A Novel Hybrid Approach for Feature Selection in Cardiovascular Risk Assessment

Ankush Hutke*, Jyoti Deshmukh

Department of Computer Engineering, MCT's Rajiv Gandhi Institute of Technology, University of Mumbai, Maharashtra, India

Received 06 July 2024; received in revised form 18 August 2024; accepted 22 August 2024

DOI: https://doi.org/10.46604/aiti.2024.13972

Abstract

Early detection of cardiac risk is crucial for accurate diagnosis and treatment of fatal cardiovascular diseases. Selecting relevant features is essential for machine learning in building an effective decision support system of cardiovascular risk assessment, ensuring accuracy of high-dimensional data. This study aims to propose a novel hybrid feature selection approach, termed ant colony optimization with hill climbing (ACOHC), integrating ant colony optimization (ACO) and hill climbing (HC) algorithms. The accuracy metric and various classifiers are deployed to evaluate the effectiveness. Additionally, comparisons are made with nine alternative feature selection techniques. The feature subset identified through the ACOHC attains a classification accuracy of 95.1% with the support vector machine classifier.

Keywords: machine learning, support vector machine, ACOHC, hill climbing, ant colony optimization

1. Introduction

Cardiovascular diseases (CVDs) continue to be the leading cause of mortality worldwide, posing significant challenges for healthcare systems and necessitating effective risk assessment and management strategies [1]. Therefore, accurate prediction of cardiovascular risk, which requires the integration and analysis of complex and multifaceted data, is crucial for early intervention and prevention [2]. The high-dimensional and diverse nature of cardiovascular data, which may include clinical, genetic, lifestyle, and environmental aspects, hinders traditional feature selection approaches despite their value [3-4].

Many diverse domains employ machine learning, an expanding topic in computer science, to develop various decision support systems. Practically, grappling with high-dimensional data emerges as a prevalent challenge. This type of data can escalate complexity and compromise system accuracy [5]. Feature selection techniques tackle this problem by eliminating insignificant features and keeping the relevant ones. This reduction improves system accuracy and simplifies its complexity. Additionally, removing redundant and noisy features helps to decrease computation time [6].

Three categories may be used to group feature selection techniques:

(1) Filter methods: These methods determine the importance of each feature apart from the learning process. Statistical measurements or heuristic techniques are pervasively employed to rank, or score features according to their correlation with the target variable. The chi-square test and information gain are examples of common methods.

* Corresponding author. E-mail address: ankush.hutke@mctrgit.ac.in

English language proofreader: Chih-Wei Chang

(2) Wrapper methods: These methods appraise the performance of a subset of features using the predictive power of a particular machine learning algorithm. They entail iterative search procedures, such as backward elimination, forward selection, or recursive feature elimination (RFE), to find the optimal subset of features, thereby providing the optimal performance for the specified model.

(3) Embedded methods: These techniques include feature selection while creating the model. Feature selection occurs internally within the algorithm during training. Examples include decision trees, least absolute shrinkage and selection operator (LASSO), and feature importance scores derived from ensemble models like random forests [7].

Researchers have examined multifarious feature selection strategies and classifiers on different heart disease datasets. Diagnosing diseases using computer-based systems encompasses processing and analyzing high-dimensional and diverse data. Such data can incur model overfitting and prolonged training times. Feature selection, a dimensionality reduction strategy, removes redundant features that do not significantly impact classifier performance, thereby reducing the data to a manageable size. Several effective feature selection methods have been created recently to reduce the negative effects of high dimensionality. The influence of several feature selection techniques is examined in this study. An experimental approach is used, which includes extensive testing on actual cardiac disease-related datasets obtained from the University of California, Irvine (UCI). The study aims to identify the most effective predictive models for forecasting heart disease and aiding the medical community. Feature selection is assessed alongside accuracy, precision, and recall as key performance indicators for the predictive models.

The following parts of the article are structured as Section 2 gives a detailed related work, including work done by various researchers and research gaps. Section 3 delves into the specifics of the proposed methodologies. The results are then shown in Section 4, followed by a comprehensive analysis. Section 5 offers a concise concluding remark of the study with some last reflections.

2. Related Work

This section summarizes the methods utilized for selecting features in the heart disease dataset.

(1) Chi-square algorithm: It is a filter-based feature selection approach, which computes the chi-squared score between each attribute and target class that measures the difference between observed and expected values. In addition, the chi-square algorithm measures the dependency between the categorical input feature and the categorical target variable. Features with high chi-square statistics and low p-values are considered more relevant to the target variable [8]. The chi-square value is calculated for each feature as shown below:

$$x^2 = \sum \frac{(o-e)^2}{e} \tag{1}$$

where o represents the observed value and e denotes the expected value.

(2) Analysis of variance (ANOVA): It is a statistical method used to examine the differences in means among the groups. In feature selection, the ANOVA assesses the relationship between continuous input features and a categorical target variable. Furthermore, the ANOVA calculates the F-value and associated p-value for each feature, indicating the significance of the feature's effect on the target variable [9].

(3) Forward selection algorithm (FSA): It is a wrapper technique that adds features to the feature subset incrementally, one at a time. After assessing the performance of the model with the new feature, it chooses the top-performing feature subset at each stage. This procedure is carried out repeatedly until the performance of the model evinces no further signs of improvement. When selecting a subset of features with a support vector machine (SVM) as the learning algorithm, it is crucial to stratify the data to ensure that each class is adequately represented [10].

(4) Backward elimination algorithm (BEA): Backward elimination is a different wrapping strategy that starts with the entire feature set and removes each feature individually. It chooses the top-performing feature subset at each step, by assessing the performance of the model with the deleted feature. This process continues until further removal of features results in decreased model performance [10].

(5) Mutual Information (MI): The amount of knowledge about one variable learned through the other variable is measured by MI. MI gauges the degree of dependability between the target variable and the input features throughout the feature selection process. To predict the target variable, features rendering high MI values are thought to be more informative [11].

(6) L2 regularization ridge regression (L2): L2 is an embedded method incorporating feature selection within the model training process. The conventional regression objective function is extended with a penalty term (L2 regularization) to reduce the coefficients of less significant characteristics to zero. Features with smaller coefficients are effectively down-weighted, leading to automatic feature selection during model training [12].

(7) Particle swarm optimization (PSO): It is a population-based stochastic optimization method that draws inspiration from fish schools and flocks of birds for their social behaviors. In feature selection, PSO optimizes a population of candidate feature subsets by repeatedly updating the positions of particles in the search space. The goal is to identify the ideal feature subset that minimizes or maximizes some objective function, which tends to be pertinent to the performance of the model [13].

(8) Ant colony optimization (ACO): It is a metaheuristic optimization method that draws inspiration from ants' foraging habits. In feature selection, ACO constructs a graph representation of the feature space, where features are nodes and edges represent the interactions between features. Ants iteratively build solutions by selecting features based on pheromone trails and heuristic information to find an optimal feature subset [13].

(9) Hill climbing (HC) algorithm: It is a local search optimization algorithm that iteratively explores the neighboring solutions within the search space. In feature selection, HC starts with an initial feature subset and iteratively modifies it by evaluating neighboring feature subsets. Subsequently, it moves towards the neighboring solution that improves the objective function (e.g., model performance), continuing until no further improvement is possible [14].

Jabbar et al. [15] employed feature selection using the chi-square method on the Cleveland Heart Disease dataset. The chi-square method is a filter-based feature selection technique that assesses the relationship between the target variable and each feature using the chi-square statistic. Wiharto et al. [16] worked on the Cleveland dataset to employ feature selection methods. Specifically, they utilized the Information Gain criterion to select features. In the paper by Haq et al. [17], feature selection methods were employed on the Cleveland Heart Disease dataset. Three specific techniques are utilized and listed as follows:

(1) Minimal-redundancy-maximal-relevance (MRMR) opts for the features based on the target variable and degree of redundancy. It minimizes duplication among chosen characteristics while taking into account the MI between features and the target variable.

(2) Relief is a method for feature selection that ranks features according to the differentiability between instances of various classes. It iteratively updates feature weights by comparing nearest neighbor instances belonging to the same and different classes.

(3) LASSO is a method of regression analysis that enhances the interpretability and prediction accuracy of statistical models by performing regularization and variable selection. It penalizes the absolute size of the regression coefficients, encouraging sparse solutions where irrelevant features have zero coefficients. These selected features are used to determine a subset of pertinent and useful characteristics for predicting results in heart disease.

Khourdifi and Bahaj [18] utilized the Cleveland dataset to explore feature selection methods. Specifically, they used the feature selection method as quick correlation-based that is enhanced by ant colony and PSO. This approach is likely to be involved in leveraging correlations between features and the target variable (heart disease diagnosis) to efficiently select the most relevant features subset. In the paper by Jain et al. [19], the feature selection method, PSO, is applied to the Cleveland Heart Disease dataset. Initially, PSO is a metaheuristic optimization method that draws inspiration from fish schools and bird flocking behavior. It entails updating a population of potential solutions (particles) iteratively according to both the global and personal best-known positions. Ali et al. [20] employed the chi-Square method as a filter-based feature selection technique to enhance the predictive performance of models on the Cleveland Heart Disease dataset.

The paper by Abdar et al. [21] focuses on feature selection methods applied to the dataset by Z-Alizadeh Sani using genetic algorithm (GA) and PSO. These optimization algorithms are utilized to identify the most relevant features within the dataset for improved classification performance. The GA mimics natural selection processes by iteratively creating a population of viable solutions through crossover, mutation, and selection processes. In contrast, PSO models the behavior of particles within a search space by continuously updating their positions based on both individual and collective best solutions.

In the paper by Amin et al. [22], a feature selection method, which is known as the Brute Force Method was employed on the Cleveland dataset. This method encompasses exhaustively evaluating all possible feature subsets to determine the optimal combination for the given dataset and classification task. By systematically testing every feature subset, the Brute Force Method ensures that the most relevant features are selected, potentially improving the accuracy and efficiency of the classification model. Notably, this method might be computationally demanding, especially for datasets having more number of features. Notwithstanding its computational demands, the Brute Force Method offers a rigorous and comprehensive approach to feature selection, it may be precious in domains where accuracy is paramount, such as medical diagnostics.

The authors, Gárate-Escamila et al. [8], employed two feature selection methods in their work on the Cleveland and Hungarian heart disease datasets, i.e., chi-square and principal component analysis (PCA). To evaluate the significance of categorical features containing the target variable, they used chi-square. The datasets are made less dimensional by using PCA, enabling the efficient capture of the most relevant characteristics in a lower-dimensional space. These feature selection methods are applied to the Cleveland and Hungarian heart disease datasets in their study. The paper by Theerthagiri and Vidya [23] focused on methods for feature selection applied to a heart disease dataset. Specifically, the paper employed RFE as the feature selection technique. RFE is a wrapper method that iteratively selects subsets of features by training the model multiple times and eliminating insignificant features in each iteration. This method continues until the target number of features is obtained or until a predetermined performance parameter is optimized. RFE considers the interactions among features, ensuring that the most relevant ones are retained for classification or prediction tasks pertinent to heart disease.

Tian and Shi [24] employed feature selection methods on the Cleveland dataset using Modified Particle Swarm Optimization (MPSO). The main goal of the research was to improve feature selection such that heart disease prediction models perform better apropos classification. The MPSO in which the selection process likely involved iteratively evaluating different feature subsets to identify the most informative ones for classification. The study probably covered the performance of the MPSO-based feature selection strategy in comparison to other approaches to increase classification accuracy and decrease computing complexity.

3. Proposed Methodology

The proposed research aims to enhance the accuracy of a system by selecting the most relevant feature subset in a dataset related to heart disease. Fig. 1 illustrates the system framework. The essential elements of the framework are comprised of data collection, data preprocessing, feature optimization, and performance evaluation. Subsequent sections detail the foundational elements of the suggested framework.

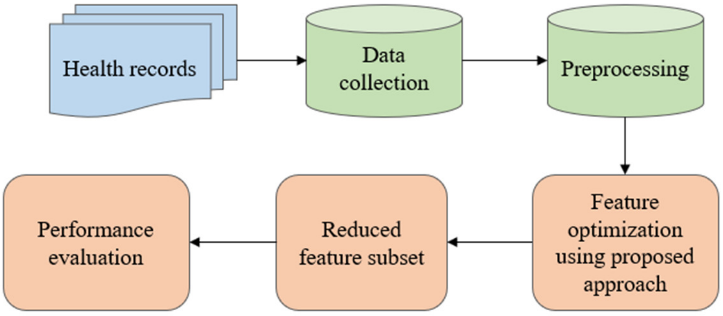

Fig. 1 Framework for the system

3.1. Data collection

The dataset used for this study is Cleveland Heart Disease datasets from the UCI web repository. This dataset initially contains 303 instances and 75 attributes. Table 1 presents a thorough overview of the dataset. The target feature label contains two values to determine the presence of cardiac disease.

Table 1 Feature subsets using feature optimization method

Sr. No.	Feature	Name	Values
1	Age	age	29 to 77
2	Sex	sex	1: Male 0: Female
3	Chest pain type	cp	1: Typical angina 2: Atypical angina 3: Non-angina pain 4: Asymptomatic
4	Blood pressure at rest	trestbps	from 94 mm Hg to 200 mm Hg
5	Serum cholesterol	chol	from 126 mg/dl to 564 mg/dl
6	Fasting blood sugar	fbs	FBSR > 120 mg/dl (True: 1, False: 0)
7	Resting electrocardiographic results	restecg	0: Normal 1: ST-T wave- abnormality 2: Hypertrophy
8	Maximum heart rate achieved	thalach	from 71 to 202
9	Exercise-induced angina	exang	1: Yes 0: No
10	ST depression induced by exercise relative to rest	oldpeak	1: Up sloping 2: Flat 3: down sloping
11	Slope at the peak exercise ST segment	slope	from 0 to 6.2
12	No. of major vessels	ca	from 0 to 3
13	Thallium	thal	3: Normal 6: Fixed defect 7: Reversible defect
14	Target	tar	1: Heart disease 0: No Heart disease

3.2. Data preprocessing

Processing the dataset is essential to accurately represent the quality of data. There are various methods to handle missing values, including ignoring them, replacing them with a numeric value, using the most frequent value for the feature, or substituting them with the mean value of the attribute. In this study, the initial step is to eliminate records containing missing values. To enhance the comparability and performance of machine learning algorithms, standardization is applied to the features. This process encompasses removing the mean and scaling to unit variance, aligning the features with a standard normal distribution. Given that machine learning algorithms preponderantly perform better when features adhere to a standard normal distribution, standardization is especially instrumental.

3.3. Feature Optimization

The experiment is conducted in this phase regardless of feature selection to assess its effects. Feature optimization is intended to select the important features related to heart diseases. Additionally, feature optimization facilitates the further establishment of precise models by removing or reducing the significance of irrelevant features, thus reducing training time and improving learning performance. This experiment examines the performance of various feature selection methods across filter, wrapper, and embedded categories. In this section, the proposed ant colony optimization with hill climbing (ACOHC) method is utilized for feature optimization.

The ACOHC algorithm combines two powerful algorithms into a hybrid feature selection method: the ACO algorithm and the HC algorithm. Fig. 2 depicts the flowchart of this approach. The ACOHC method commences by creating k artificial ants to explore the feature space. Initially, these ants search randomly within this space. Throughout a series of iterations, each ant selects certain features and develops its solution during each iteration t. The gathered subsets are then evaluated. By iteratively evaluating feature subsets and updating pheromone levels, ACO guides the search toward regions of higher performance.

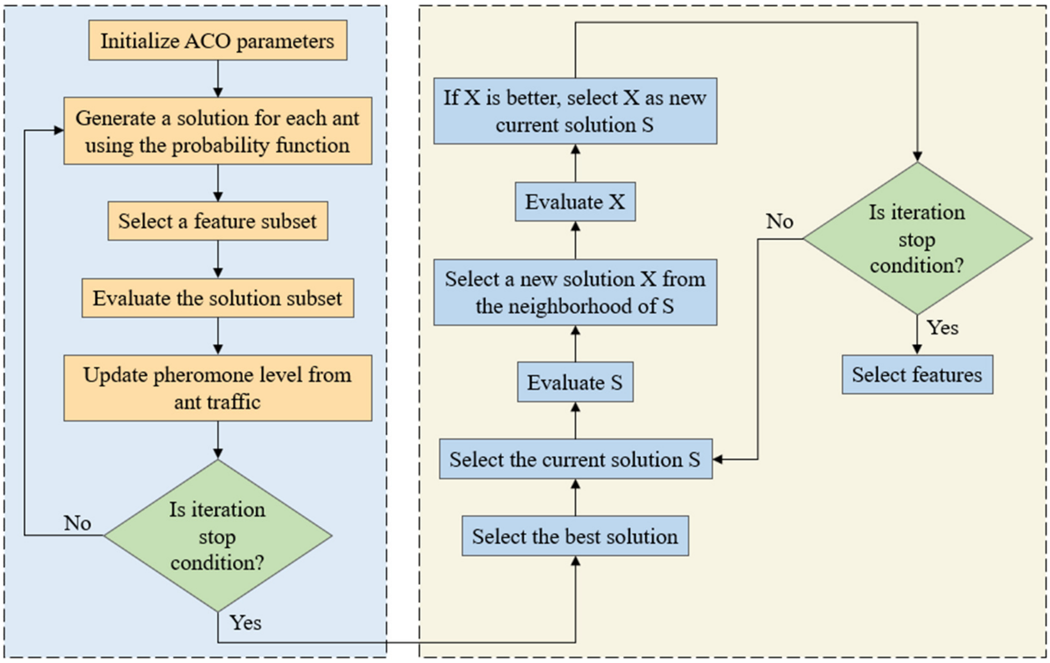
Initialize ACO parameters
Generate a solution for each ant using the probability function
Select a feature subset
Evaluate the solution subset
Update pheromone level from ant traffic
Is iteration stop condition?
No
Yes
If X is better, select X as new current solution S
Evaluate X
Select a new solution X from the neighborhood of S
Evaluate S
Select the current solution S
Select the best solution
No
Is iteration stop condition?
Yes
Select features

Fig. 2 Flowchart of feature optimization method

After the ACO identifies the initial feature subset, the chosen subset undergoes further enhancement through the application of the HC algorithm. HC is a heuristic local search technique that systematically investigates neighboring solutions to enhance the existing feature subset. By continuously optimizing an objective function, often denoted as the performance metric, HC gradually enhances the feature set towards greater classification accuracy. This iterative process of refinement empowers the algorithm to enhance the performance, as compared to the original feature subset identified by ACO. The ACOHC method uniquely combines ACO and HC to balance global exploration and local exploitation. ACO explores diverse solutions through probabilistic decisions and pheromone updates, while HC refines these solutions by local optimization. This hybrid approach leverages the exploration strength of ACO and the fine-tuning capability of the HC, yielding higher-quality solutions and reduced risk of premature convergence

A. Construction of feasible solutions

In ACOHC, each ant begins constructing a solution by randomly choosing an initial feature. Subsequently, it selects the subsequent feature from the pool of unchosen features based on a specified probability. The probability that an ant_k, currently at feature i will move to feature j at time t is:

$$P_i^k(t) = \frac{(\tau_i(t))^{\alpha}(\eta_i(t))^{\beta}}{\sum_{j\in N_j^k}(\tau_j(t))^{\alpha}(\eta_j(t))^{\beta}},\ j\in N_j^k \tag{2}$$

where the characteristics that ant_k has not selected yet but has the option to select from N_j^k, which is the viable neighborhood for ant_k. It functions as the memory of an ant. At time t, the heuristic information of feature i is represented by $\eta_i(t)$. The feature-related pheromone value at time t is shown by $\tau_i(t)$. The coefficients α and β represent the effect of pheromone τ and heuristic information η, respectively. The parameters α and β are adjusted to real positive values following the parameter setting guidelines.

B. Pheromone updating

The pheromone update equations, which specify how to adjust the pheromone levels, are used by all ants to increase the pheromone values and update $\tau_i(t)$. These equations are determined by:

$$(t+1) = \tau_i(t)\times(1-\rho)+\Delta\tau_i(t) \tag{3}$$

$$\Delta\tau_i(t) = \begin{cases} Q,\ \textit{if the ant selects the feature} \\ 0,\ \textit{otherwise} \end{cases} \tag{4}$$

where, $\Delta\tau_i^k$ is the pheromone left behind by an ant k and found as an efficient solution for the present iteration, $\tau_i(t)$ is the amount of pheromone on the path i at time t, and ρ represents the features pheromone evaporation rate ($0 < \rho < 1$).

Given the phenomenon that ants tend to communicate with one another using the pheromone value in the ACO algorithm, each ant utilized the obtained information to propose better solutions, thereby increasing the efficiency of the solutions found after several iterations. In Eq. (3), the left term indicates the evaporation of pheromone across all edges, while the right term represents the increase in pheromone intensity due to deposition.

The time complexity of ACO is given by:

$$O(A\times I_{\text{ACO}}\times O(E)) \tag{5}$$

where A is the number of ants, I_{ACO} is the number of iterations of ACO, and $O(E)$ is the complexity of evaluating the objective function for a subset of features.

After ACO selects a subset of features, HC is applied to further refine this subset. Assuming the number of iterations for HC is I_{HC}, the time complexity of HC is presented as:

$$O(I_{\text{HC}}\times O(E_{\text{HC}})) \tag{6}$$

where I_{HC} is the number of iterations of HC, and $O(E_{\text{HC}})$ is the complexity of evaluating the objective function for the feature subset selected by ACO.

The combined time complexity of the hybrid approach, where ACO is followed by HC, is the sum of the time complexities of the two phases:

$$O(A\times I_{\text{ACO}}\times O(E)+O(I_{\text{HC}}\times O(E_{\text{HC}})) \tag{7}$$

The summary of the ACOHC algorithm is mentioned as follows:

```
Input: Initial features of the dataset
Output: Optimized feature set
Initialization: [number of generations, number of ants, pheromone value (r),
maximum features, heuristic value (q), pheromone evaporation rate ($\rho$), $\alpha$, $\beta$]
    repeat for each iteration
```

```
For each ant
    Select distinct features randomly
    Calculate probabilities for selecting features based on pheromone levels
    The selected features are appended to the ant_solutions list
End for
Evaluate the performance (accuracy) of each solution
Find the index of the ant solution with the highest accuracy using np.argmax
If the accuracy of this best ant solution > the best accuracy
    best_accuracy = ant_accuracies[best_ant_index]
    best_solution = ant_solutions[best_ant_index]
End if
(After evaluating all ant solutions and selecting the best one)
updates the pheromone
Return best_solution that represents the highest accuracy feature subset
End for
```

Initially, the feature set is optimized using the ACO algorithm. To further reduce the number of features, the optimized set of features will be input into the HC algorithm. The pseudocode of the HC algorithm is given below:

```
Input: Initial set of features obtained in ACO
Output: Final optimized feature set
Current solution = initial solution
repeat
    for all neighbors of the current solution do
      Obtain a random neighbor
        if accuracy > best accuracy then
          best accuracy = neighbor solution
          best solution = index of the neighbor solution
        end if
    end for
until the end of the iterations
```

The rationale for choosing ACO and HC specifically for feature selection in cardiovascular risk assessment lies in their complementary strengths. ACO renders a robust global search capability, efficiently handling high-dimensional data and avoiding local optimum, while HC offers effective local optimization, refining the feature sets identified by ACO. This combination ensures the selection of high-quality feature subsets, facilitating more accurate and reliable predictive models for cardiovascular risk assessment. This study operates under the following assumptions:

(1) The dataset is of high quality and representative, containing relevant features for heart disease prediction.

(2) Parameter settings for both ACO and HC are optimal or near-optimal, enhancing the feature selection process.

4. Results and Discussion

In the experiment, the parameters are set empirically as pheromone constant (updating) $\alpha = 1$, heuristic information $ß = 1$, number of ants = 10, number of iterations = 50, pheromone decay (trail evaporates) $\rho = 0.5$. The effectiveness of the proposed approach for classification is assessed using logistic regression (LR), decision tree (DT), random forest (RF), and SVM classifiers. The experimental configuration entails implementing the proposed method using Python 3.0 programming language. The interactive coding environment is provided by Google Colab to perform experiments in Python. LR employed the 'lbfgs' solver with L2 regularization and a high number of iterations (max_iter = 1000). RF utilized 100 estimators and the 'gini' criterion for node impurity calculation. DT utilized the 'gini' criterion for splitting, without any depth restriction. SVM employed the default 'rbf' kernel for non-linear separation and regularization parameters. Feature subsets obtained through the hybrid ACOHC are assessed with LR, DT, RF, and SVM classifiers, and the final feature subset is selected based on the ACOHC method. The performance is evaluated based on metrics including accuracy, precision, recall, F-score, and specificity.

The feature selection method proposed is evaluated against nine alternative feature selection techniques for comparison. The reduced feature subsets, which are produced by the various feature selection techniques, are displayed in Table 2. In the table, every existing feature selection method is tested with different sizes of feature subsets, such as 5, 6, and 7 features. Each feature is represented by a value of either 1 or 0. According to the attribute sequence below, if a selected feature is included in the feature subset, it will be represented by 1; otherwise, if the selected feature is excluded from the subset, it will be represented by 0. After feature optimization using the proposed ACOHC method the selected features are: ‘cp’, ‘trestbps’, ‘thalach’, ‘ca’, and ‘thal’.

Table 2 Feature subsets using feature optimization method

Feature selection method	age	sex	cp	trest-bps	chol	fbs	rest ecg	thal-ach	exang	old-peak	slope	ca	thal	Size of feature subset
CS	0	0	1	0	0	0	0	1	1	1	0	1	0	5
	1	0	1	0	1	0	0	0	1	1	0	1	0	6
	1	0	1	0	1	0	0	1	1	1	0	1	0	7
ANOVA	0	0	1	0	0	0	0	1	1	1	0	1	0	5
	0	0	1	0	0	0	0	1	1	1	1	1	0	6
	0	0	1	0	0	0	0	1	1	1	1	1	1	7
FSA	0	1	0	0	0	1	0	0	1	0	0	1	1	5
	0	1	0	1	0	1	0	0	1	0	0	1	1	6
	1	1	0	1	0	1	0	0	1	0	0	1	1	7
BEA	0	1	0	1	0	0	0	0	1	0	1	1	0	5
	1	2	0	1	0	0	0	0	1	0	1	1	0	6
	1	1	0	0	1	1	0	0	1	0	1	1	0	7
MI	0	1	0	0	0	0	0	1	0	1	0	1	1	5
	0	0	0	0	0	0	0	1	1	1	1	1	1	6
	0	0	1	0	0	0	0	1	1	1	1	1	1	7
L2	0	1	1	0	0	0	1	0	1	0	0	1	0	5
	0	1	1	0	0	0	0	0	1	0	1	1	1	6
	0	1	1	0	0	0	0	0	1	1	1	1	1	7
PSO	0	0	1	1	1	0	1	1	0	0	0	0	0	5
	0	1	1	1	1	0	1	1	0	0	0	0	0	6
	0	1	1	1	0	0	1	1	1	0	0	0	1	7
ACO	0	0	1	0	0	1	0	0	0	0	1	1	1	5
	1	0	1	1	0	0	0	0	1	1	0	1	0	6
	0	0	1	1	1	0	0	1	0	1	1	0	1	7
HC	0	0	1	1	0	0	0	1	0	0	0	1	1	5
	0	1	1	1	0	0	0	1	0	1	0	1	0	6
	0	1	1	1	1	1	0	1	0	0	0	1	0	7
ACOHC	0	0	1	1	0	0	0	1	0	0	0	1	1	5

The classification accuracies obtained using feature sets optimized by different feature selection techniques are shown in Table 3. The analysis demonstrates that models built from optimized feature subsets consistently outperform those built from the original feature set. Initially, training the original feature set with LR, DT, RF, and SVM yields a maximum accuracy of 54.0%, precision of 55.7%, sensitivity of 50.8%, and f-measure of 50.8%. However, after applying feature selection techniques, a significant improvement in classifier accuracy is observed across all models.

Table 3 Classification accuracies of reduced feature subsets

Feature selection method	No. of features	Accuracy			
		LR	DT	RF	SVM
---	13	0.540	0.557	0.508	0.508
CS	5	0.885	0.770	0.803	0.688
	6	0.885	0.770	0.770	0.819
	7	0.885	0.737	0.852	0.803

Table 3 Classification accuracies of reduced feature subsets (continued)

Feature selection method	No. of features	Accuracy			
		LR	DT	RF	SVM
ANOVA	5	0.885	0.786	0.819	0.688
	6	0.868	0.803	0.803	0.688
	7	0.885	0.836	0.836	0.852
FSA	5	0.836	0.786	0.843	0.852
	6	0.836	0.836	0.836	0.819
	7	0.819	0.704	0.770	0.786
BEA	5	0.803	0.836	0.803	0.737
	6	0.819	0.803	0.786	0.819
	7	0.818	0.789	0.871	0.814
MI	5	0.819	0.639	0.754	0.819
	6	0.819	0.704	0.704	0.836
	7	0.819	0.688	0.770	0.803
L2	5	0.721	0.819	0.836	0.737
	6	0.721	0.789	0.836	0.836
	7	0.721	0.704	0.737	0.786
PSO	5	0.819	0.688	0.770	0.786
	6	0.819	0.704	0.770	0.786
	7	0.868	0.786	0.852	0.885
ACO	5	0.803	0.819	0.868	0.803
	6	0.868	0.704	0.868	0.885
	7	0.901	0.819	0.852	0.868
HC	5	0.868	0.836	0.836	0.803
	6	0.836	0.786	0.852	0.803
	7	0.819	0.803	0.819	0.836

Table 4 evinces the performance of the proposed approach, compared to other currently used feature selection strategies for LR, DT, RF, and SVM classifiers on the Cleveland dataset irrespective of the number of features. The proposed ACOHC method exhibits a significant improvement in the performance of classifiers, in contrast to alternative methods of feature selection.

Table 4 Performance of various feature selection algorithms

Feature selection method	Accuracy			
	LR	DT	RF	SVM
CS	0.885	0.770	0.852	0.819
ANOVA	0.885	0.836	0.836	0.852
FSA	0.836	0.836	0.843	0.852
BEA	0.818	0.836	0.871	0.819
MI	0.819	0.704	0.770	0.836
L2	0.721	0.819	0.836	0.836
PSO	0.868	0.786	0.852	0.885
ACO	0.901	0.819	0.868	0.885
HC	0.868	0.836	0.852	0.836
ACOHC	0.902	0.869	0.836	0.951

The performance indicators for the manifold classifiers, which are assessed using the proposed approach, are displayed in Table 5. The table provides a detailed comparison of metrics such as accuracy, precision, recall, F1-score, specificity, and area under the receiver operating characteristic (AUROC) highlighting the effectiveness of each classifier. This comparison aids in selecting the most suitable classifier for accurate prediction.

Table 5 Performance summary of ACOHC

Classifiers	LR	DT	RF	SVM
Accuracy	0.902	0.869	0.836	0.951
Precision	0.897	0.862	0.862	0.931

Table 5 Performance summary of ACOHC (continued)

Classifiers	LR	DT	RF	SVM
Recall	0.897	0.862	0.806	0.964
F1-score	0.897	0.862	0.833	0.947
Specificity	0.906	0.875	0.867	0.939
AUROC	0.913	0.837	0.923	0.957

Fig. 3 presents the accuracies of heart disease prediction using multifarious feature optimization techniques to identify the optimal feature set. The results highlight differences in performance across methods, showing the impact of feature selection on model accuracy. This comparison aids in pinpointing the most effective technique for enhancing predictive outcomes in heart disease diagnosis.

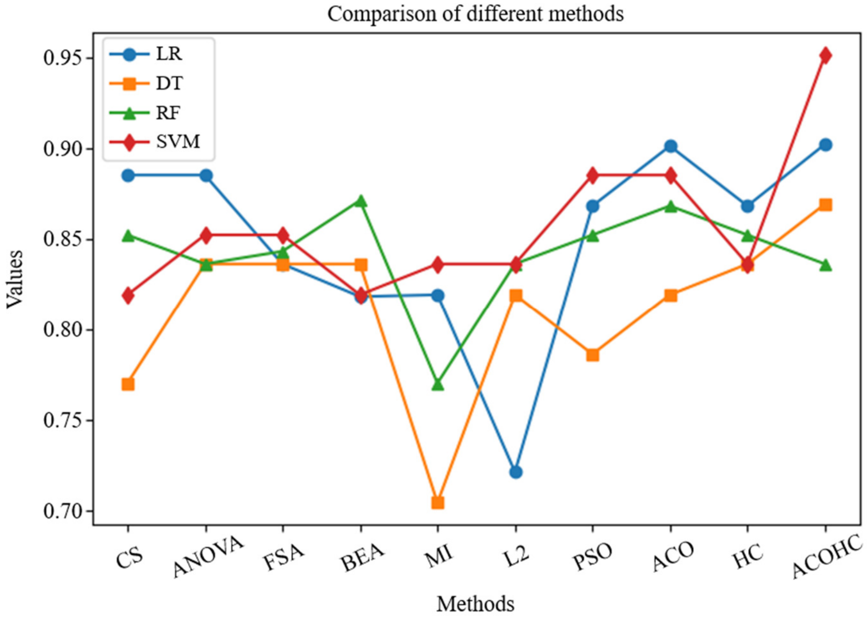

Fig. 3 Comparison of feature selection methods

Figs. 4-7 presents the AUROC curve scores for the classifiers evaluated on the Cleveland dataset. Given the provision of a single scalar value for evaluating the performance of the classification model across all threshold levels, AUROC is perceived to be advantageous. The scores indicate the ability of the model to dichotomize the patients according to the suffering of a certain disease. The AUROC scores demonstrate that the SVM has gained the highest performance, closely followed by RF, LR, and DT classifiers. All classifiers exhibit AUROC scores above 0.80, indicating robust performance.

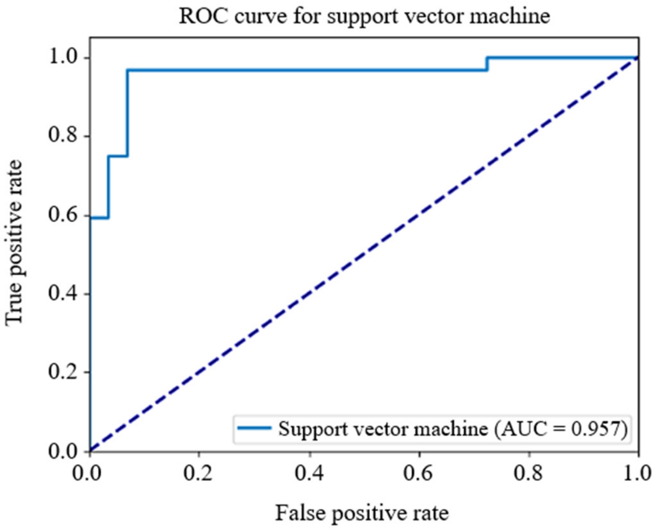

Fig. 4 AUROC analysis for SVM

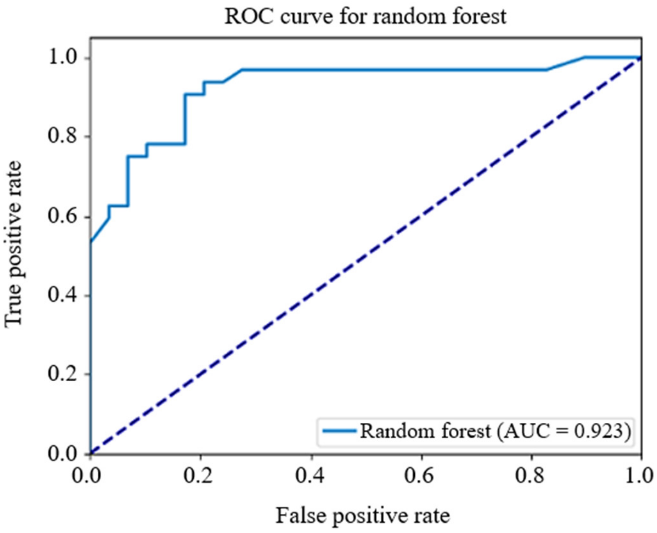

Fig. 5 AUROC analysis for RF

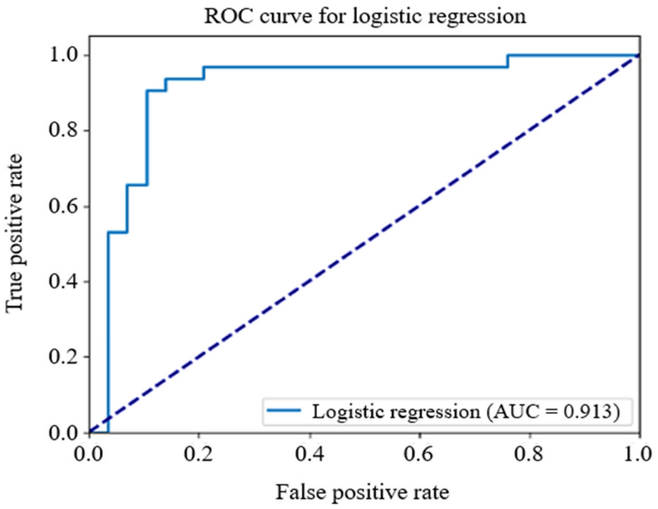

Fig. 6 AUROC analysis for LR

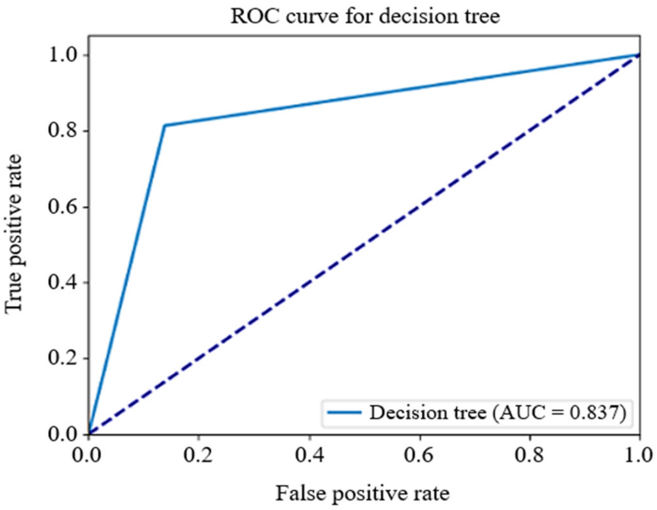

Fig. 7 AUROC analysis for DT

5. Conclusions

This paper aims to investigate the rationale for the prediction accuracy of heart disease affected by the feature selection techniques. Technically, this study is conducted against a collection of different features that were extracted from widely used Cleveland Heart Disease datasets that are available at the UCI using a range of feature selection approaches. Experiments have been conducted both including and excluding feature selection to determine the influence of feature selection.

Chi-square, ANOVA, FSA, BEA, MI, L2, PSO, ACO, and HC are utilized as algorithms for feature selection. Four techniques for classification are analyzed: SVM, RF, DT, and LR. The best result, using the DT classifier, yields 55.7% model accuracy without feature selection. Subsequently, feature selection is used to experiment. The highest accuracy value without feature selection is 55.7%; with the use of ACOHC and SVM classifier, this value is raised to 95.1%. The findings from the experiment suggest that feature selection algorithms could identify the disease accurately with less number of features. The additional key points are:

(1) The model achieves accuracies of 90.1%, 88.5%, and 83.6%, for the LR, DT, and RF classifiers respectively.

(2) The ACOHC technique attained a precision of 93.1%, specificity of 93.9%, f-score of 94.7%, and AUROC score of 95.7% using an SVM classifier.

A hybrid approach combines several feature selection strategies, ultimately enabling the extraction of the best feature subsets for model building. Future work can focus on using real-time medical records from different regions that may help to improve the accuracy of heart disease prediction algorithms. A limitation of this work is that if the data is not acquired properly or contains a high number of missing values, it may impact the quality of the features and, consequently, the performance of the system.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] V. P. Kavitha, V. Janarthanan, T. Annamalai, and M. Arumugam, “Enhancing Healthcare in the Digital Era: A Secure E-Health System for Heart Disease Prediction and Cloud Security,” Expert Systems with Applications, vol. 255, part A, article no. 124479, December 2024.

[2] O. Gaidai, Y. Cao, and S. Loginov, “Global Cardiovascular Diseases Death Rate Prediction,” Current Problems in Cardiology, vol. 48, no. 5, article no. 101622, May 2023.

[3] R. Bharti, A. Khamparia, M. Shabaz, G. Dhiman, S. Pande, and P. Singh, "Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning," Computational Intelligence and Neuroscience, vol. 2021, article no. 8387680, 2021.

[4] A. Hutke and J. Deshmukh, "A Systematic Review of Machine Learning Approaches and Missing Data Imputation Techniques for Predicting Heart Disease," International Conference on Advanced Computing Technologies and Applications, pp. 1-5, October 2023.

[5] H. Rao, X. Shi, A. K. Rodrigue, J. Feng, Y. Xia, M. Elhoseny, et al., "Feature Selection Based on Artificial Bee Colony and Gradient Boosting Decision Tree," Applied Soft Computing, vol. 74, pp. 634-642, January 2019.

[6] A. Bhattacharya, R. T. Goswami, K. Mukherjee, and N. G. Nguyen, "An Ensemble Voted Feature Selection Technique for Predictive Modeling of Malwares of Android," International Journal of Information System Modeling and Design, vol. 10, no. 2, pp. 46-69, 2019.

[7] M. Ahsan and Z. Siddique, "Machine Learning-Based Heart Disease Diagnosis: A Systematic Literature Review," Artificial Intelligence in Medicine, vol. 128, article no. 102289, June 2022.

[8] A. K. Gárate-Escamila, A. H. El Hassani, and E. Andrès, "Classification Models for Heart Disease Prediction Using Feature Selection and PCA," Informatics in Medicine Unlocked, vol. 19, article no. 100330, 2020

[9] U. Moorthy and U. D. Gandhi, "A Novel Optimal Feature Selection Technique for Medical Data Classification Using ANOVA Based Whale Optimization," Journal of Ambient Intelligence and Humanized Computing, vol. 12, no. 3, pp. 3527-3538, March 2021.

[10] R. Aggrawal and S. Pal, "Sequential Feature Selection and Machine Learning Algorithm-Based Patient's Death Events Prediction and Diagnosis in Heart Disease," SN Computer Science, vol. 1, no. 6, article no. 344, November 2020.

[11] M. A. Sulaiman and J. Labadin, "Feature Selection Based on Mutual Information," 9th International Conference on IT in Asia, pp. 1-6, August 2015.

[12] S. Kaushik, A. Choudhury, A. K. Jatav, N. Dasgupta, S. Natarajan, L. A. Pickett, et. al., "Comparative Analysis of Features Selection Techniques for Classification in Healthcare," Lecture Notes in Computer Science, in press.

[13] S. Tabakhi, P. Moradi, and F. Akhlaghian, "An Unsupervised Feature Selection Algorithm Based on Ant Colony Optimization," Engineering Applications of Artificial Intelligence, vol. 32, pp. 112-123, June 2014.

[14] Purushottam, K. Saxena, and R. Sharma, "Efficient Heart Disease Prediction System," Procedia Computer Science, vol. 85, pp. 962-969, 2016.

[15] M. A. Jabbar, B. L. Deekshatulu, and P. Chandra, "Prediction of Heart Disease Using Random Forest and Feature Subset Selection," Proceedings of the 6th International Conference on Innovations in Bio-inspired Computing and Applications, pp. 187-196, December 2015.

[16] W. Wiharto, H. Kusnanto, and H. Herianto, "Interpretation of Clinical Data Based on C4.5 Algorithm for the Diagnosis of Coronary Heart Disease," Healthcare Informatics Research, vol. 22, no. 3, pp. 186-195, 2016.

[17] A. U. Haq, J. P. Li, M. H. Memon, S. Nazir, and R. Sun, "A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms," Mobile Information Systems, vol. 2018, article no. 3860146, 2018.

[18] Y. Khourdifi and M. Bahaj, "Heart Disease Prediction and Classification Using Machine Learning Algorithms Optimized by Particle Swarm Optimization and Ant Colony Optimization," International Journal of Intelligent Engineering and Systems, vol. 12, no. 1, pp. 242-252, 2019.

[19] A. Jain, S. Tiwari, and V. Sapra, "Two-Phase Heart Disease Diagnosis System Using Deep Learning," International Journal of Control and Automation, vol. 12, no. 5, pp. 558-573, 2019.

[20] L. Ali, A. Rahman, A. Khan, M. Zhou, A. Javeed, and J. A. Khan, "An Automated Diagnostic System for Heart Disease Prediction Based on $\chi 2$ Statistical Model and Optimally Configured Deep Neural Network," IEEE Access, vol. 7, pp. 34938-34945, 2019.

[21] M. Abdar, W. Książek, U. R. Acharya, R. S. Tan, V. Makarenkov, and P. Pławiak, "A New Machine Learning Technique for an Accurate Diagnosis of Coronary Artery Disease," Computer Methods and Programs in Biomedicine, vol. 179, article no. 104992, October 2019.

[22] M. S. Amin, Y. K. Chiam, and K. D. Varathan, "Identification of Significant Features and Data Mining Techniques in Predicting Heart Disease," Telematics and Informatics, vol. 36, pp. 82-93, March 2019.

[23] P. Theerthagiri and J. Vidya, "Cardiovascular Disease Prediction Using Recursive Feature Elimination and Gradient Boosting Classification Techniques," Expert Systems, vol. 39, no. 9, article no. e13064, November 2022.

[24] D. Tian and Z. Shi, "MPSO: Modified Particle Swarm Optimization and Its Applications," Swarm and Evolutionary Computation, vol. 41, pp. 49-68, August 2018.