scientific reports

Check for updates

OPEN

A hybrid approach with metaheuristic optimization and random forest in improving heart disease prediction

Geetha Narasimhan & Akila Victor✉

Cardiovascular diseases (CVD) a major cause of morbidity and mortality among the world's non-communicable disease incidences. Though these practices are in use for diagnostics of different CVDs in clinical settings, need improvement because they are solving the purpose of only 57% of the patients in emergency. Due to this cost of diagnosis for heart disease is increasing which is the reason for analyzing heart disease and predicting it as early as possible. The main motive of this paper is to find an intelligent method for predicting disease effectively by means of machine learning (ML) and metaheuristic algorithms. Optimization techniques have the merit of handling non-linear complex problems. In this paper, an efficient ML model along with metaheuristic optimization techniques is evaluated for heart disease dataset to enhance the accuracy in predicting the disease. This will help to reduce the death rate due to the severity of heart disease. The SelectKBest feature selection is applied to the Cleveland Heart dataset and overall rank is obtained. Accuracy is measured. The optimization techniques namely Genetic Algorithm Optimized Random Forest (GAORF), Particle Swarm Optimized Random Forest (PSORF), and Ant Colony Optimized Random Forest (ACORF) are applied to the Cleveland dataset. Classification algorithms are performed before and after optimization. The output of the experiment explains that the GAORF performed better for the dataset considered. Also, a comparison is made along with the SelectKBest filter methods. The proposed model achieved better accuracy which is the maximum among other optimization and classification techniques.

Keywords Metaheuristic, Feature selection, Optimization techniques, Genetic algorithm (GAO), Particle swarm optimization (PSO), Ant colony optimization (ACO), Random Forest (RF)

Healthcare Challenges are misdiagnosis and delayed diagnosis are prevalent, impacting 5–15% of patients globally. Conventional methods miss early detection and struggle with complex diseases. Machine Learning to the rescue offers improved accuracy and handles large data volumes for better diagnosis. ML can predict and prevent diseases, potentially reducing mortality. CVD leading cause of death worldwide, responsible for approximately 17.9 million fatalities each year[1–5]. Current methods like ECGs and echocardiograms only accurately diagnose 57% of CVD patients in emergencies. Many CVD patients are asymptomatic or have undetectable changes, making early diagnosis crucial. The solution would be integrating medical algorithms with existing techniques that improve accuracy and facilitate better decision-making in diagnosis, treatment, and research. Leverages patient data for disease identification, risk prediction, and refined diagnoses[6–10].

The challenges to using medical algorithms are lack of awareness, uncertain capabilities, complex results, and access issues that hinder adoption. Automation could improve accuracy, and data sharing, and reduce administrative burden. The methodology used is Data Acquisition to collect patient data (age, weight, blood pressure, etc.). Feature Selection method is used select relevant features relevant features. Optimization Algorithms like Genetic Algorithm: Evolves solutions like natural selection. Particle Swarm Optimization: Particles learn from each other to find the best solutions. Ant Colony Optimization: Pheromone trails guide "ants" to the best solutions. Performance Evaluation to measure classifier accuracy on unseen data. Comparison to select the most accurate algorithm for the dataset. The results are the optimized features for heart disease classification and the potential for diagnostic tools and improved treatments. The model uses the Cleveland

School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, Tamilnadu, India. ✉email: akilavictor@vit.ac.in

Dataset (303 records, 13 features). Compared 3 optimization algorithms (PSO, ACO, Genetic). Random Forest was chosen as the best-fitting ML algorithm.

The work is approached with a Hybrid model that uses the PSO for optimizing the local solution ACO is adopted for continuous optimization of the local solution. This adaptation reduces the complexity and of complex datasets accuracy is improved. The rank of the selected system is obtained using the SelectKBest filter feature and thus fed to classification algorithms. When it comes to classification algorithms Support Vector Machine (SVM), Random Forest (RF), Naïve Bayes(NB), Logistic Regression(LR), Decision tree(DT), Random Forest with Grid Search (RFGS) and Extended Gradient Boost (XGB)[12]. By optimizing the data with GAO, PSO, and ACO which is fed to RF to obtain the best possible accuracies with improved performances. In this study, the feature selection with three different techniques and an overall ranking method is applied rather than relying on a single technique. Also, three different optimization techniques are applied to find the best method for the dataset and adopted with machine learning techniques. The study suggested that GAO with RF performs better for the dataset considered for the study. The list of symbols used in the study are listed in Table 1.

This article highlights the boundaries of current diagnostic approaches for heart disease, emphasizing the need for improvement. It emphasizes the importance of early prediction for better patient outcomes and reduced mortality rates. The model introduces the concept of using ML and metaheuristic algorithms for improved heart disease prediction. It showcases the potential of using optimization techniques like Genetic Algorithm (GAO), Particle Swarm Optimization (PSO), and Ant Colony Optimization (ACO) to enhance the accuracy of machine learning models like Random Forest. This methodology provides a specific example of using the SelectKBest feature selection technique on the Cleveland Heart Disease dataset. It demonstrates the process of comparing the performance of different optimization techniques (GAORF, PSORF, ACORF) with a baseline model (Random Forest). The proposed model establishes a benchmark for accuracy achieved using the proposed GAORF model on the Cleveland dataset. It encourages further research by mentioning the possibility of achieving even better accuracy with different techniques.

Our proposed work is categorized into 5 different sections "Related work" , which refers to the literature on ML algorithms for machine learning processes that detect CVDs, "Materials and methods" section, which discusses the framework for the detection of CVDs, design setup, and features selected and "Optimization algorithms" section, discusses different machine learning algorithms for optimization of a better algorithm for the dataset. In the final section, results extracted from different models are done with performance evaluation that decides the best algorithm from the work. In "Materials and methods" section, machine learning algorithms like the PSO, GA, ACO are applied to the selected datasets[44,45].

Need for hybrid ml model for diagnosis of cardiovascular diseases

Extensive literature is reported in predicting heart disease considering a single classification or combination of algorithms known as hybrid techniques[13]. Despite this very few researchers have focused on optimization techniques and however, researchers have failed to analyze the difference between these optimization techniques when applied for a single dataset.

In this proposed framework, the SelectKBest Method is used to find the rank of each feature using Chi-square, Mutual information, and F-statistics. The overall rank is considered and based on that features are selected and given to classification techniques. GAO, ACO and PSO optimize the dataset and given to classification techniques is given to classification techniques. A comparison of classification algorithms is evaluated before and after optimization to determine an optimized model for the selected dataset that gives better accuracies where the workflow for the hybrid model is shown in Fig. 1.

Symbol	Meaning	Description
ML	Machine Learning	A field within artificial intelligence and computer science that focuses on utilizing data and algorithms to create AI systems that mimic human behavior.
GAORF	Genetic Algorithm Optimized Randon Forest	A technique for addressing both constrained and unconstrained optimization problems by simulating the natural selection process inspired by biological evolution.
PSORF	Particle Swarm Optimized Random Forest	This is a population-based selection. This algorithm is to obtain global optimum fitness function in a given area.
ACORF	Ant Colony Optimized Random Forest	This is probabilistic method for solving computational problems which can be reduced to find good paths through graphs.
CVD	Cardiovascular diseases	Is a general term for conditions that affect heart or blood vessels.
RF	Random Forest	This algorithm combines multiple decision tree's output to reach one result.
SVM	Support Vector Machine	This algorithm solves complex classifications, and other regression, outlier detection by carrying optimal data transformation.
NB	Naïve Bayes	This algorithm solves classifications where high dimensional datasets are trained.
LR	Logistic Regression	This algorithm performs binary classification by estimating the probability of a given outcome, event, or observation.
DT	Decision Tree	This algorithm is used to make predictions based on the decision trees.
RFGS	Random Forest Grid Search	This method is used to tune the hyperparameters of Random Forest algorithm in ML
XGB	Extended Gradient Boosting	A ML algorithm used to solve regression, classification, user defined predictions ranking problems.
KNN	K Nearest Neighbor	This algorithm determines classifications or predictions by analyzing the proximity of an individual data point to others.
UCI	University of California	A machine learning repository which consists of 665 datasets from various fields.

Table 1. List of symbols used in the manuscript and its explanation.

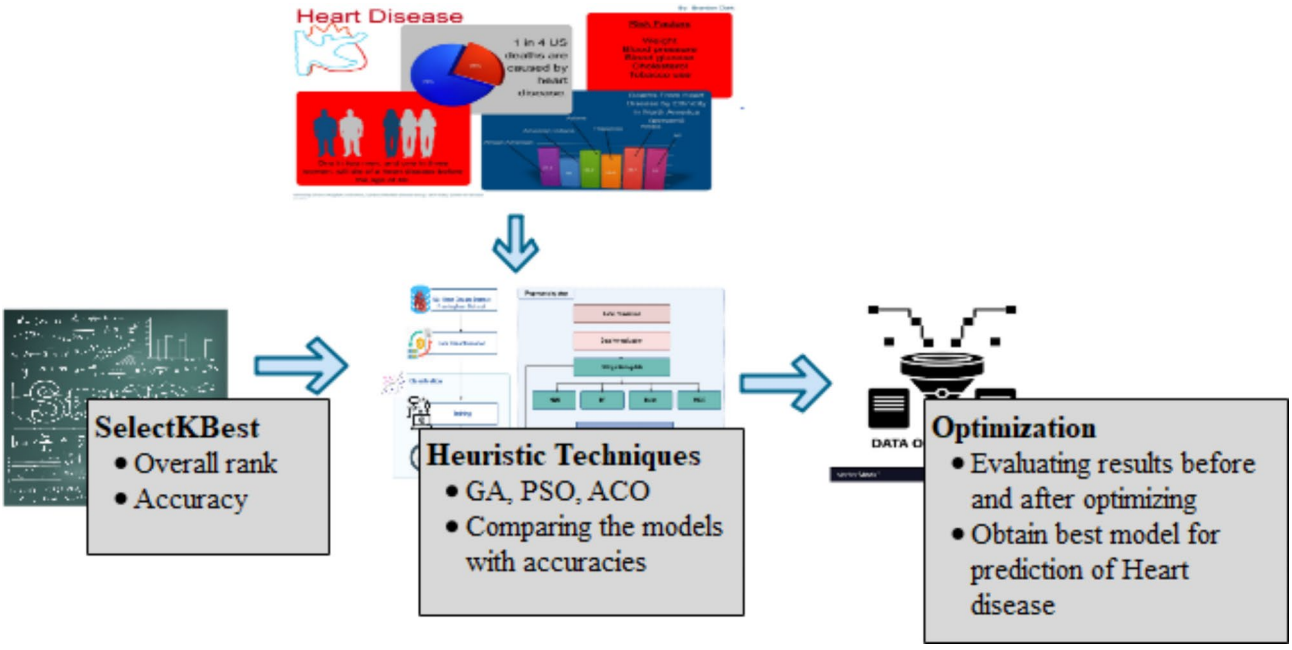

Fig. 1. Optimization model for cardiovascular disease diagnosis[49–51].

Related work

Various situation

The objective is to predict heart disease by applying swarm intelligence and ANN for which researchers used UCI dataset and applied PSO algorithm for feature selection[6]. ANN is applied after features are selected and found PSO performed better in predicting heart disease. In the case of heart diseases, the prediction using PSO and ACO along with other ML techniques, the PSO using feed-forward has resulted in better accuracies[12]. A combined method of ML using Mutual information(MI) and Binary PSO referred to as MI_BPSO is used for optimization to predict heart disease[3]. It is found that random forest with integrated Ant colony particle swarm optimization works better and achieves better performance after the PSO. Similarly[14] worked on heart disease prediction using optimization techniques. The PSO algorithm is combined with logistic regression for classifying heart disease binary, feature selection using dragonfly algorithm and found that have high accuracies[13,15,16]. In[17], a model was created for predicting heart disease. The model used a binary artificial bee colony for finding the best features and KNN was used for the classification technique. The model achieved 92.4% accuracy.

Khourdifi, Y., & Bahaj, M in 2019 used PSO along with MLP for predicting heart disease[9]. The dataset is used from the UCI repository namely the Cleveland dataset. The model achieved around 84.61% accuracy, whereas in[18] modified Artificial bee colony-dependent feature selection is applied for predicting heart disease. In[19] an optimal feature subset selection based on ACO is used to predict breast cancer and liver disorder with the model achieving 96.56% and 92.44% respectively. In[20] a cuckoo-inspired algorithm for feature selection for predicting heart disease. In this method, the Cuckoo search algorithm performed better than the Cuckoo optimization algorithm. Asadi et al., in 2021 used ACO for classifying data and the proposed model used swarm optimization for feature selection[21]. Velswamy K et al. used random forest along with multi-objective swarm optimization for predicting heart disease in 2021[22].

Jabbar M.A, in 2013 used a model for the classification of heart disease and Modified bee algorithms for feature selection. SVM, Naïve Bayes, DT, and RF are used for classification[23]. Deekshatulu and Chandra. P in 2013 has also used Associative classification and genetic algorithms for predicting heart disease[24]. Another work of[25] used KNN and genetic algorithms for the classification of heart disease. Bahassine et al., in 2020 developed a classification algorithm for predicting heart disease and for feature selection using a GA.

Gene expression datasets offer vast amounts of information for analyzing biological processes; however, the presence of redundant and irrelevant genes makes it challenging to identify key genes in high-dimensional data. To overcome this issue, several feature selection (FS) methods have been introduced. Enhancing the efficiency and accuracy of FS techniques is crucial for extracting significant genes from complex multidimensional biological data[46,47]. The authors created a new method called CSSMO that combines two algorithms to select a smaller set of important genes from the data. Using fewer genes allows machine learning algorithms to work faster. The selected genes should be informative for cancer prediction. The authors tested their method on eight datasets and found it to be more accurate than other existing methods. Overall, this research introduces a new technique for selecting genes that might be useful for building more accurate and efficient models for early cancer prediction[48]. This paper specifically looks at how machine learning algorithms are used with medical data to categorize different cancers and even predict their outcomes. The paper dives into supervised, unsupervised, and reinforcement learning, explaining their strengths and weaknesses. By accurately classifying cancers, predicting patient outcomes, and identifying potential treatment targets, machine learning has the potential

to revolutionize cancer diagnosis and treatment. This review equips readers with the latest advancements in machine learning for cancer classification, allowing them to decide on its use in clinical settings. The paper also discusses the potential for even more powerful machine learning systems in the future. Based on the background study a model is proposed for predicting heart disease.

Materials and methods

Proposed methods

The proposed model gives a detailed idea of the model and its behaviour. In this paper, the model proposed consists of gathering the dataset, preprocessing the dataset, applying optimization techniques, and evaluating the accuracy of the classification algorithms. The study utilizes the Cleveland dataset, which comprises 13 features and a single target class. The framework for the proposed model is given in Fig. 2. The dataset is preprocessed using the SelectKBest feature and election is applied where overall rank is obtained with the accuracies. Then the dataset will be given to optimization techniques namely GAO, PSO, ACO. These techniques are applied to the dataset and a subset of features are selected ignoring the unrelated features. The subset of features is then used by classification algorithms. The performance metrics are evaluated from which RF performed better before optimization. Therefore, RF along with a subset of features selected by each optimization is evaluated and accuracy is found. The higher accuracy model is selected and proved to be the best optimization technique for the selected dataset.

Dataset preparation

The dataset is used for training ML model and metaheuristic algorithms the dataset is much needed. The dataset is available publically in repositories like UCI and Kaggle. The Cleveland dataset which has 303 records[10] shown in Table 2. The Cleveland dataset comprises 13 independent variables and one target feature with 2 classes namely heart disease and no heart disease.

This dataset is considered the standard dataset for research purposes and it has very minimum missing values.

Feature selection-ethical consideration

The initial step is finding the accuracy using SelectKBest feature selection which is obtained by applying the Chi-square, Mutual information, and F-statistics. The chi-square is used for categorical data. In statistics, the chi-square test is utilized for testing the event's independence. The formula is given in Eq. (1)[27]

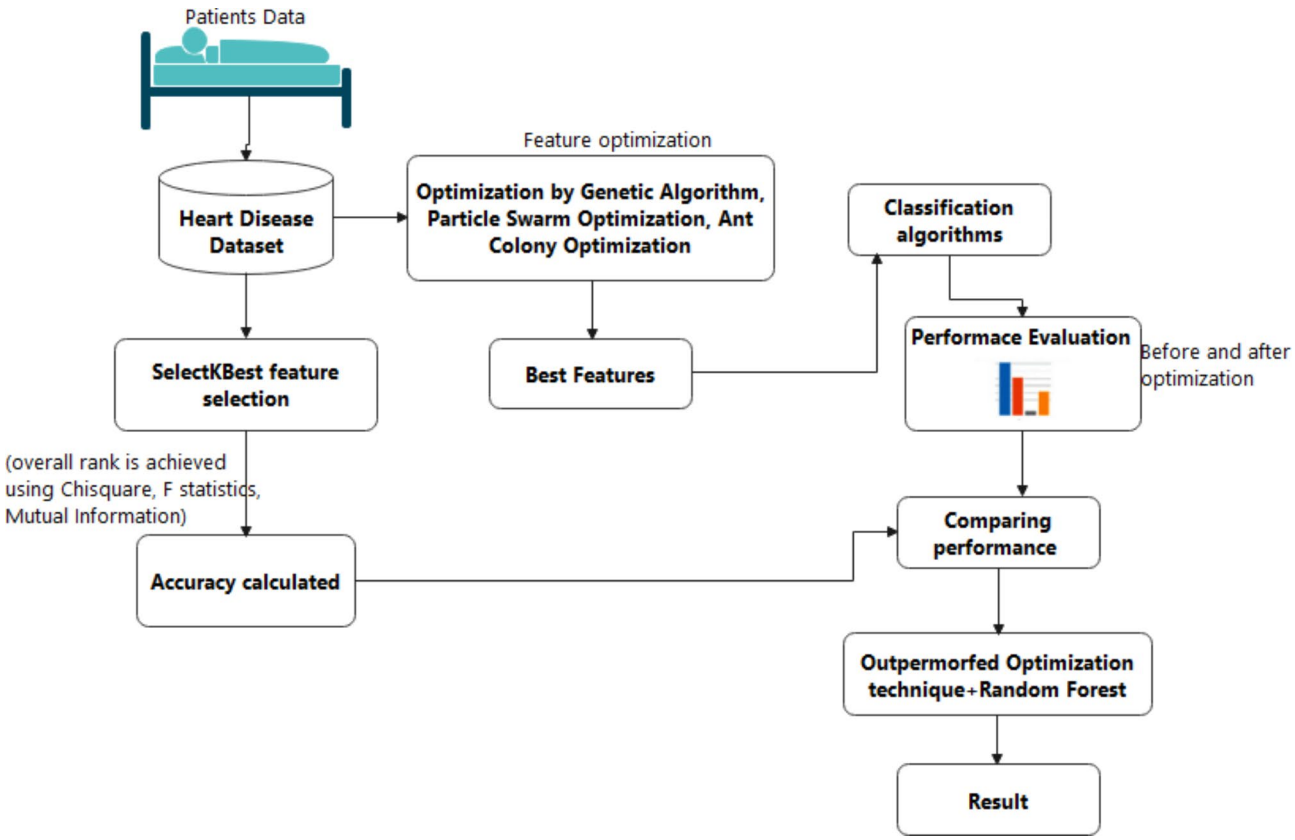

Fig. 2. Workflow of the proposed model. The proposed model follows a structured process, starting with the dataset obtained from the UCI repository, which undergoes preprocessing. The refined dataset is then processed through GAO, PSO, and ACO algorithms for feature optimization. The optimized features are subsequently utilized by machine learning techniques, and the model's performance is assessed.

Attribute	Description	Type	Value Range
Age in years	Continuous value	Numeric	29 < age > 77
Sex	1 represents male 0 represents female	Nominal	1-male 0-female
CP-Chest pain	The chest pain type is categorized as follows: 0 represents typical angina, 1 indicates atypical angina, 2 corresponds to non-anginal pain, and 3 signifies asymptomatic cases.	Nominal	4-asymptomatic 3-non anginal pain 2-atypical angina 1-typical angina
trestbps	Resting blood pressure	Numeric	> 160: very high 140–160: High 120–140: Unusual 90–120: Normal
chol	Serum cholesterol in mg/dl	Numeric	> 250: very high 110–200: Normal 200–240: borderline high 240–250: High
Fasting Blood sugar	Fasting blood sugar > 120 mg/dl 1	Nominal	1-true 0-false
Restecg—resting electrocardiographic (ECG)	resting electrocardiographic results are classified as follows: 0 indicates a normal result, 1 signifies the presence of ST-T wave abnormalities, and 2 suggests possible or definite left ventricular hypertrophy	Nominal	0-normal 1-STTwave abnormality 2-showing probable
thalach	heart rate achieved	Numeric	60-100-normal > 100- tachycardia
Exangexercise-induced angina	Blood supply when you exercise 1 represent yes 0 represents no	Nominal	1 0
Old peak	ST depression induced by exercise relative to rest	Numeric	0-6.2
Sl: slope	The slope of the peak exercise ST segment 0 represents upsloping 1 represents flat 2 represents downsloping	Nominal	1-upsloping 2-flat 3-downsloping
Ca	number of vessels (0–3) coloured by fluoroscopy	Nominal	0, 1, 2, 3
thal	Thallium stress test 1 represents normal; 2 represents fixed defect; 3 represents a reversible defect	Nominal	3-normal 6-fixed 7-reversible
target	The predicted attribute 0 represents no chances of heart failure 1 represents chances of heart failure	Class variable	0-no heart disease 1-heart disease

Table 2. Attributes of the heart dataset from uci with feature details [38].

$$\Sigma \chi_c^2 = \frac{(O_i - E_i)^2}{E_i} \tag{1}$$

Where O is observed cases, E is the expected values, χ_c^2 is Chi-square value Mutual information is the measurement of two random variables that calculates the information obtained about one variable through another random variable. The probability distribution function (p.d.f) of mutual information p (x, y) is calculated using Eq. (2)[27]. These values are integral parts of the feature but mentioning them here is to justify understanding that data for selected features fits into the distribution.

$$I(X;Y) = \int_X \int_Y p(x,y) \log \frac{p(x,y)}{p(x)\,p(y)} dxdy \tag{2}$$

Where p (x, y) denotes the joint probability density function, the marginal density function is given by p(x) and p(y). The F-statistics is the value obtained after conducting the ANOVA test which is the mean of two populations. The F-statistics are given in Eqs. (3) (4) (5)[28,29].

$$\mathrm{F} = \frac{Sum\ of\ squares\ between\ groups/df1}{sum\ of\ squares\ within\ groups/df2} \tag{3}$$

This study utilizes the SelectKBest method, which picks the top K features based on their scores calculated using three different functions: Chi-squared: Analyzes the dependence between a feature and the target variable. Higher scores indicate stronger relationships. F-statistic: Compares the variance between groups defined by the target variable. Higher scores suggest features that better differentiate groups. Mutual information: Measures the shared information between a feature and the target variable[52]. Higher scores imply features that provide more information about the target. Each feature receives a rank based on its score for each method. The top-ranked feature gets a rank of 1, and so on. These individual ranks are then combined to generate an overall rank for each feature, indicating its collective importance across all scoring methods. The results of feature selection are shown in Tables 3, 4 and 5 respectively. These tables explain the features alongside their corresponding scores, aiding in understanding which features contribute most to accurate heart disease prediction.

Measure	Formula	Description
Precision	$\frac{TP}{TP+FP}$	The ratio of correctly predicted positive cases to the total number of actual positive cases
Recall	$\frac{TP}{TP+FN}$	The proportion of correctly identified positive cases relative to the total sample size
Specificity	$\frac{TN}{TN+FP}$	The number of correctly classified negative cases out of the total non-disease samples
Accuracy	$\frac{TP+TN}{TP+FP+FN+TN}$	The overall accuracy, representing the number of correct predictions out of all predictions made.
F-score	$2^{*}(\frac{Precision*Recall}{Precision+Recall})$	The F-score is primarily utilized for addressing classification challenges, particularly when dealing with imbalanced class distributions.

Table 3. Different evaluation criteria for assessing performance[37].

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Specificity (%)	Classification error rate (%)
LR	85.25	86	88	87	81	14.75
KNN	68.85	70	76	73	59	31.15
SVM	70.49	70	82	76	55	29.51
NB	85.25	84	91	87	78	14.75
RF with Grid Search	86.89	88	88	88	81	13.11
RF	90.16	89	94	91	85	9.84
XGB	81.97	83	85	84	77	18.03
DT	83.61	88	82	85	85	16.39
NN	83.61	80	94	86	43	16.39

Table 4. Performance analysis for cleveland dataset before feature selection.

Cleveland dataset features	chi-square	MI	F-statistics	Total	overall rank
Age in years	4.501	0.001	3.452	7.954	11
Sex	1.465	2.802	5.525	9.792	9
CP-Chest pain	12.1	16.351	14.945	43.396	3
trestbps	2.866	4.232	1.383	8.481	10
chol	4.627	6.612	0.472	11.711	8
Fasting Blood sugar	0.039	1.315	0.051	1.405	13
Restecg—resting electrocardiographic	0.576	2.344	1.237	4.157	12
thalach	36.403	7.365	13.949	57.717	1
Exangexercise-induced angina	7.522	8.872	15.198	31.592	5
Old peak	14.042	15.663	14.684	44.389	2
Sl: slope	1.895	7.99	8.761	18.646	7
Ca	12.843	11.819	11.687	36.349	4
thal	1.12	14.365	8.655	24.14	6

Table 5. Cleveland dataset feature score table.

Optimization algorithms

Genetic algorithm (GAO)

GA is mainly used for selecting the feature subset from the given dataset. This method consisted of terms like selection, chromosome, genes, individual, nutation, crossover, and fitness function. This method was developed by Goldberg in 1989[30] and the method is used to generate a population of fixed size called search space. The entities from the population are selected randomly, combined, and mutated, later they can be ignored are selected based on their fitness levels. To create the next level generation, the selection operation is performed. A combination of genes is called a chromosome. DNA blocks are genes. The population is the individuals with the same chromosome length, and the value assigned to them is fitness[32]. The flow of GAO is given in Fig. 3. The fitness function is denoted by Eqs. (4),

$$\mathrm{f}(\mathrm{x}) = \frac{1}{1+g(x)} \tag{4}$$

the objective function given by g(x) and the fitness function given by f(x).

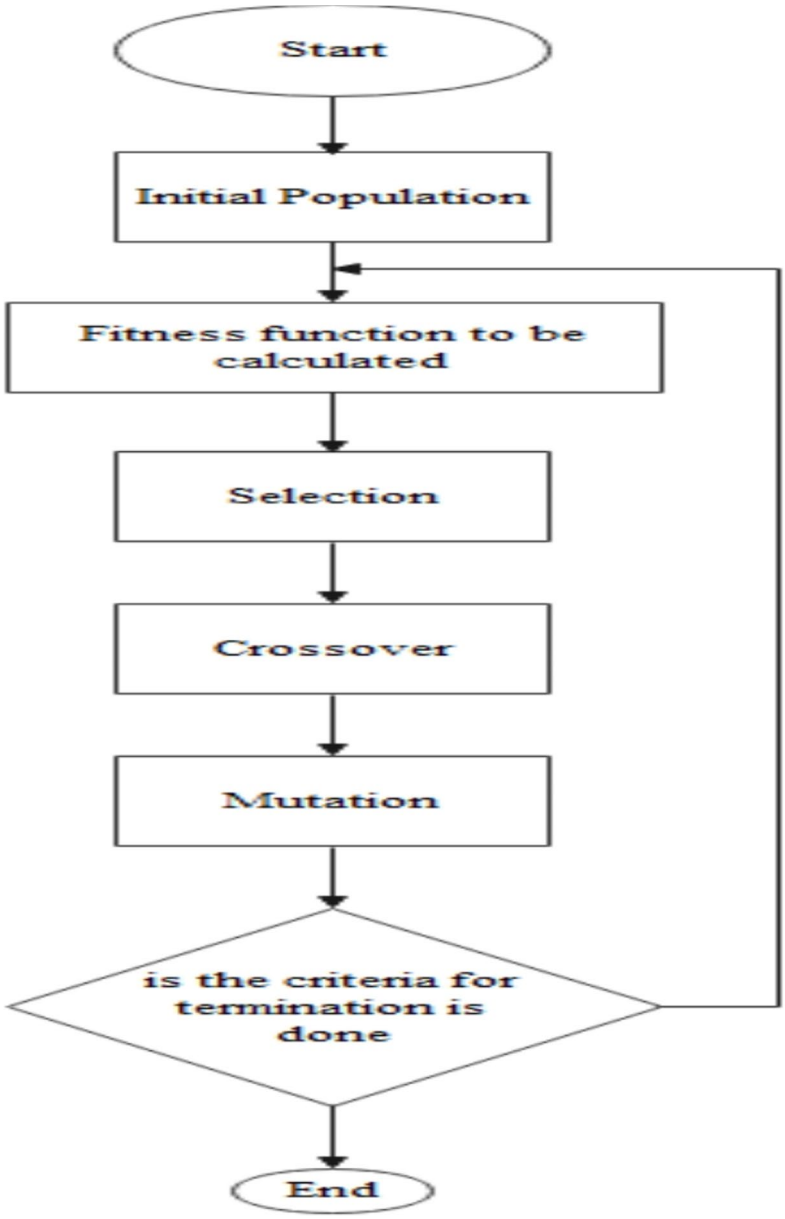

Fig. 3. Genetic algorithm flowchart[31].

The value assigned to the individual is carried by fitness function f(x). Genes from parents combine to form new chromosomes in a crossover, randomly changing the gene is a mutation, and creating the next generation is selection. Algorithm 1 presents the pseudocode for GAO.

Input: Dataset from UCI repository
Output: Optimized Features
initialize t to 0
initpopulation Pop(p)
evaluate Pop(p)
while not done do
    p=p+1
    S1=SelectParents S(p)
    Recombine S1 (p)
    mutate S1 (p)
    evaluate S1 (p)
    S = survive S, S1(p)
end while
end GAO
Input: optimized features R, sample size r, next node n
Output: RF
    Generate N bootstrap samples from the optimized dataset.
    For each node, randomly select a subset of m features, where m < M (total features).
    Construct a split using the m selected features and determine the optimal split point to identify the k node.
    Continue splitting the tree iteratively until a leaf node is reached, ensuring the tree is fully grown.
    Train the algorithm independently on each bootstrapped sample.
    Collect predicted data from n trained trees using a voting mechanism in tree classification.
    Develop the final Random Forest (RF) model based on the features with the highest votes.
    Return the RF model's accuracy as the final output
end

Algorithm 1. Genetic algorithm optimization–Random Forest (GAORF).

Ant colony optimization (ACO)

This section explains the feature selection procedure carried out by Ant Colony Optimization. The flowchart of the ACO is shown in Fig. 4. In this model, the features of the dataset are independent nodes. Based on the probability of the selection $p_k(i)$, features are selected[33]. Eq. (5) gives the probability of selection.

$$P_k(i) = \frac{[\tau(i)]^{\alpha}\,[\eta(i)]^{\beta}}{\sum_{l \in N_{i\,to\,k}} [\tau(l)]^{\alpha}\,[\eta(i)]^{\beta}} \tag{5}$$

Where $\eta(i)$ is feature frequency and it denotes the number of features in the training and it also denotes the heuristic information that the ants have. The feasible neighbourhood of ants is denoted by $N_{i\,to\,k}$. The feature i pheromone trail value is mentioned by $\tau(i)$. ACO[34] parameters are initialized as α and β. The trial of pheromone is updated based on the global update rule and mentioned in Eq. (6).

$$\tau(i) = \rho\,\tau(i) + \sum_{k=1}^{n} \Delta\tau(i)(i) \tag{6}$$

ρ denotes the evaporation parameter that decays the pheromone trail and the number of ants is denoted by η. The proposed ACO works for feature selection. Features have similar probability selection $P_k(i)$. The roulette wheel selection algorithm[35] is used for selecting n features.

This method is more suitable for feature selection problems, as this method has no heuristic that guides the optimal search all the time. Considering the graph, ants will identify the best combinations of features[11]. The pseudocode for the ACO is mentioned in Algorithm 2.

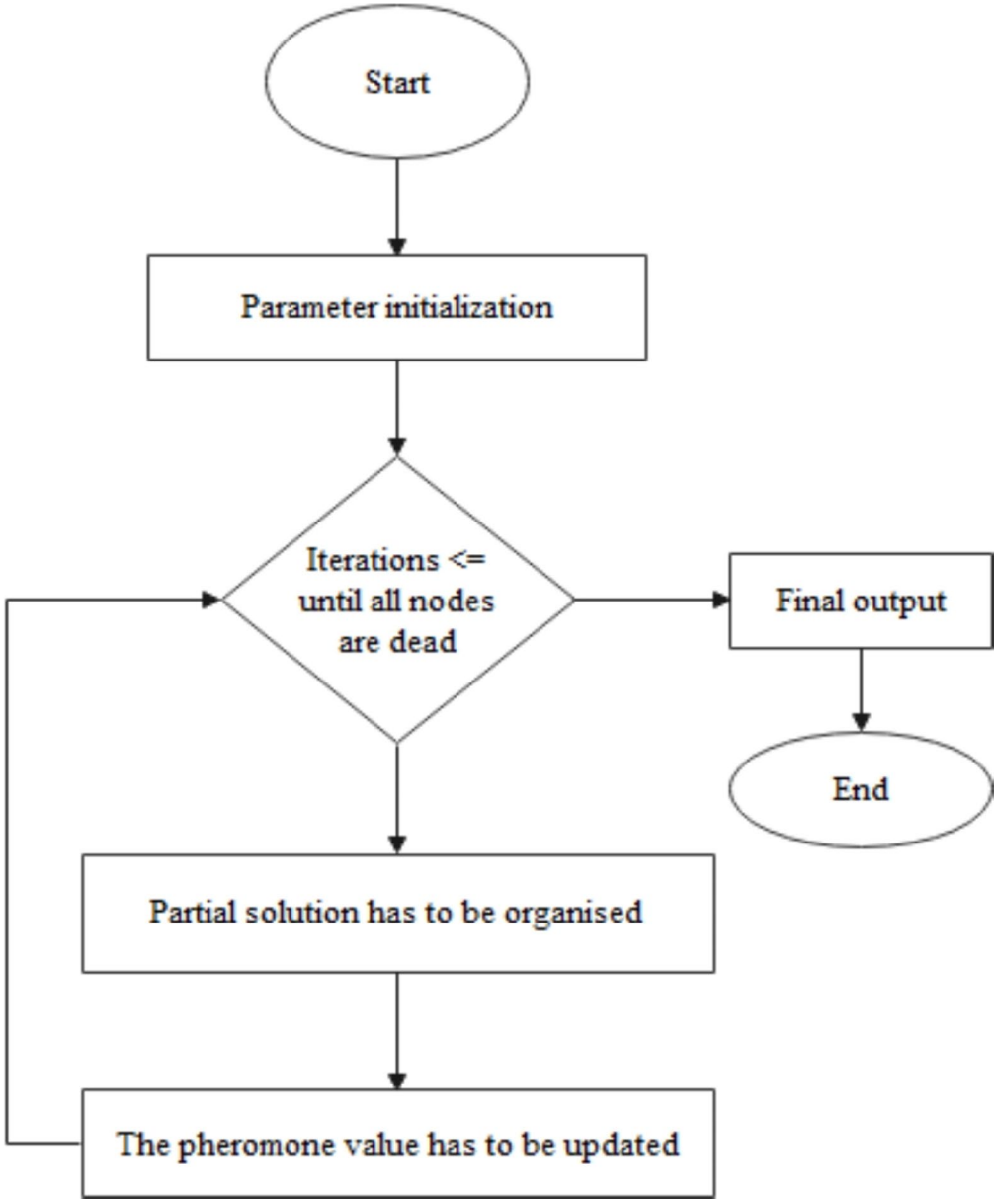

Fig. 4. Ant colony optimization flowchart[11].

Input: Instance z ϵ J of $\prod_{opt}$
Set algorithm parameters ()
n, m=0
for m=1 to colonies do
Output: ST_{test} "candidate" to be the best-found solution z ϵ J
Ant ST_0 = create sub colony and release agent
while not-termination conditions
On sub-colony do
n=n+1
Man_ants' activity ()
Man_Pheromone ()
Man_Demon Action ()
Selection Procedure ()
Compute solution Quality ()
end while
m=m+1
ST_{test}= optimal solution candidate
Update pheromone on arc ()
end for
end ACO
Input: M optimized features, m sample size, k next node
Output: RF
Generate N bootstrap samples from the optimized dataset.
For each node, randomly select a subset of m features, where u < U (total features).
Construct a split using the m selected features and determine the optimal split point to identify the k node.
Continue splitting the tree iteratively until a leaf node is reached, ensuring the tree is fully grown.
Train the algorithm independently on each bootstrapped sample.
Collect predicted data from n trained trees using a voting mechanism in tree classification.
Develop the final Random Forest (RF) model based on the features with the highest votes.
Return the RF model's accuracy as the final output
end

Algorithm 2. Pseudocode for Ant Colony Optimization—(ACORF).

Particle swarm optimization (PSO)

Solving the complex problem by interacting between simple agents and their environment is carried out by particle swarm optimization. Russel 1995 developed this algorithm from the inspiration of a metaheuristic setup. Each article keeps moving at every iteration.

The PSO has two equations, the velocity update and particle position. Based on dimensions the fitness value is calculated[37]. The updated equation and position are given below in Eq. (7), and (8).

$$v_{id}^{t+1} = w_i v_{id}^t + c_1\, rand\,()\ (p_{id}^t - x_{id}^t) + c_2\, rand\,()\ (p_{nd}^t - x_{id}^t) \tag{7}$$

$$x_{id}^{t+1} = x_{id}^t + v_{id}^{t+1} \tag{8}$$

Here w is inertia weight, $c_1\ and\ c_2$ are acceleration constants, a random function denoted by rand (). The personal best at *t*th iteration is given by Eq. (9)

$$P_i^t = [P_{i1}^t, P_{i2}^t, \ldots P_{id}^t, \ldots P_{im}^t] \tag{9}$$

and local best is given by Eq. (10)

$$P_n^t = [P_{n1}^t, P_{n2}^t, \ldots P_{nd}^t, \ldots P_{nm}^t] \tag{10}$$

The one particle close to the optimum, informs its position to get modified. The most effective is the extension of combinatorial optimization. The main code of this algorithm is to move the particle to find the optimum[36]. The pseudocode for PSO is given in Algorithm 3[37].

Input: UCI Dataset
Output: Features Optimized
For each particle
Initialize position and velocity randomly
end
c=1
do
For every particle
compute fitness function
If fitness value > qBest Then
qBest will be the current fitness value
end
update particle with rBest as best fitness value
for every particle
compute new velocity using $Velo_{id}$ (c) = w ∗$Velo_{id}$ (c – 1) + ca_1 ∗rand () ∗ ($qBest_{id}$ − X_{id} (c – 1)) + ca_2 ∗rand () ∗ ($rBest_{id}$ − X_{id} (c – 1)).
update position using $X^d{}_i$ (c) = $X^d{}_i$ (c – 1) + $Velo^d{}_i$ (c)
end
c=c+1
while (c < max iterations)
process the result
end PSO
Input: optimized features M, sample size m, next node d
Output: RF
Generate N bootstrap samples from the optimized dataset.
For each node, randomly select a subset of m features, where m < M (total features).
Construct a split using the m selected features and determine the optimal split point to identify the k node.
Continue splitting the tree iteratively until a leaf node is reached, ensuring the tree is fully grown.
Train the algorithm independently on each bootstrapped sample.
Collect predicted data from n trained trees using a voting mechanism in tree classification.
Develop the final Random Forest (RF) model based on the features with the highest votes.
Return the RF model's accuracy as the final output
end

Algorithm 3. Pseudocode for Particle Swarm Optimization—(PSORF).

Experiment results and analysis

Performance measure

The effectiveness of the proposed method is evaluated based on the selected features and the accuracy of classification. The performance is assessed using key metrics, including accuracy, sensitivity, precision, and recall. The formulas used for calculating specificity, F-score, precision, recall, and accuracy are given in Table 3.

Design setup for experiment

The proposed model is evaluated using optimization techniques and machine learning models. The process begins with data collection, followed by preprocessing to manage any missing values. Next, the dataset undergoes optimization, and the selected features are utilized for classifying heart disease patients using a machine learning model. Various evaluation metrics are applied to assess performance. The experiment is conducted on Anaconda Jupyter Notebook 6.4.8, which includes built-in machine learning packages, and runs on an Intel(R) Core (TM) i7-7600U CPU @ 2.80GHz–2.90GHz.

Dataset	classifier	Accuracy using all features	Accuracy of selected features using filter methods
Cleveland	RF	90.16	89(12)
	RFGS	86	87(8)
	NB	85.25	87(4)
	LR	85.25	90.16(4)
	XGB	81.97	87(6)

Table 6. Accuracy of the heart disease dataset after feature selection along with the selected features

Metaheuristic technique	Number of features selected	Subset of features
Genetic Algorithm	10	Age, sex, cp., trestbps, chol, fbs, thalach, oldpeak, slope, thal
Particle Swarm Optimization	10	sex, cp., trestbps, chol, fbs, restecg, thalach, oldpeak, exang, thal
Ant Colony Optimization	9	Age, cp., trestbps, chol, fbs, restecg, thalach, exang thal

Table 7. List of subsets of features selected based on metaheuristic techniques.

Feature selection results

This work investigates feature selection to optimize the accuracy of ML algorithms for heart disease prediction. The authors employed the SelectKBest method with three filter methods: Chi-square, Mutual Information, and F-statistics. Each feature's rank was determined by each filter, and an overall rank was calculated by combining these individual ranks mentioned in Table 4. The lowest-ranked features were removed, and the remaining features were fed into the classification algorithms. Using all features, Random Forest achieved the highest accuracy of 90.16% from Table 4. After applying feature selection, Random Forest experienced a slight accuracy drop. However, Logistic Regression performance improved, reaching 90.16% accuracy Table 6.

This section explores three metaheuristic algorithms GA ACO, PSO for feature selection in predicting heart disease.These algorithms aim to identify a subset of important features from the original dataset, potentially enhancing the accuracy of the prediction model. All three algorithms successfully reduced the initial 13 features to a smaller set. Genetic Algorithm and Particle Swarm Optimization selected 10 features each. Ant Colony Optimization selected 9 features shown in Table 7. This study will compare the performance of these optimized feature sets with existing models utilizing all 13 features. Each optimized feature set will be fed into a Random Forest classifier to assess its impact on prediction accuracy. The results will be presented, comparing the accuracy achieved by different feature selection methods and traditional approaches using all features. In medical predictions, accuracy is paramount. By identifying the most relevant features, these algorithms have the potential to improve the model's ability to correctly diagnose heart disease, potentially leading to better patient outcomes.

The parameters considered for GAO are.
Mutation Rate: Varied (1000, 100, 5).
Crossover Rate: Implicitly 100%.
Crossover Operator: Single-Point Crossover.
Number of Generations: 50 and 100 in different runs.
Population Size: 20.
Stopping Criteria: Number of generations.
Selection Operator: Fitness-proportional selection from a randomly chosen subset of 5 individuals.
Selection Pressure: Implicit in the selection operator.
Fitness Function: Based on the inverse of the root mean squared error from a Random Forest Regressor.
The parameters considered for PSO are.
Features = 13.
C1 = 0.5 (Cognitive factor influencing individual particle movement).
C2 = 0.5 (Social factor guiding particle interaction within the swarm).
W = 0.9 (Inertia weight controlling the influence of previous velocity).
K = 30 (Number of top-performing particles or neighboring particles considered).
$P = 2$ (Parameter defining the search space dynamics).
Iterations = 1000.
The parameters considered for ACO are.
iteration = 100.
n_ants = 5 (Total number of ants in the system).
n_citys = 5 (Total number of cities or nodes to be visited).
e = 0.5 (Evaporation rate, determining pheromone decay over time).
alpha = 1 (pheromone influence factor, controlling the impact of pheromone trails).
beta = 2 (visibility factor, emphasizing the importance of heuristic information in decision-making).

Based on the parameters used by each optimization algorithm, the features are selected and are listed in Table 7.

Performance metrics	RF	GAORF	PSORF	ACORF
Recall	93	93.1	94	91.18
Precision	88.88	93.1	91.43	86.11
Accuracy	90.16	92	91	86.88
Specificity	85.18	91.59	88.89	81.48
AUC	90	94	93	92

Table 8. The result of performance metrics for different optimization techniques along with classification algorithm in terms of percentage (%).

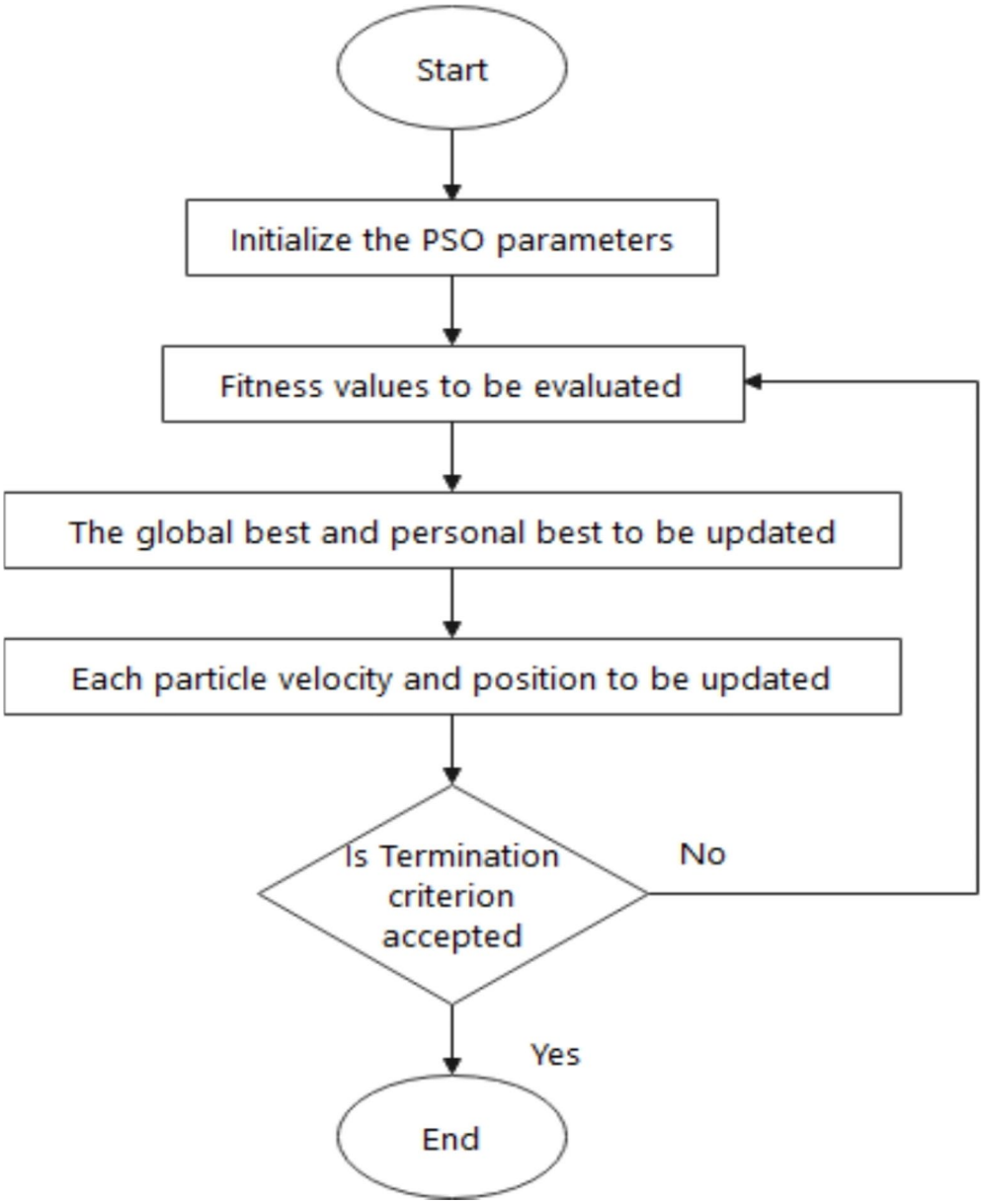

Fig. 5. Particle swarm optimization flowchart[11].

RF algorithm is performed for the features selected by GA, ACO, PSO. Table 8 records the performance metrics achieved after the feature selection by metaheuristics. Random forest with Genetic algorithm achieves 92% accuracy, precision, and recall with 94.1%, specificity of 92.59%, and Area under the curve of 95%.

Random forest with PSO achieves 91.8% accuracy, 94% recall 91.43% precision, specificity 88.89%, and Area under curve 93%. Random forest with ACO achieves only 86.88% accuracy, 91.18% recall, 86.11% precision,

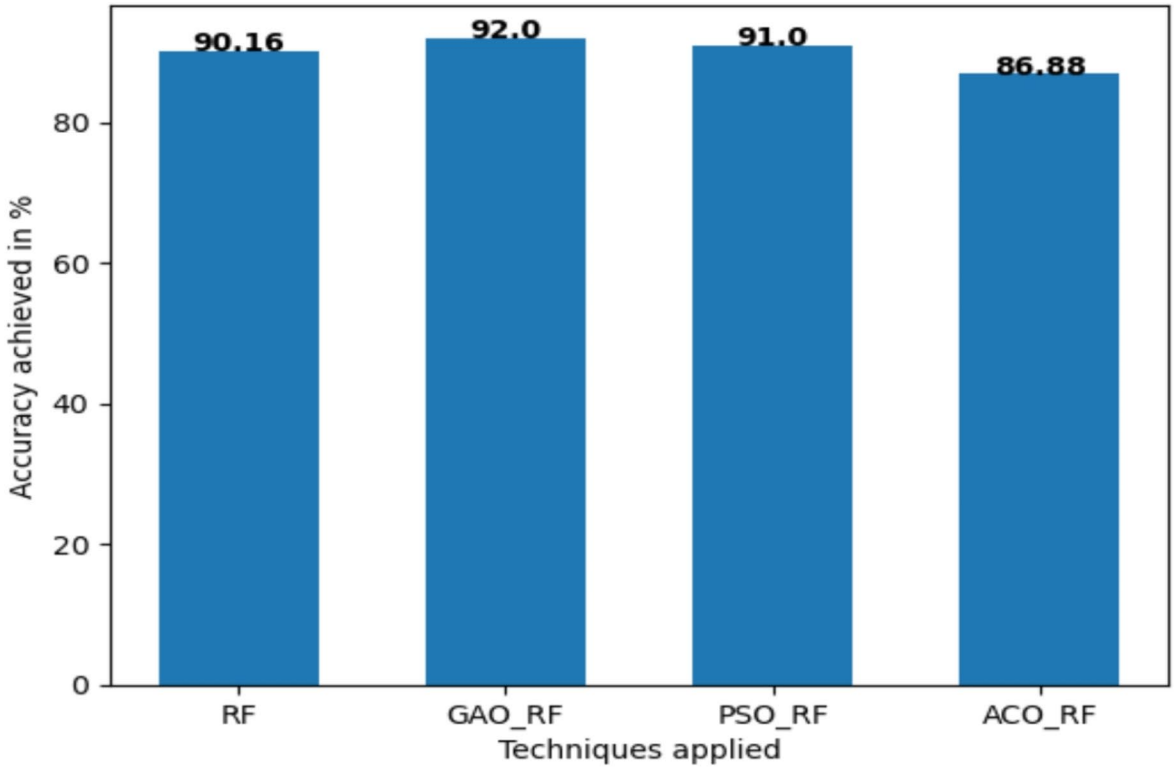

Fig. 6. Comparing the accuracy of the optimization techniques along with RF for the Cleveland dataset.

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)	AUC (%)
Logistic Regression	85.25	86	88	81	91
KNN	68.85	70	76	59	73
SVM	70.49	70	82	55	81
NB	85.25	84	91	78	91
RF with Grid Search	86.89	88	88	81	90
RF	90.16	88.88	93	85.18	90
XGB	81.97	83	85	77	91
DT	83.61	88	82	85	84
GAORF	92	93.1	93.11	91.59	94

Table 9. Comparing the performance of GAORF along with different classification algorithms.

and specificity with 81.48% and 92% area under the curve. From Table 8, it is clear that GAORF performs better compared with other techniques and achieved 92%. The same is visualized with the help of a bar graph shown in Fig. 5.

GAORF Achieves Highest Accuracy in Heart Disease Prediction. Accuracy is a key metric in evaluating the effectiveness of prediction models. In this study, Fig. 6 visualizes the accuracy achieved by different algorithms applied to the Cleveland dataset. The proposed Genetic Algorithm-optimized Random Forest (GAORF) model outperformed all other techniques with a remarkable 92% accuracy. This result underscores the efficiency of GAORF in both feature selection and classification for heart disease prediction. Table 9 summarizes the performance of different machine learning algorithms. Random Forest (RF) alone achieved 90.16% accuracy, demonstrating its strong potential. Other methods like LR (85.25%), KNN (68.85%), SVM (70.49%), and NB (85.25%) displayed lower accuracy. Notably, Grid Search optimization slightly improved RF accuracy to 86.89%, but remained below GAORF.

The different performance metrics for the proposed model GAORF are compared with varied Machine learning classification techniques. The performance is visualized with the help of Table 9. Figure 7 explains the accuracy variation between features selected using filter methods and optimization techniques. The line graph clearly explains that optimization techniques perform better compared with filter methods. Among the

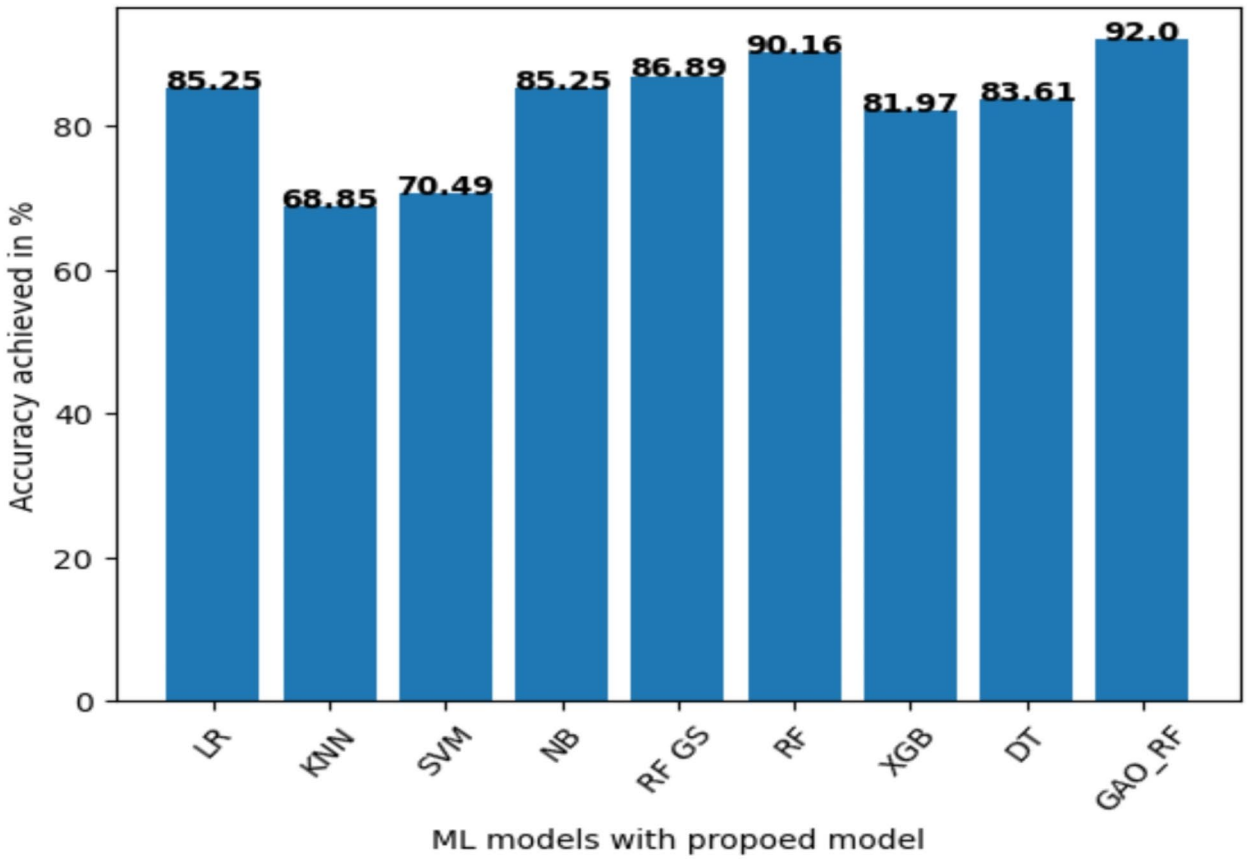

Fig. 7. Comparing the accuracy of the proposed model along with the ML classification algorithm for the Cleveland dataset.

Dataset	classifier	Accuracy using all features	Accuracy of selected features using the filter method
Cleveland	RF	90.16	89(12)
	RFGS	86	87(8)
	NB	85.25	87(4)
	LR	85.25	90.16(4)
	XGB	81.97	87(6)
	GAORF	92	
	PSORF	91	
	ACORF	86.11	

Table 10. Comparing the accuracy of GAO + RF along with different classification algorithms before and after feature selection.

Technique	Strengths	Limitations
GAORF	Handles non-linear complex problems	Can be computationally expensive. Tuning hyperparameters for GA can be challenging. May not be easily interpretable
PSORF	Relatively easy to implement. Well-suited for continuous optimization problems	May struggle with high-dimensional datasets (many features). Can get stuck in local optima (suboptimal solutions)
ACORF	Inspired by real-world ant behaviour for efficient searching	Performance can be sensitive to parameter settings. May converge slowly on complex problems
SelectKBest Feature Selection	Reduces computational cost and training time	May discard potentially useful features. Doesn't consider feature interactions

Table 11. Proposed model strength and limitations.

Models	Accuracy
Proposed Model	92%
BABC–K-NN produced[16]	86.76%
GA + SVM[25]	87%
ACO + SVM[20]	87.77%
Best F measure[26]	90.50%
ACO[21]	85.21%
Modified Bee algorithm + RF[22]	85.20%
92HGDGWO[39]	80%
NB + FS[40]	84%
Fisher score algorithm + SVM[41]	81.91%
Hybrid (NB, BN, RF, and MP)[42]	85.45%
Modified GWO + LR[43]	86.91%
Modified GWO + KNN[43]	87.46%

Table 12. Proposed model vs existing models

optimization techniques, GA achieved an accuracy of 92% in predicting heart disease. The performance metrics values are mentioned in Table 10.

Discussion

While the proposed approach using GAORF achieved promising results for heart disease prediction, it's important to acknowledge the inherent limitations associated with machine learning and metaheuristic optimization techniques. The following Table 11 highlights some key limitations of the methods explored in this paper.

This study seeks to enhance heart disease prediction accuracy by integrating feature selection methods with machine learning algorithms. The work consists of 2 phases. Phase 1 is Feature Selection. The SelectKBest method with three filter methods (Chi-square, Mutual Information, and F-statistics) is employed to identify the most relevant features from the dataset. This reduced the initial feature set while potentially maintaining or even improving prediction accuracy. Initially, Random Forest performed best with 90.16% accuracy using all features. However, after applying feature selection, Logistic Regression showed better performance (90.16%). Phase 2 is applying Metaheuristic Optimization for Feature Selection. Observing the impact of feature selection on different algorithms, we explored metaheuristic techniques for further optimization. The study utilized GAO, PSO, and ACO to select even more impactful feature subsets. Each optimized feature set was then fed into a Random Forest classifier to assess its effectiveness. The combination of GAO and Random Forest (GAORF) achieved the highest accuracy of 92%, surpassing all other techniques. This demonstrates the effectiveness of GAORF in selecting optimal features for heart disease prediction. By combining metaheuristic optimization with machine learning, we were able to significantly improve prediction accuracy compared to using all features or basic feature selection methods. The existing model with the proposed model comparison is shown in Table 12.

Conclusion

This research proposes a hybrid approach for predicting heart disease, focusing on maximizing accuracy and surpassing existing methods. An efficient machine learning model along with metaheuristic optimization techniques is applied to the heart disease dataset to enhance the accuracy in predicting the disease. The model is applied to the Cleveland Heart dataset. The optimization techniques namely GAO, PSO, and ACO are compared with the classification algorithms of machine learning. The model performed before and after optimization. Also, the model is compared with SelectKBest filter methods namely Chi-square, Mutual information, and F-statistics. The accuracy achieved by selecting features based on the overall rank of features is analyzed with the metaheuristic techniques. The output of the experiment explains that the GAORF performed better for the dataset considered. The proposed model achieved higher accuracy which is the maximum among other classification techniques. This study is limited to heart disease and its dataset. Future work can be done with other metaheuristic techniques and also different datasets available for predicting heart disease. The same methods can be applied to diabetes prediction and cancer prediction at the initial level.

Data availability

The datasets generated and/or analyzed during the current study are available in the UCI repository and can be accessed from (https://archive.ics.uci.edu/dataset/45/heart+disease).

Received: 7 February 2024; Accepted: 23 September 2024
Published online: 31 March 2025

References

1. Richens, J. G., Lee, C. M. & Johri, S. Improving the accuracy of medical diagnosis with causal machine learning. *Nat. Commun.*. https://doi.org/10.1038/s41467-020-17419-7 (2020).
2. Hameed, B. Z. et al. Engineering and clinical use of artificial intelligence (AI) with machine learning and data science advancements: Radiology leading the way for future. *Ther. Adv. Urol.***13**, 17562872211044880 (2021).
3. Al Bataineh, A. & Jarrah, A. High-performance implementation of neural network learning using swarm optimization algorithms for EEG classification based on brain wave data. *Int. J. Appl. Metaheuristic Comput. (IJAMC)*. **13**(1), 1–17 (2022).
4. Yoo, I. et al. Data mining in healthcare and biomedicine: a survey of the literature. *J. Med. Syst.***36**, 2431–2448 (2012).
5. Oikonomou, E. K. et al. A novel machine learning-derived radiotranscriptomic signature of perivascular fat improves cardiac risk prediction using coronary CT angiography. *Eur. Heart J.***40**(43), 3529–3543 (2019).
6. Tsao, C. W. et al. Heart disease and stroke statistics: a report from the American Heart Association. *Circulation***145**, 153–639 (2022).
7. Mendis, S., Puska, P., Norrving, B. E. & World Health Organization. *Global Atlas on Cardiovascular Disease Prevention and Control* (World Health Organization, 2011).
8. Kumar, Y., Koul, A., Singla, R. & Ijaz, M. F. Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework, and future research agenda. *J. Ambient Intell. Humaniz. Comput.*https://doi.org/10.1007/s12652-021-03612-z (2022).
9. Yazdani, A., Varathan, K. D., Chiam, Y. K., Malik, A. W., Ahmad, W. & W. A A novel approach for heart disease prediction using strength scores with significant predictors. *BMC Med. Inf. Decis. Mak.***21**(1), 1–16 (2021).
10. Khourdifi, Y. & Bahaj, M. Heart disease prediction and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization. *Int. J. Intell. Eng. Syst.***12**(1), 242–252 (2019).
11. Gaddala, L. K. & Rao, N. N. An analysis of heart disease prediction using swarm intelligence algorithms. *Int. J. Innov. Eng. Technol.***6**(3). (2018).
12. Dubey, A. K., Sinhal, A. K. & Sharma, R. An improved auto categorical PSO with ML for heart disease prediction. *Eng. Technol. Appl. Sci. Res.***12**(3), 8567–8573 (2022).
13. Saeed, N. A. & Al-Ta'i, Z. T. M. Heart disease prediction system using optimization techniques. In *New Trends in Information and Communications Technology Applications: 4th International Conference, NTICT 2020, Baghdad, Iraq, June 15, 2020, Proceedings 4*, 167–177 (Springer International Publishing, 2020).
14. Mafarja, M. M., Eleyan, D., Jaber, I., Hammouri, A. & Mirjalili, S. Binary dragonfly algorithm for feature selection. In *2017 International Conference on New Trends in Computing Sciences (ICTCS)*, 12–17 (IEEE, 2017).
15. Majeed, N. M. & Ramo, F. M. Implementation of features selection based on dragonfly optimization algorithm. *Technium***4**(10), 44–52 (2022).
16. Subanya, B. & Rajalaxmi, R. R. A novel feature selection algorithm for heart disease classification. *Int. J. Comput. Intell. Inf.***4**(2), 117–124 (2014).
17. Yaqoob, M. M. et al. Modified artificial bee colony based feature optimized federated learning for heart disease diagnosis in healthcare. *Appl. Sci.***12**(23), 12080 (2022).
18. Sabeena, S. & Sarojini, B. Optimal feature subset selection using Ant Colony Optimization. *Indian J. Sci. Technol.***8**(35) (2015).
19. Usman, A. M., Yusof, U. K. & Naim, S. Cuckoo-inspired algorithms for feature selection in heart disease prediction. *Int. J. Adv. Intell. Inf.***4**(2), 95–106 (2018).
20. Dwivedi, R., Kumar, R., Jangam, E. & Kumar, V. An ant colony optimization-based feature selection for data classification. *Int. J. Recent. Technol. Eng.***7**, 35–40 (2019).
21. Asadi, S., Roshan, S. & Kattan, M. W. Random forest swarm optimization-based for heart disease diagnosis. *J. Biomed. Inform.***115**, 103690 (2021).
22. Velswamy, K., Velswamy, R., Swamidason, I. T. J. & Chinnaiyan, S. Classification model for heart disease prediction with feature selection through modified bee algorithm. *Soft. Comput.*, 1–9. (2021).
23. Jabbar, M. A., Deekshatulu, B. L. & Chandra, P. Heart disease prediction using lazy associative classification. In *2013 International Mutli-Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s)*, 40–46 (IEEE, 2013).
24. Deekshatulu, B. L. & Chandra, P. Classification of heart disease using k-nearest neighbour and genetic algorithm. *Procedia Technol.***10**, 85–94 (2013).
25. Kanwal, S., Rashid, J., Nisar, M. W., Kim, J. & Hussain, A. An effective classification algorithm for heart disease prediction with a genetic algorithm for feature selection. In *2021 Mohammad Ali Jinnah University International Conference on Computing (MAJICC)*, 1–6 (IEEE, 2021).
26. Bahassine, S., Madani, A., Al-Sarem, M. & Kissi, M. Feature selection using an improved chi-square for arabic text classification. *J. King Saud Univ.-Comput. Inform. Sci.***32**(2), 225–231 (2020).
27. Bommert, A., Sun, X., Bischl, B., Rahnenführer, J. & Lang, M. The benchmark for filter methods for feature selection in high-dimensional classification data. *Comput. Stat. Data Anal.***143**, 106839 (2020).
28. Van Cauwenberge, L. Top 10 machine learning algorithms. *Data Sci. Cent.*, **6**. (2015).
29. Uma, S. M., Gandhi, K. R. & Kirubakaran, E. A hybrid PSO with dynamic inertia weight and GA approach for discovering classification rules in data mining. *Int. J. Comput. Appl.***40**(17), 32–37 (2012).
30. Albadr, M. A., Tiun, S., Ayob, M. & Al-Dhief, F. Genetic algorithm based on natural selection theory for optimization problems. *Symmetry*. **12**(11), 1758 (2020).
31. Katoch, S., Chauhan, S. S. & Kumar, V. A review of the genetic algorithm: past, present, and future. *Multimedia Tools Appl.***80**, 8091–8126 (2021).
32. Saraç, E. & Özel, S. A. An ant colony optimization-based feature selection for web page classification. *Sci. World J.2014*. (2014).
33. Dorigo, M., Birattari, M. & Stutzle, T. Ant colony optimization. *IEEE Comput. Intell. Mag.***1**(4), 28–39 (2006).
34. Back, T. *Evolutionary Algorithms in Theory and Practice: Evolution Strategies, Evolutionary Programming, Genetic Algorithms* (Oxford University Press, 1996).
35. Pasala, S., Kumar, B. N. & Satapathy, S. C. A study of the roulette wheel and elite selection on ga to solve job shop scheduling. In *Proceedings of the International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA)*, 477–485 (Springer, 2013).
36. Li, J., Sun, Y. & Hou, S. Particle swarm optimization algorithm with multiple phases for solving continuous optimization problems. *Discrete Dyn. Nat. Soc***2021**, 1–13 (2021).
37. Jangle, P. & Narayankar, S. Alternating decision trees for early diagnosis of Heart Disease. *Int. J. Eng. Comput. Sci.***05**(16070), 16070–16072. https://doi.org/10.18535/ijecs/v5i4.02 (2016).
38. UCI Heart Disease Dataset. https://archive.ics.uci.edu/ml/datasets/heart+disease
39. Kitonyi, P. M. & Segera, D. R. Hybrid gradient descent grey wolf optimizer for optimal feature selection. *BioMed Res. Int.***2021**. (2021).
40. Nassif, A. B., Mahdi, O., Nasir, Q., Talib, M. A. & Azzeh, M. Machine learning classifications of coronary artery disease. In *2018 International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, 1–6 (IEEE, 2018).
41. Saqlain, S. M. et al. Fisher score and Matthews correlation coefficient-based feature subset selection for heart disease diagnosis using support vector machines. *Knowl. Inf. Syst.***58**, 139–167 (2019).
42. Latha, C. B. C. & Jeeva, S. C. Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. *Inf. Med. Unlocked*. **16**, 100203 (2019).

43. Mohiddin, S. K. et al. A modified grey wolf optimizer algorithm for feature selection to predict heart diseases. *Ijfans Int. J.***2319** 1775 https://doi.org/10.48047/IJFANS/V11/I12/180 (2023).
44. Victor, A., Gandhi, B. S., Ghalib, M. R. & Jerlin, A. M. A review on skin cancer detection and classification using infrared images. *Int. J. Eng. Trends Technol.***70**(4), 403–417 .
45. Victor, A., Gandhi, B. S. & Ghalib, M. R. Detection and classification of breast cancer using machine learning techniques for Ultrasound images. *Int. J. Eng. Trends Technol.***70**(3), 170–117 (2022).
46. Yaqoob, A., Verma, N. K. & Aziz, R. M. Optimizing gene selection and cancer classification with hybrid sine cosine and cuckoo search algorithm. *J. Med. Syst.***48**(1), 10 (2024).
47. Mahto, R. et al. A novel and innovative cancer classification framework through a consecutive utilization of hybrid feature selection. *BMC Bioinform.***24**(1), 479 (2023).
48. Yaqoob, A., Aziz, M., Verma, N. K. & R., & Applications and techniques of machine learning in cancer classification: a systematic review. *Human-Centric Intell. Syst.***3**(4), 588–615 (2023).
49. Louridi, N., Douzi, S. & El Ouahidi, B. Machine learning-based identification of patients with a cardiovascular defect. *J. Big Data*. **8**, 133. https://doi.org/10.1186/s40537-021-00524-9 (2021).
50. *Math & Stat Reviews: Stat Online*. PennState: Statistics Online Courses. (2024). https://online.stat.psu.edu/statprogram/reviews
51. Iconsdom *Data optimization icon black sign vector image on VectorStock*. VectorStock. (2020). https://www.vectorstock.com/royalty-free-vector/data-optimization-icon-black-sign-vector-35192565
52. Elemam, T. & Elshrkawey, M. A highly discriminative hybrid feature selection algorithm for cancer diagnosis. *Sci. World J.***2022**. https://doi.org/10.1155/2022/1056490 (2022).

Author contributions
Geetha Narasimhan wrote the main manuscript, implementation, prepared figures and tables. Akila Victor reviewed the manuscript.

Funding
Open access funding provided by Vellore Institute of Technology.

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to A.V.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.