS in their proposed methods, little attention has been given to optimizing fitness function in GA. In our proposed approach, we used RF as a fitness function in GA to get the maximum value of classification accuracy. We tuned the hyper-parameters of RF using grid search; after that, we applied pruning and pre-pruning. Besides that, we implemented the experiments in GA different hyper-parameters and different types of crossover, mutation and selection as shown in Fig. 5. In the crossover, we applied uniform as it achieved better results than one point. Moreover, we used tournament selection as it produced better results than roulette wheel.

Table 14 summarizes the performance-evaluation results for both 10-fold and holdout (TR = 70%) validations on the Statlog dataset. As seen from Table 14, our proposed approach with the optimal selected features achieved a better performance than that achieved upon using all the features at once. Additionally, our proposed approach increases the average accuracy on

Table 12 Benchmarking our approach with others in the literature on the Statlog dataset

Study	Method			Accuracy (%)
	Pre-Processing	Feature Selection	Classification	
El-Bialy et al. [16]	–	manual	C4.5 algorithm and fast decision tree	76.60 (10-fold)
Long et al. [30]	*min--max* normalization	Chaos-based firefly algorithm & rough set	type-2 fuzzy logic	78.78 (holdout)
Rado et al. [40]	remove the missing values	Correlation-based FS & Feature importance & Recursive feature elimination	SVM	84 (10-fold)
Mukherjee et al. [34]	–	–	Multi-layer perceptron ensembles & SVM & Generalized additive model	85 (10-fold)
Yazid et al. [54]	–	–	Flower pollination neural network	89.60 (holdout)
Karthikeyan and Kanimozhi. [24]	–	–	CNN & Deep belief network algorithm	90 (holdout)
GAPSO-RF (ours)	*min--max* normalization	GA-PSO	RF	**87.78 (10-fold)**
GAPSO-RF (ours)	*min--max* normalization	GA-PSO	RF	**91.40 (holdout)**

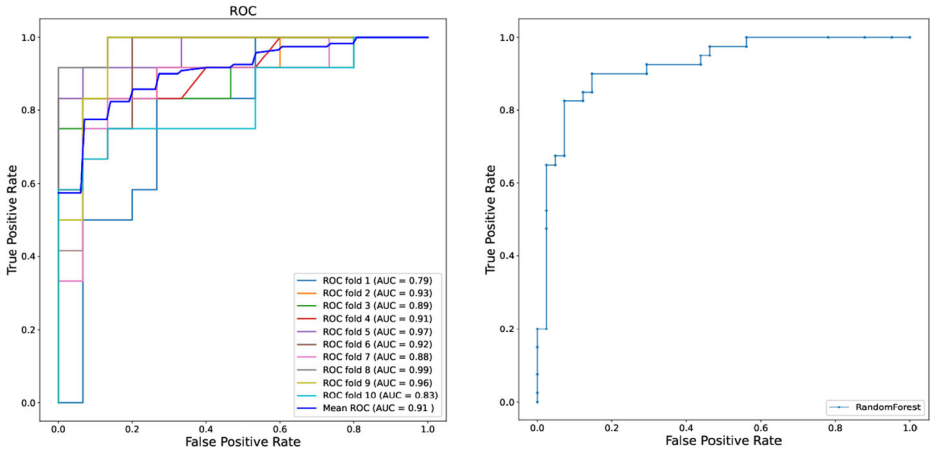


Fig. 4 ROC curve of the Statlog dataset for the **a** 10-fold and **b** holdout (TR = 70%) validations

the Statlog dataset by 5.93% and 7.45% in the 10-fold and holdout (TR = 70%) validations, respectively.

4.4 Effectiveness of FS

In this subsection, the performance of the FS process in GA is evaluated. FS improves the performance of the proposed approach compared with using all the features at once. The features sets of the Cleveland and Statlog datasets are reduced by 46.15% and 30.77%, respectively. From Table 15, it is evident the proposed approach decreased the number of features. The features selected in Cleveland dataset are 7 features (*Cp, Fbs, Restecg, Exang, Slope, Ca,* and *Thal*) and for statlog dataset the features selected are 9 features (*Age, Sex, Cp, Fbs, Thalach, Exang, Slope, Ca,* and *Thal*). The overall measurement results for the Cleveland dataset both with and without FS on GA are summarized in Table 16. As previously mentioned, the experiment was performed twice (10-fold and holdout (TR = 70%)). From the experimental results in Table 16, one can see that the proposed approach with selected optimal features achieves better performance than that achieved upon using all the features at once. Our proposed approach increases the average accuracy on the Cleveland dataset by 4.33% and 6.59% in the 10-fold and holdout (TR = 70%) validations, respectively.

Table 13 Compare the subset features selected from GA with other classifiers for Statlog dataset

Model	Accuracy (%)
Decision trees	74.07
Logistic regression (SVM)	83.95
SVM	83.95
NB	88.89
Proposed RF	**91.40**

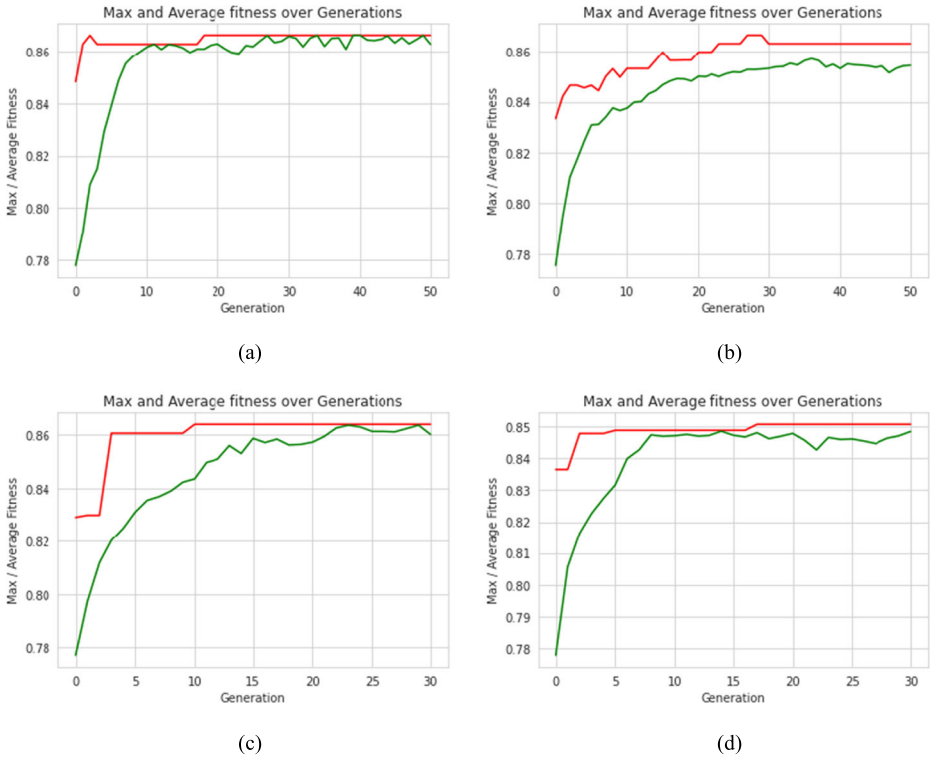


Fig. 5 Different hyper-parameters which affect the efficiency of GA. Red and green curves represent the max and average fitness, respectively. **a** number of generations = 50, population size = 50, crossover rate = 0.5 and mutation rate = 0.07. **b** number of generations = 50, population size = 50, crossover rate = 0.5 and mutation rate = −0.4. **c** number of generations = 30, population size = 50, crossover rate = 0.5 and mutation rate = 0.07. **d** number of generations = 30, population size = 50, crossover rate = 0.5 and mutation rate = 0.08

4.5 Time complexity

In this subsection, we compare the time complexity of our proposed approach GAPSO-RF with GA-based different models. Table 17 shows the comparison of computational cost among the approaches on the Cleveland dataset. This table records the complexity time (i.e., FS and classification), number of generations, and prediction accuracy. We can see that the complexity time of our approach GAPSO-RF is not the best. However, as concluded from Table 17, our proposed approach achieves the best prediction accuracy compared to other GA-based methods. In addition, the proposed approach reached the best rate of 87.80% by the minimum number of generations 30.

Table 14 Evaluation of the FS used in the proposed approach on the Statlog dataset

	FS	Accuracy (%)	Precision (%)	Recall (%)	AUC (%)
10-fold	Without	81.85	82.17	86.00	89.50
	With	**87.78**	**87.26**	**91.33**	**91.00**
Holdout	Without	83.95	86.36	84.44	91.90
	With	**91.40**	**89.58**	**95.56**	**92.60**

Table 15 Feature-dimension details for the Cleveland and Statlog datasets

DataSet	Original	Selected Feature Subset	Reduced (%)
Cleveland	13	7	46.15
Statlog	13	9	30.77

Table 16 Evaluation of the FS employed in the proposed approach on the Cleveland dataset

	FS	Accuracy (%)	Precision (%)	Recall (%)	AUC (%)
10-fold	Without	83.17	83.72	78.26	91.30
	With	**87.80**	**89.76**	**82.61**	**92.00**
Holdout	Without	89.01	87.97	84.78	91.00
	With	**95.60**	**97.44**	**92.68**	**94.00**

Table 17 Comparison of computational time of GAPSO-RF with other GA based different models

Parameter Comparison	Proposed GAPSO-RF	Conventional GA	GA-SVM	GA-NB	GA-CNN
Number of generations	30	100	100	100	30
Optimum classification (%)	87.80	87.13	85.64	83.68	86.39
Time (second)	2559.10	7499.98	148.05	111.41	9185.88

First, the proposed GAPSO-RF approach improves the conventional GA as follows: 1) Reduce the number of generations by 70%, 2) Execution time improved by 65.87%, and 3) Improve the prediction accuracy by 0.67%. Second, a convolutional neural network (CNN) is good in feature selection, as mentioned in [17, 52]. However, it consumes a higher computation cost. Results show that GA-based CNN has a higher computational cost than others. In addition, the proposed GAPSO-RF outperforms GA-CNN in prediction accuracy and execution time. In the future, the CNN model can be used for different types of datasets (e.g., electrocardiogram (ECG) signals and images). The implementation of a discriminate mutation strategy in GA-based statistical analysis and the implementation of the PSO in local search are reasons to reduce the number of generations and thus reduce the execution time. Despite this, the execution time in our approach is relatively large. In the future, we intend to develop an efficient feature selection method with low complexity and high performance. Third, GA-NB achieves the best execution time due to the number of iterations in PSO and RF. Nevertheless, we outperform this approach by 5.88% in the average prediction accuracy.

5 Conclusion

We presented a GAPSO-RF-based FS approach with an RF classifier as the base of a fitness function to select significant features to increase the accuracy of heart-disease diagnosis. The proposed approach achieved high accuracy of 95.6% and 91.4% on the Cleveland and Statlog datasets, respectively. After that, the results of the proposed FS method are compared with the

results without using FS and found that it outperforms in accuracy. Moreover, it outperformed the existing state-of-the-art methods on the same datasets. Furthermore, a comparative analysis is performed between GAPSO-RF and conventional GA and found that our proposed approach outperformed conventional GA. Additionally, we protected our model from overfitting by using the RF algorithm for classification. Hence, our experimental results confirmed that the proposed approach enhanced the decision-making process of the practitioners during heart-disease diagnosis.

The following are some of the limitations of this research: First, more classifiers should be evaluated to have a more extensive evaluation of the results. Second, the proposed model's key drawbacks are its high computational cost and temporal complexity, as it is based on the wrapper feature selection strategy. More studies need to be conducted to address the limitations of the proposed approach in the future. First, multi-objective genetic algorithm can be applied. Second, to overcome small-data limitations in heart-disease prediction, we plan to use in future work surrogate data or merging different heart-disease datasets. Finally, for electrocardiogram (ECG) signals and images, further study can be carried out for improving the features selection by using convolutional neural network (CNN). Moreover, we intend to develop an efficient classifier to improve the performance.

Funding Open access funding provided by The Science, Technology & Innovation Funding Authority (STDF) in cooperation with The Egyptian Knowledge Bank (EKB).

References

1. Abdel-Basset M, Gamal A, Manogaran G, Long HV (2019) A novel group decision making model based on neutrosophic sets for heart disease diagnosis. Multimed Tools Appl:1–26
2. Adler ED, Voors AA, Klein L, Macheret F, Braun OO, Urey MA, Zhu W, Sama I, Tadel M, Campagnari C, Greenberg B, Yagil A (2020) Improving risk prediction in heart failure using machine learning. Eur J Heart Fail 22:139–147
3. Ali L, Niamat A, Khan JA, Golilarz NA, Xingzhong X, Noor A, Nour R, Bukhari SAC (2019) An optimized stacked support vector machines based expert system for the effective prediction of heart failure. IEEE Access 7:54007–54014
4. Ali L, Rahman A, Khan A, Zhou M, Javeed A, Khan JA (2019) An automated diagnostic system for heart disease prediction based on x2 statistical model and optimally configured deep neural network. IEEE Access 7:34938–34945
5. Amin MS, Chiam YK, Varathan KD (2019) Identification of significant features and data mining techniques in predicting heart disease. Telematics Inform 36:82–93
6. Asoh H, Mühlenbein H (1994) On the mean convergence time of evolutionary algorithms without selection and mutation. In: International conference on parallel problem solving from nature, pp 88–97
7. Atal DK, Singh M (2020) A dictionary matrix generation based compression and bitwise embedding mechanisms for ECG signal classification. Multimed Tools Appl 79:13139–13159

8. Banerjee D, Thompson C, Kell C, Shetty R, Vetteth Y, Grossman H et al (2017) An informatics-based approach to reducing heart failure all-cause readmissions: the Stanford heart failure dashboard. J Am Med Inf Assoc 24:550–555
9. Beasley D, Bull DR, Martin RR (1993) An overview of genetic algorithms: part 1, fundamentals. Univ Comput 15:56–69
10. Benjamin EJ, Muntner P, Bittencourt MS (2019) Heart disease and stroke statistics-2019 update: a report from the American Heart Association. Circulation 139:e56–e528
11. Buettner R, Schunter M (2019) Efficient machine learning based detection of heart disease, presented at the IEEE international conference on E-health networking, Application & Services (HealthCom), Bogota, Colombia, Colombia
12. Chitra R, Seenivasagam V (2015) Heart disease prediction system using intelligent network. In: Power electronics and renewable energy systems. Springer, pp 1377–1384
13. Dua D, Graff C (2017) UCI Machine Learning Repository. Available: http://archive.ics.uci.edu/ml
14. Durairaj M, Sivagowry S (2014) A pragmatic approach of preprocessing the data set for heart disease prediction. Int J Innov Res Comput Commun Eng 2:6457–6465
15. Dwivedi AK (2018) Performance evaluation of different machine learning techniques for prediction of heart disease. Neural Comput Appl 29:685–693
16. El-Bialy R, Salamay MA, Karam OH, Khalifa ME (2015) Feature analysis of coronary artery heart disease data sets. Procedia Comput Sci 65:459–468
17. Garcia-Garcia A, Orts-Escolano S, Oprea S, Villena-Martinez V, Garcia-Rodriguez J (2017) A review on deep learning techniques applied to semantic segmentation. arXiv preprint arXiv:1704.06857
18. Gokulnath CB, Shantharajah SP (2018) An optimized feature selection based on genetic approach and support vector machine for heart disease. Clust Comput 22:1–11
19. Gupta V, Mittal M (2019) R-peak detection in ECG signal using yule–Walker and principal component analysis. IETE J Res 67:1–14
20. Gupta V, Mittal M, Mittal V (2020) Performance evaluation of various pre-processing techniques for R-peak detection in ECG signal. IETE J Res:1–16
21. Halder B, Mitra S, Mitra M (2019) Classification of complete myocardial infarction using rule-based rough set method and rough set explorer system. IETE J Res:1–11
22. Ismaeel S, Miri A, Chourishi D (2015) Using the extreme learning machine (ELM) technique for heart disease diagnosis. In: IEEE Canada international humanitarian technology conference (IHTC2015), Ottawa, ON, Canada, pp 1–3
23. Jha SK, Pan Z, Elahi E, Patel N (2019) A comprehensive search for expert classification methods in disease diagnosis and prediction. Expert Syst 36:e12343–e12343
24. Karthikeyan T, Kanimozhi V (2017) Deep learning approach for prediction of heart disease using data mining classification algorithm deep belief network. Int J Adv Res Sci Eng Technol 4:3194–3201
25. Kaur P, Kumar R, Kumar M (2019) A healthcare monitoring system using random forest and internet of things (IoT). Multimed Tools Appl 78:19905–19916
26. Kennedy J, Eberhart R (1995) Particle swarm optimization. In: Proceedings of ICNN'95-international conference on neural networks, pp 1942–1948
27. Khan RU, Zhang X, Kumar R, Sharif A, Golilarz NA, Alazab M (2019) An adaptive multi-layer botnet detection technique using machine learning classifiers. Appl Sci 9:2375
28. Kohli R, Garg A, Phutela S, Kumar Y, Jain S (2021) An improvised model for securing cloud-based E-Healthcare systems. In: IoT in Healthcare and ambient assisted living. Springer, pp 293–310
29. Krishnaiah V, Narsimha G, Chandra NS (2015) Heart disease prediction system using data mining technique by fuzzy K-NN approach, vol 337. Springer, Cham
30. Long NC, Meesad P, Unger H (2015) A highly accurate firefly based algorithm for heart disease prediction. Expert Syst Appl 42:8221–8231
31. Luo M, Wu K (2020) Heart rate prediction model based on neural network. IOP Conf Ser Mater Sci Eng 715:012060–012060
32. Mathan K, Kumar PM, Panchatcharam P, Manogaran G, Varadharajan R (2018) A novel Gini index decision tree data mining method with neural network classifiers for prediction of heart disease. Des Autom Embed Syst 22:225–242
33. Mitchell M (1998) An introduction to genetic algorithms. MIT press
34. Mukherjee S, Kapoor S, Banerjee P (2017) Diagnosis and identification of risk factors for heart disease patients using generalized additive model and data mining techniques. J Cardiovasc Dis Res 8:137–144
35. Paul AK, Shill PC, Rabin MRI, Akhand MAH (2016) Genetic algorithm based fuzzy decision support system for the diagnosis of heart disease. In: 5th international conference on informatics, electronics and vision (ICIEV), Dhaka, Bangladesh, pp 145–150

36. Polat K, Şahan S, Güneş S (2007) Automatic detection of heart disease using an artificial immune recognition system (AIRS) with fuzzy resource allocation mechanism and k-nn (nearest neighbour) based weighting preprocessing. Expert Syst Appl 32:625–631
37. Prado R, García-Galán S, Yuste AJ, Expósito JM (2010) A fuzzy rule-based meta-scheduler with evolutionary learning for grid computing. Eng Appl Artif Intell 23:1072–1082
38. Priyatharshini R, Chitrakala S (2019) A self-learning fuzzy rule-based system for risk-level assessment of coronary heart disease. IETE J Res 65:288–297
39. Purnomo A, Barata MA, Soeleman MA, Alzami F (2020) Adding feature selection on Naïve Bayes to increase accuracy on classification heart attack disease. J Phys Conf Ser 1511:012001–012001
40. Rado O, Ali N, Sani HM, Idris A, Neagu D (2019) Performance analysis of feature selection methods for classification of healthcare datasets. In: Intelligent computing-proceedings of the computing conference, pp 929–938
41. Reddy GT, Khare N (2017) Hybrid firefly-bat optimized fuzzy artificial neural network based classifier for diabetes diagnosis. Int J Intell Eng Syst 10:18–27
42. Reddy GT, Khare N (2017) An efficient system for heart disease prediction using hybrid OFBAT with rule-based fuzzy logic model. J Circ Syst Comput 26:1750061
43. Reddy GT, Reddy MPK, Lakshmanna K, Rajput DS, Kaluri R, Srivastava G (2020) Hybrid genetic algorithm and a fuzzy logic classifier for heart disease diagnosis. Evol Intel 13:185–196
44. Revett K, Gorunescu F, Salem A-B, El-Dahshan E-S (2009) Evaluation of the feature space of an erythematosquamous dataset using rough sets. Ann Univ Craiova-Math Comput Sci Ser 36:123–130
45. Saifudin A, Nabillah UU, Yulianti, Desyani T (2020) Bagging technique to reduce misclassification in coronary heart disease prediction based on random forest. J Phys Conf Ser 1477:032009–032009
46. Saqlain SM, Sher M, Shah FA, Khan I, Ashraf MU, Awais M, Ghani A (2019) Fisher score and Matthews correlation coefficient-based feature subset selection for heart disease diagnosis using support vector machines. Knowl Inf Syst 58:139–167
47. Saxena K, Sharma R, others (2016) Efficient heart disease prediction system. Procedia Comput Sci 85:962–969
48. Shah SMS, Batool S, Khan I, Ashraf MU, Abbas SH, Hussain SA (2017) Feature extraction through parallel probabilistic principal component analysis for heart disease diagnosis. Phys A: Stat Mech Appl 482:796–807
49. Subanya B, Rajalaxmi RR (2014) Feature selection using Artificial Bee Colony for cardiovascular disease classification. In: 2014 International Conference on Electronics and Communication Systems (ICECS), pp 1–6
50. Suresh P, Ananda Raj MD (2018) Study and analysis of prediction model for heart disease: an optimization approach using genetic algorithm. Int J Pure Appl Math 119:5323–5336
51. Vivekanandan T, Iyengar NCSN (2017) Optimal feature selection using a modified differential evolution algorithm and its effectiveness for prediction of heart disease. Comput Biol Med 90:125–136
52. Voulodimos A, Doulamis N, Doulamis A, Protopapadakis E (2018) Deep learning for computer vision: a brief review. Comput Intell Neurosci 2018:1–13
53. Wang Z, Zhu Y, Li D, Yin Y, Zhang J (2020) Feature rearrangement based deep learning system for predicting heart failure mortality. Comput Methods Prog Biomed 191:105383–105383
54. Yazid MHBA, Talib MS, Satria MH (2019) Flower pollination neural network for heart disease classification. In: IOP Conference Series: Materials Science and Engineering, pp 012072–012072