Evolutionary Computing-Based Feature Selection for Cardiovascular Disease: A Review

J. Jasmine Gabriel and L. Jani Anbarasi

Abstract Cardiovascular diseases (CVDs) are one of the most lives threatening primary reasons in reducing the mortality of human beings, where one-third of global death is due to CVDs. People who fall under the age group of 75 were mostly affected by CVD. It causes quite 30% of the deaths throughout the world between the ages of 30 and 70. It is high time to identify the disease at the early stage to enhance the expectancy rate through the accuracy of disease prediction. To develop a better prediction, the input to the system needs to refine first, i.e., the selection of appropriate, relevant features. Feature selection is important in the detection of CVDs for better classification resulting in better prediction. However, to meet the challenges in the feature selection process, the evolutionary computing method obtained more attention with improved results comparatively. This paper describes a survey on evolutionary computing-based feature selection techniques, its merits, demerits, and contribution in the classification of CVDs. This detailed, comprehensive work might help the researcher a better understanding of evolutionary computing in heart disease prediction.

Keywords Heart disease dataset · Feature selection · Evolutionary computing · CVDs · Prediction

1 Introduction

Living in this epoch and alter of lifestyle play an important role in affecting the human's physical state and mental condition. Any disorder or malfunctioning of the body or mind that destroys healthiness is known as a disease. Diseases are caused due to various reasons. Every disease has certain traits to spot the categories of diseases.

J. Jasmine Gabriel (✉) · L. Jani Anbarasi
School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, India
e-mail: jasminegabriel.j2019@vitstudent.ac.in

L. Jani Anbarasi
e-mail: janianbarasi.l@vit.ac.in

R. J. Kannan et al. (eds.), *International Virtual Conference on Industry 4.0*,
Lecture Notes in Electrical Engineering 355,
https://doi.org/10.1007/978-981-16-1244-2_4

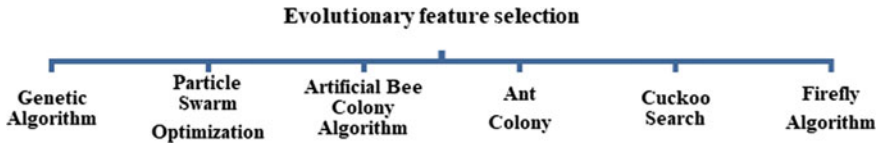

Fig. 1 Evolutionary computing feature selection techniques

World Health Organization defined some virulent diseases in humans are cardiovascular disease a heart disease, cancer, lower respiratory infections, HIV, trachea, bronchus, tuberculosis, and diabetes mellitus. CVDs are a collection of malfunctions in the heart, and the blood vessels connected to it, i.e., the heart does not function properly due to fatty deposits on the stream of blood vessels. Fat prevents the flow of blood to the heart or brain [1]. The traditional methods of diagnosis of CVD are electrocardiogram (ECG) (uneven heartbeat), Holter monitoring (handy ECG device), echo cardiogram (ultrasound of your chest) [2, 3], stress test (raising your heart rate with exercise), cardiac catheterization (heart abnormalities is checked by catheter into an artery), multidetector computerized tomography (CT) scans (different angled X-ray picture of heart and chest), and cardiac magnetic resonance imaging (MRI magnetic field pictures of the heart).

This traditional way of diagnosis may lead to false results due to human error and delayed results due to the number of different tests we take based on the availability of the doctor and the radiologist. The limitation also includes side effects and high cost. It is high time the CVD has to be diagnosed and treated at the earliest. The severity of the disease depends on how early it was detected to increase the mortality of humans. Early detection can be automated by using the heart disease dataset and use them for prediction mechanism. Many researchers and academicians have already proposed various prediction methods. The demanding need for a better prediction model with high accuracy of disease prediction with less cost and less time is required for early diagnosis.

In machine learning, input data is preprocessed by choosing relevant feature selection, and missing value imputation [4] plays an essential role in some real-life applications. Of which, identifying relevant features helps in different fields like false detection, text mining, disease diagnosis, face detection, image processing, bioinformatics, and socio-industrial application. In those applications, the quantity and the quality of features are more critical in the classification and prediction process. The higher numbers of features are challenging to visualize. Irrelevant features create noisy data that affect the accuracy of the prediction model [5]. The traditional feature selection techniques are filter, wrapper, and embedded method. FS can also be a hybrid method (the combination of filter and wrapper methods) and the ensemble method (combining many classifiers into one desired model by aggregating the results) where those methods of feature selection are popular among the researchers in selecting the better feature. Another fast-growing feature selection technique is evolutionary computing, which is inspired by the natural behavior of animals and birds.

Figure 1 shows the 6 meta-heuristic evolutionary computing like genetic algorithm (GA), particle swarm optimization (PSO), artificial bee colony algorithm (ABCA), ant colony optimization (ACO), cuckoo search (CS), and firefly algorithm (FA) specifically carried on in this survey.

2 Literature Survey

Researchers analyzed numerous techniques in developing an automated method for finding a global optimal solution for heart disease prediction by improving its accuracy. This section gives a brief literature survey of different research works using the evolutionary computing-based feature selection specifically for heart disease prediction.

Ismail et al. [6] proposed a novel FS method using BPSO in treating coronary heart disease with EST dataset (ECG data) of 23 attributes of 480 instances. Two feature selection methods like GA and binary particle swarm optimization (BPSO) methods are used for finding the relevant features. The selected features were trained with three different classifiers namely BPSO with support vector machine (SVM) kernel, GA with SVM kernel, and simple SVM method. The BPSO with SVM classifier showed better results in terms of selecting 11 relevant features and improved classification accuracy of 81.46%.

Jabbar et al. [7] proposed a GA-based heart disease prediction for Andhra Pradesh dataset with 12 features using GA-based association rule. Each feature is represents an individual items with the frequent itemset that are derived with minimum support of 6. Fitness function using a G-mean measure generates frequent itemset with two different sets of 7 frequent itemsets are found to be relevant. Frequent itemset creates a lesser number of rules for disease prediction. The author prefers GA for a higher level of prediction and better feature interaction than the greedy algorithm.

Subanya et al. [8, 9] proposed a novel feature selection method with ABC and SVM classifiers. The classifier sets the processing parameters manually for the number of iteration, limit, number of dimensions, and number of employer bee and onlooker bee values. The proposed method showed good classification accuracy compared with forward ranking and reverse ranking. The ABC-SVM showed improved accuracy with seven features. The author identified features like age and fasting blood sugar as notable features in improving the accuracy of 86.76%. The same author enhanced the model with binary artificial bee colony algorithm-K-nearest neighbor (BABC-KNN), for heart disease classification. This model uses the BABC method for feature selection which resulted in selecting 6 prominent features. The selected features are trained using the KNN model with eight existing models for comparison and achieved improved accuracy of 92.4%

Long et al. [10] proposed a new diagnosis method for feature selection using fuzzy classification. The feature selection was carried out by two different methods chaos firefly algorithm rough set-attribute reduction (CFARS-AR) and BPSORS-AR. Comparatively, CFARS-AR is efficient with only four relevant features tested

with Cleveland dataset and SPECTF. Fuzzy c-means clustering is used on selected features to obtain the required number of fuzzy and structure of the fuzzy rule. Those values are input to the newly defined approach, interval type-2 fuzzy logic system (IT2FLS) classifier. The proposed model showed improved accuracy of 88.3% over other classifiers like Naïve Bayes, SVM, and artificial neural networks (ANN).

Verma et al. [11] developed a hybrid model with correlation-based FS (CFS) and PSO as the feature selection method for the real-time dataset from Indira Gandhi medical college, Shimla and Cleveland dataset. The k-means clustering tests the selected feature for eliminating incorrectly classified clusters. The resultant features are classified using multilayer perceptron (MLP), fuzzy unordered rule induction algorithm (FURIA), C4.5, and multinomial logistic regression (MLR) classifier. MLP classifier with 7 features achieved an accuracy of 90.28%.

Thippa Reddy et al. [12] proposed a diagnosis model using cuckoo search and classification with fuzzy logic. Cuckoo search (CS) and rough set (RS) helped to pick the most exceptional feature through fitness function without losing its precision value. CS with RS shows reduced 6 features when compared with FA with RS, BAT with RS, and locality preserving projection (LPP). The selected features are fed into the fuzzy classifier to obtain its fuzzy score. The author used three different datasets for comparison. The proposed method CS with RS feature selection with RS classifier of Cleveland heart disease dataset resulted in improved accuracy with 91.5% when compared with other fuzzy models and other datasets used.

Iftikhar et al. [13] proposed a new healthcare model for identifying heart disease risk factors using Cleveland dataset. The proposed model includes 4 test scenarios. In the first and second methods, the model tests with the least square linear method (SVM-LS) and sequential minimum optimization (SVM-SMO-RBF) without feature selection. The third and fourth method tests GA-SVM-SMO-RBF and PSO-SVM-SMO-RBF with feature selection. Of all the four methods, the GA-SVM-SMO-RBF method showed improved accuracy with only seven selected features.

Dulhare [14] developed a prediction system with Cleveland dataset using PSO and GA for feature selection. PSO eliminated features with low PSO values resulting in the selection of 7 relevant features which is better than GA. The selected relevant features are fed into NB classifier for 100 iterations to obtain better accuracy.

Gokulnath et al. [15] defined a wrapper-based feature selection with GA-SVM. The feature selection of GA is compared with existing algorithms like relief, correlation-based, filtered subset, information gain, consistency subset, chi-squared, one attribute-based, filtered attribute, and gain ratio. GA algorithm efficiently chooses 7 relevant features out of 13. The selected features use Z score optimization for preprocessing. The system is tested with classifiers like SVM, MLP, J48, and KNN. SVM classifier showed improved accuracy of 88.34% when compared with the other classifier.

The literature survey focuses on existing EC-based feature selection for diagnosing the heart disease dataset, a lesser number of relevant features, classifier used, accuracy achieved, and the different types of heart disease dataset used. Table 1 explains the overall comparison of this survey.

Table 1 Summarized of evolutionary computing-based feature selection

Reference	FS	Classifier used	SF/TF	List of selected features	Accuracy (%)	Advantage	Disadvantage	Dataset used
Ismail et al. [6]	BPSO	SVM	11/23	NA	81.46	• Minimal feature • Less training and testing time	200 iterations for BPSO	EST dataset
Jabbar et al. [7]	GA	Association rule based	7/12	Sex, blood pressure, resting ECG, smoking, alcohol, family history of CAD, rural/urban	–	• Higher level prediction • Better feature interaction	No performance metric was evaluated	Andhra Pradesh dataset
Subanya et al. [8]	ABC	SVM	7/13	Age, Chol, Fbs, Restecg, Thalach, slope, ca	86.76	Good classification	Parameters are manually chosen	Cleveland
Subanya et al. [9]	BABC	KNN	6/13	Cp, Trestbps, Chol, Thalch, Slope and Thal	92.4	Improved accuracy	–	Cleveland
Long [10]	CFARS- (AR)	IT2FLS	4/13	NA	88.3	• High accuracy • Reduce computation expenses with high dimension data	• Cost is more when used on high dimensional dataset • Training time is slow	Cleveland

(continued)

Table 1 (continued)

Reference	FS	Classifier used	SF/TF	List of selected features	Accuracy (%)	Advantage	Disadvantage	Dataset used
Verma [11]	CFS + PSO	Clustering + MLP	7/13	cp, thalch, exang, old peak, slope, ca, and thal	90.28	• K-means clustering is used	–	Cleveland, Switzerland, and Hungarian
Gadekallu [12]	CS + RS	Fuzzy	6/13	NA	91.5	Less computational cost	• Space complexity can be considered	Cleveland, Switzerland, and Hungarian
Iftikhar et al. [13]	GA	SVM-SMO-RBF	7/13	Cp, Restecg, Thalach, Exang, Slope, Ca, Thal,	88.10	• Better accuracy • Reduced search space	–	Cleveland
Dulhare [14]	PSO	NB	7/13	cp, Restecg, Thalach, Exang, Old peak, Ca, Thal	87.91	• Minimal FS • Better classification	Features were selected at the 70^{th} iteration	Statlog
Gokulnath [15]	Wrapper-based GA	SVM	7/13	Cp, Restecg, Thalach, Exang,Oldpeak, Ca, Thal	88.34	• ROC curve show good performance with SVM	Accuracy is not more than 90%	Cleveland

3 Results and Observation on EC-Based FS

The following are the datasets used by various researchers in this survey for preprocessing the heart disease are EST dataset, Andhra Pradesh dataset, Cleveland, Switzerland, Hungarian, and Statlog heart disease dataset. Table 1 gives the summary of literature survey performed on 10 different papers on heart disease dataset, where all the authors aimed at selecting the lesser number of feature through EC-based FS method for improving the classification accuracy with a suitable classifier.

Figure 2 shows the comparison of the achieved accuracy by processing the heart disease dataset by various researchers. It shows that the BABC method of EC-based feature selection has the highest accuracy of 92.4% with a minimum number of 6 relevant features. Moreover, the accuracy rate ranges from 87 to 92%. Still, a better way of using the EC method is needed to improve the accuracy of heart disease prediction.

Based on the observation, the most of the datasets have these 13 common features in the heart disease dataset. 10 out of 7 authors use the Cleveland heart disease dataset for their proposed work. The attribute of the dataset are age, gender, cp (chest pain), trestps (level of blood pressure at resting), chol (serum cholesterol), fbs (fasting blood sugar), restecg (resting value of ECG), thalach (maximum heart rate), exang (exercise -induced angina), oldpeak (St depression induced by exercise vs. rest), slope (ST segment in terms of a slope), ca (number of major vessels colored by fluroscopy), and thal (defect type). Figure 3 shows the most selected feature using evolutionary computing methods as thalach, cp, restecg, ca, and thal. Dataset referred in this paper is available in the UCI repository [16, 17].

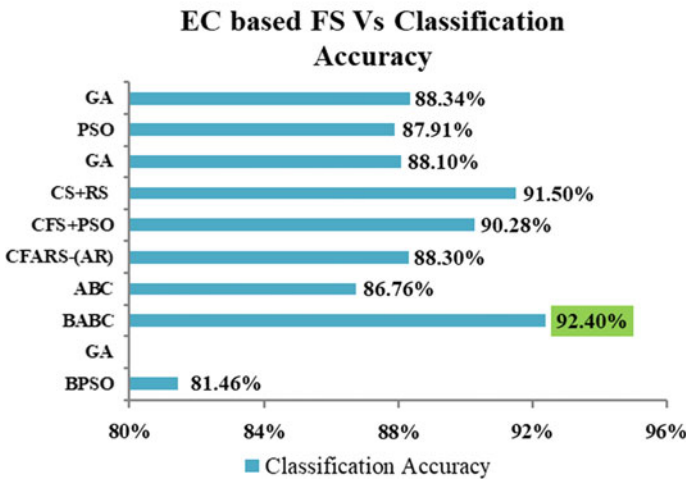

Fig. 2 EC-based FS versus classification accuracy

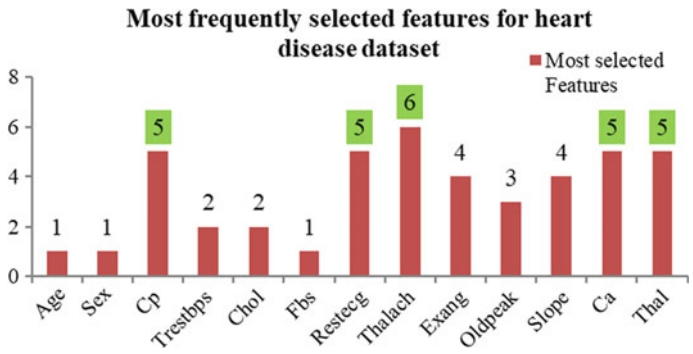

Fig. 3 Most frequently selected features based on the survey

Table 2 gives the insight knowledge about the methods with their strength and weakness of EC-based FS techniques. This brief view can help the researchers in choosing the best method for their problem. Particularly, in dealing with a complex problem and high dimensional dataset, the EC method approach would make a better choice for feature selection.

4 Conclusion and Future Work

The comprehension survey states, more number of features might result in misclassification and overfitting. Most of the researchers focused on two important issues. A suitable feature selection method to obtain less number of relevant features and a better classifier for improving accuracy. The evolutionary computing-based feature selection method met the expectation of a researcher in solving the mentioned issues with efficient identification of global optimal solution, better interaction between features, higher level of prediction, flexibility, and better visual output. The evolutionary computing-based feature selection built a better classifier, a better classifier results in better prediction. Hence, the growing need for handling high dimensional dataset with better prediction automated systems can be constructed for early detection and diagnosis of CVDs. In future work, a fully automated EC-based feature selection can be carried out with a real-time dataset for better heart disease prediction.

Table 2 Strength and weakness of evolutionary computing methods

EC methods		Key points
GA	Strength	• Identify the best attribute by a fitness function • Higher level prediction and find search space easily • Evaluate each subset by executing the model • Better interaction among features
	Weakness	• Sometime, result is in invalid states • Computationally expensive • Poor performance with a high dimensional dataset • Does not guarantee minimal features
PSO	Strength	• Deal with real numbers optimization problems • Implementation is simple and easy continuous optimization • Lesser time complexity
	Weakness	• Does not guarantee minimal features • In high dimensional dataset can fall into the local optimum solution • Slow during iteration
ABCA	Strength	• Fewer control parameters • Flexible and fast during iteration
	Weakness	• Premature convergence where cannot produce offspring's • Low accuracy
ACO	Strength	• Can use for dynamic approach, inherit parallelism • Positive feedback information and discrete optimization
	Weakness	• Challenging to work with high dimension data • Too many parameters were theoretically difficult • Distribution change over iteration
CS	Strength	• Aim to find the global optimal solution • Use fewer parameters • Use of levy flight process finds the optimal solution
	Weakness	• Continuous optimization • Difficult in solving a complex problem
FA	Strength	• Simple mathematical operator • Find global optima faster with a higher success rate • Find the globally optimal solution • Does not worry about initial iteration and velocity as PSO
	Weakness	• Computational expensive both by memory and runtime

References

1. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC530409
2. Sharon JJ, Anbarasi LJ (2018) Diagnosis of DCM and HCM heart diseases using neural network function. Int J Appl Eng Res 13(10):8664–8668
3. Sharon JJ, Anbarasi LJ, Edwin Raj B (2018) DPSO-FCM based segmentation and classification of DCM and HCM heart diseases. In: 2018 Fifth HCT information technology trends (ITT). IEEE
4. Jasmine Gabriel J, Valarmathie P (2012) Unified clustering technique for microarray gene expression data. In: International conference on computing and control engineering. ISBN: 978-1-4675-2248-9

5. Hira ZM, Gillies DF (2015) A review of feature selection and feature extraction methods applied on microarray data. Adv Bioinform 1–13
6. Babaoglu I, Findik O, Ulker E (2010) A comparison of feature selection models utilizing binary particle swarm optimization and genetic algorithm in determining coronary artery disease using support vector machine. Expert Syst Appl 37(4):3177–3183
7. Jabbar Akil M, Deekshatulu BL, Chandra P (2012) Heart disease prediction system using associative classification and genetic algorithm. Elsevier 1, 183–192
8. Subanya B, Rajalaxmi RR (2014) A novel feature selection algorithm for heart disease classification. Int J Comput Intell Inf 4(2):117–124. ISSN: 2349-6363
9. Subanya B, Rajalaxmi R (2014) Feature selection using artificial bee colony for cardiovascular disease classification. In: 2014 International conference on electronics and communication systems (ICECS)
10. Long NC, Meesad P, Unger H (2015) A highly accurate firefly based algorithm for heart disease prediction. Expert Syst Appl 1–9. Verma L, Srivastava S, Negi PC (2016) A hybrid data mining model to predict coronary artery disease cases using non-invasive clinical data. J Med Syst 40(7):1–7
11. Gadekallu TR, Khare N (2017) Cuckoo search optimized reduction and fuzzy logic classifier for heart disease and diabetes prediction. Int J Fuzzy Syst Appl 6(2):25–42
12. Iftikhar S, Fatima K, Rehman A, Almazyad AS, Saba T (2017) An evolution based hybrid approach for heart diseases classification and associated risk factors identification. Biomed Res Int J Med Sci 28(8)
13. Dulhare UN (2018) Prediction system for heart disease using Naive Bayes and particle swarm optimization. Biomed Res 29(12)
14. Gokulnath CB, Shantharajah SP (2018) An optimized feature selection based on genetic approach and support vector machine for heart disease. Cluster Comput 1–11
15. Espejo PG, Ventura S, Herrera F (2010) A survey on the application of genetic programming to classification. IEEE Trans Syst Man Cybernet Part C App Rev 40(2):121–144
16. The UCI machine learning repository [online]. Available at: https://archive.ics.uci.edu/ml/datasets/Heart+Disease
17. http://archive.ics.uci.edu/ml/datasets/Statlog+%28Heart%29