



New hybrid method for heart disease diagnosis utilizing optimization algorithm in feature selection

Jalil Nourmohammadi-Khiarak¹ · Mohammad-Reza Feizi-Derakhshi² · Khadijeh Behrouzi³ · Samaneh Mazaheri⁴ · Yashar Zamani-Harghalani² · Rohollah Moosavi Tayebi⁴

Received: 15 October 2019 / Accepted: 10 November 2019 / Published online: 3 December 2019
© The Author(s) 2019

Abstract

The number and size of medical databases are rapidly increasing, and the advanced models of data mining techniques could help physicians to make efficient and applicable decisions. The challenges of heart disease data include the feature selection, the number of the samples; imbalance of the samples, lack of magnitude for some features, etc. This study mainly focuses on the feature selection improvement and decreasing the numbers of the features. In this study, imperialist competitive algorithm with meta-heuristic approach is suggested in order to select prominent features of the heart disease. This algorithm can provide a more optimal response for feature selection toward genetic in compare with other optimization algorithms. Also, the K-nearest neighbor algorithm is used for the classification. Evaluation result shows that by using the proposed algorithm, the accuracy of feature selection technique has been improved.

Keywords Classification · Feature selection · Data mining · Imperialist competitive algorithm · K-nearest neighbor

1 Introduction

Heart diseases or cardiovascular diseases (CVD) are one type of diseases that include heart or vessels (veins and arteries). Ten percent of total death in the early twentieth century resulted from heart diseases [1], and the death rate due to these diseases were increased by 25% in the late twentieth century. Heart diseases mainly affect individuals with 65 years old and older, and it has taken the place of infectious diseases as the main reason of death in the world [2]. It was believed that heart diseases are problematic in developed countries, but

now they are also expanding in developing countries, since there is no proper health care in these countries.

Recently, World Health Organization has reported that the most common reason of death in the world is CVD. People die from CVD more than any other factor. It has been estimated that 17.1 million individuals died from CVD in 2004, which is 29% of total death in the world. The reason of 7.2 million of these deaths was coronary artery disease that is one of the most common CVD. 5.7 million Of deaths occurred due to cardiac trauma. Countries with low and moderate incomes are affected adversely, 82% of death caused by CVD occurs in these countries, and this rate is similar in both male and female. It is anticipated that until 2030 about 23,600,000 individuals would mostly die from heart diseases and brain stroke. The most increase of death will occur in the eastern Mediterranean region, and most deaths will occur in the Southeast Asia because of changes of life styles, food habits, and occupational culture. Therefore, according to the reports of the WHO, utilizing accurate methods and efficient periodical examination of heart to diagnose heart diseases are very crucial [3].

Heart diseases are the most common reason of death in Iran as well; since, the Iranian ministry of health reported that 46% of death is ceased by heart diseases [4]. The significant growth of these diseases and their complications, and their high costs

✉ Jalil Nourmohammadi-Khiarak
Jalil.Nourmohammadi@elka.pw.edu.pl

¹ Faculty of Electronics and Information Technology, Warsaw University of Technology, 00-661 Warszawa, Poland

² Faculty of Electrical and Computer Engineering, University of Tabriz, Tabriz, Iran

³ Faculty of Electrical and Computer Engineering, Islamic Azad University, Meimand, Iran

⁴ Faculty of Science, Computer Science, University of Ontario, Institute of Technology, Ontario, Canada

adversely affect the societies and impose a lot of financial and physical burden on the international community. Therefore, using effective methods to do prevention is very vital.

The construct of a heart disease that includes complete medical evaluation, history taking, and examination with early diagnosis of heart diseases can decrease the mortality rate. One of the best ways to diagnose heart disease is echocardiogram. Echocardiogram or cardiac echo is a painless test using acoustic waves to create images from the heart [5].

However, interpretation of cardiac echo images is a challenge as there is no accurate rule available to infer the data properly. Data analysis of cardiac echo by experts is very time-consuming, and also there are not many experts to do so. Therefore, the approach of automatic cardiac echo interpreter to minimize the human efforts is very important in diagnosis of heart disease. There is a method to extract the hidden data from a huge set of data collected in the past to solve the problems related to diagnosis of diseases. Data mining can make rules from this huge set of data in order to use them in the concept of cardiac echo [6]. Data mining can play an important role to extract data from huge datasets, and it has gained a crucial position in healthcare, recently. The process of extraction is related to classification, clustering, and relationship law discovery, instead of data analysis.

Two datasets are used in the paper, namely; the UCI Machine Learning Repository and Tehran Shahid Rajaei hospital which have 303 samples and 13 features [7] and 303 random patients with 59 features respectively.

When there are features in data, some of them are not useful even though they cause to get bad result therefore, the main aim of this study is to use a combined method to improve the classification and better feature selection which will lead to better diagnosis of heart disease. In this study, imperialist competitive algorithm with meta-heuristic approach is used to optimize the selection of important features in heart disease. This algorithm can provide a more optimal response for feature selection towards genetic and other optimization algorithms. After feature extraction, the features have been supplied to K-nearest neighbor (KNN) for classification proposes. Therefore, using the combination of these two methods can lead to improve the result of heart disease diagnosis and their different aspects. In other words, we are trying to improve classification accuracy on heart disease diagnosis. Using ICA and KNN was our idea which has never been done before. Moreover, according to the result section 4 the proposed method has achieved better result in comparison with other algorithms, which has had two advantages, first decreasing number of features, second, increasing classification accuracy. Objective of this research are as following:

- Data collection from new features about the heart disease.
- Prediction and classification of incidence of heart disease using the proposed method.

- Using new feature selection algorithms for the first time.
- Providing a new combined approach with higher accuracy.

This paper consists of four sections. The first section is introduction of topic, and introducing the aims of the study, steps needed to achieve these aims, and the results of the previous related studies in the field of heart disease. In the second part the studies about diagnosis of CVD using data mining are assessed. The third part explains the suggested methods in detail. In this section, first the suggested method and its required data are explained, and then it is assessed for heart disease using standard criteria. The last section explains the conclusion and suggests some works about the things that can be done to continue the optimization of accuracy, and feature selection.

2 Related works

There are a lot of medical data about different kinds of diseases nowadays. Medical centers collect these data for different purposes. One of the aims of using these data is study to obtain useful results and approaches about the diseases. The huge size of these data leads to confusion and impedes achieving appropriate results. A lot of studies have been performed about the data mining techniques for cardiovascular patients, which applied different methods such as decision tree and neural networks. Based on the wide spreading heart disease and its high costs, some studies have searched for solutions for prevention, efficient early treatment with lower costs, and as a result decreasing the number of tests.

In [8] a method has been presented to diagnose the heart disease using Particle Swarm Optimization as well as Neural Network Feed Forward Back-Propagation. The main focus was decreasing the number of features and costs. Also, in [9], a Naïve Bayes classification method was used based on the patients' history. In this study, first contemporary community rules were extracted to preprocess the data in order to obtain high quality patterns. Then, a pattern recognition algorithm was presented to identify contemporary community rules by the identification of most of temporal relationships in temporal abstractions (TAs). Finally, periodical contemporary community rules were combined as the feature for classification.

In [10] decision tree used for data mining in heart disease. One of the objectives of this research is to extract the hidden knowledge from huge datasets of heart disease in order to create a predictor model for heart disease using decision tree. Dataset that is used in this study consisted of 2346 unique samples, which included 1159 healthy individuals and 1187 patients with heart disease.

Feature selection [27] and removing the effects of redundant features in heart diseases can be a good solution to diagnose the disease. In this regard, [11] used medical tests as input in order to present a method, then a set of features are extracted by decreasing dimension, and a diagnosis system for heart disease is provided. This study has used Probabilistic Principal Component Analysis to extract prominent features with high effects. Then, Support Vector Machines (SVM) classifier that is based on the radial basis function (RBF) is used to classify data.

In [12], researchers tried to apply data mining methods to diagnose heart diseases. Different classification methods such as neural networks and decision tree are used to predict heart disease and to identify of its most important factors. Whereas, the discovery of relationship law to identify the effects of diet, life style, and environment on heart disease is used in this study. Clustering algorithms such as k-means algorithm have been used in the datasets of heart diseases which includes clinical data screening of the patients in order to diagnose sample, especially heart attack.

In [13], a combination of decision tree and Bayes network is used. The methodology that has been presented in this study in order to predict coronary disease included understanding and selection of the goal dataset, preparation and normalization, data mining and evaluation, and conclusion and the use of knowledge. Each step has some subsets. First, the samples are selected from patients' files in order to randomly predict, and the fields or the features of samples are identified and extracted based on the opinions of experts. In second stage of methodology, some new fields that are not directly mentioned in the patients' files are cleaned, normalized, and calculated. In the 3rd stage, the aim and data mining responsibility that is the use of prediction techniques in the mentioned research is determined. Then, based on the aim the mentioned algorithms are selected in order to predict. The algorithms used in this study included Bayesian network and decision tree as they are more perceptible for the experts. In the 4th stage, results are evaluated and accuracy and precision of the prediction method are assessed. The selected method for prediction is Bayesian network that describes the conditional relationship between variables. This network can create a probable model of the variables in order to determine the probability of occurrence of the feature sets.

In [1], Jabbar used the combination of Genetic algorithm and K-nearest neighbor. The assessment of the previous studies shows that the present methods have decreased the difficulties presented in the field of heart disease, however, feature selection with high accuracy and classification of them is a big challenge. In this study the authors tried to diminish the difficulties using combined methods.

3 Method

A data mining method is presented using a combination of imperialist competitive algorithm and K-nearest neighbor for the selection of feature and classification of heart disease data, which the combination of these two methods reaches to an optimized method. This process is shown in Fig. 1.

The proposed method is started with feeding input training data. Then the data are normalized. It is proven [14] that, in comparison with other methods, Z-score has been normalized heart data in the best way. Because it maintain range (minimum and maximum) and variance and standard deviation are introduced as a dispersion of the series by it. After that, ICA start to do feature extraction which is described section 3.1. It should be mentioned that, KNN are used as a fitness function inside ICA algorithm. Therefore ICA alone with KNN has made a combination method to classify the input data. There is no mathematic formula for the proposed method because it is combination of two already existing method and they are described in detail further.

The combination of these two methods in this study is more efficient in compare previous works. First, in order to perform the suggested method the data are loaded, and then imperialist competitive algorithm is used to select the appropriate features from the loaded data, which can classify them. In the following section it is completely explained that how imperialist competitive algorithm with meta-heuristic approach will operate.

3.1 Imperialist competitive algorithm for feature selection

Imperialist Competitive Algorithm (ICA) is used to select features in the diagnosis of heart disease. In this study, it is assumed that the number of features is specified, and the aim is just to find the best features that can increase the accuracy of diagnosis of heart disease. The number of the selectable features in the implemented tests is assumed to be equal to different datasets. Similar to other developmental optimization methods, imperialist competitive algorithm also begins with an initial population, which each member of the population is called a country. These countries are divided into two groups; the countries that are colony that are subordinated to a country, and colonialist countries, i.e. they are dominating some colonies. Each colonialist country dominates colonies based on their power, and finally the most powerful country is selected as the optimized point in optimization problems [15]. Figure 1 shows the proposed method.

Indeed, initialization is the determination of the initial population to start the optimization algorithm. Optimization problems mostly look for an optimized answer based on the problem's variables. Also, in this algorithm an array of variables of the problem that will be optimized are created. The created

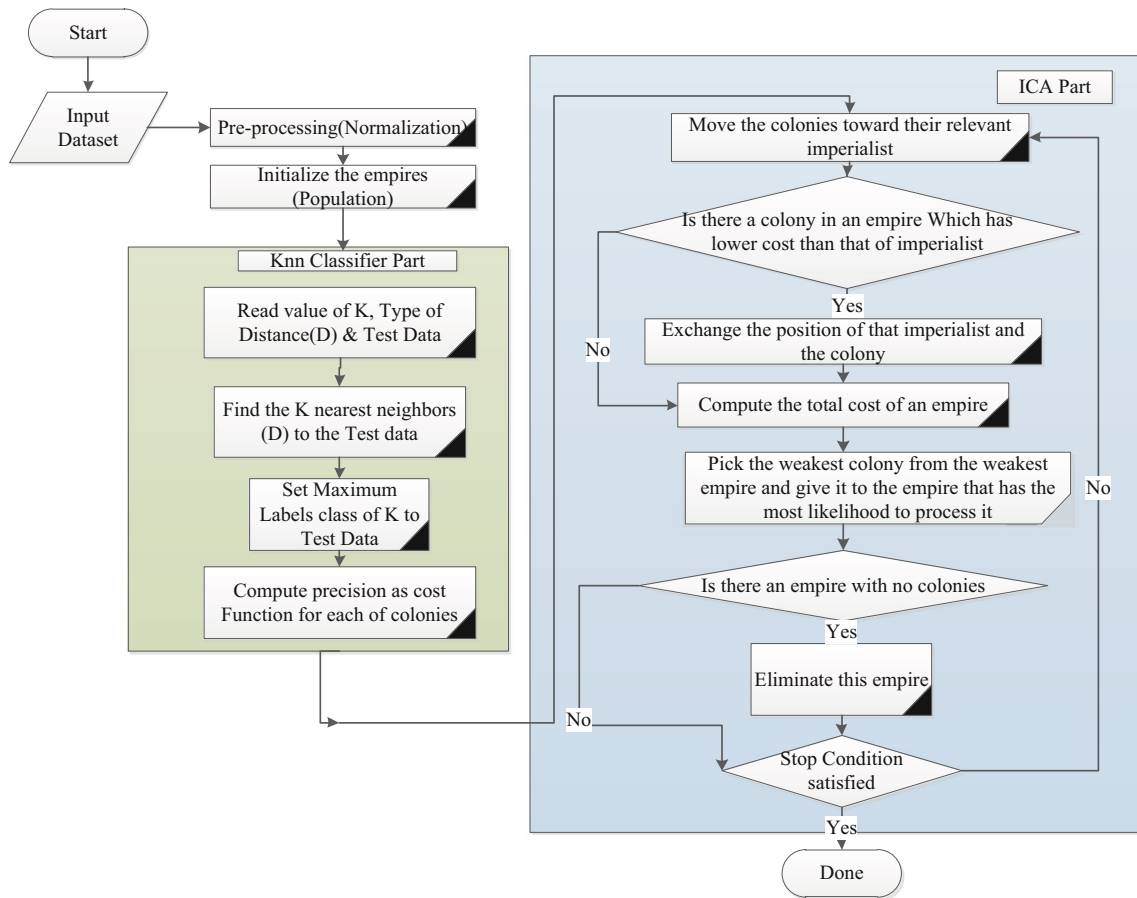


Fig. 1 A general schematic from the proposed method

array in the imperialist competitive algorithm is called a country. A country is in an array of $1 * N_{var}$. The magnitudes of the variables are shown in decimal form. The members of a country for heart disease in this algorithm can be considered as the features in this disease. A better understanding of this problem can be achieved from Fig. 2.

To design the features of heart disease with the highest accuracy in diagnosis, the obtained answers will have a stable output. For this problem a set of initial answers may be obtained, and it defines a country as follows:

$$Country_i = [Age_i, Gender_i, Clostrul_i] \tag{1}$$

To use these countries as the initialization of the algorithm, some initial countries have to be created. Therefore, the matrix of all countries is randomly formed by random magnitudes.

The cost of a country in this algorithm is calculated by the evaluation of function f in variables (p_1, p_2, \dots, p_n) as follows:

$$Cost_i = Fitness(Country_i) = Fitness(p_1, p_2, \dots, p_n) \tag{2}$$

Indeed, the cost function in this study is the accuracy of the classification of heart disease data, which the aim is to maximize the accuracy of the classification. To calculate the costs of a country, each feature of the classifier data is assumed as test data and the educational data. Then the accuracy or the cost function is calculated using these data. The sum of the answers of $ErrorValue = 1 - (count/n + Epsilon)$; is calculated as the cost of the country.

The best country (the best features with the highest accuracy) is what this study was looking for. The algorithm used in this study made an initial dataset and classified them into the

Fig. 2 Components features of a country

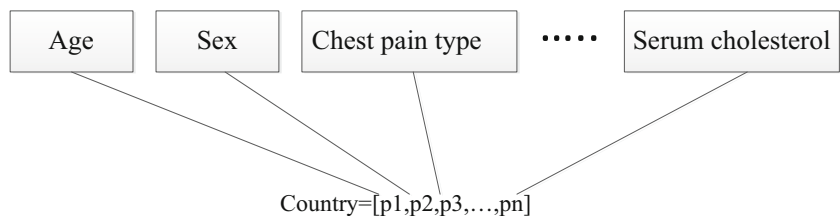


Table 1 Confusion Matrix

Predicted class	Actual class	
	Class 1	Class 2
Class 1	True positive (TP)	False positive (FP)
Class 2	False negative (FN)	True negative (TN)

form of the empires, and colonialist countries applied assimilation policy on the colonies and imperialist competitive is applied among the empires in order to find the best country. To start the imperialist competitive algorithm for heart disease, $N_{country}$ initial countries with random magnitudes were formed. The number of the best members of the population is shown by N_{imp} , and this member is selected as the emperor. The number of other countries (colonies) is shown by N_{col} . Each imperialist took some colonies based on their power. In this regard, after the calculation of the costs of all imperialists, the normalization cost of them was calculated using the following equation:

$$C_n = \max_i \{c_i\} - c_n \tag{3}$$

Equation 3 shows the power and the cost of an emperor. c_n is the cost obtained for imperialist n. $\max_i \{c_i\}$ is the maximum cost among all imperialists. And c_n is the cost spent for normalization of this imperialist. The more imperialist costs will have the less normalization costs. Colonies are divided among imperialists using the following equation:

$$P_n = \left| \frac{C_n}{\sum_{i=1}^{N_{imp}} C_i} \right| \tag{4}$$

$P_n = \frac{C_n}{\sum_{i=1}^{N_{imp}} C_i}$ In the other words, the ratio of the colonies of a normalized power is an imperialist that is directed by that imperialist. Therefore, the initial number of the colonies for an imperialist is calculated using the following equation:

Table 2 Required information to describe component of three criteria

Abbreviation	Description
TP	The number of samples of class C1 which has been correctly classified
TN	The number of samples of class C2 which has been correctly classified
FN	The number of samples of class C1 which has been falsely classified as C2
FP	The number of samples of class C2 which has been falsely classified as C1

Table 3 Explanation of features presented in the datasets

Features Name
Age
Sex
Chest pain type
Resting blood pressure
Serum cholesterol
Fasting blood sugar
Resting electrocardiographic results
Maximum heart rate achieved
Exercise induced angina
St depression induced by exercise relative
The slop of peak exercise at segment
Number of major vessels colored by fluoroscopy
Thal

$$N.C_n = |\text{Round}\{P_n \cdot (N_{col})\}| \tag{5}$$

In this equation, $N.C_n$ is the initial number of the colonies, and N_{col} is the total number of the colonies. $N.C_n$ is the number of the initial colonies that are selected randomly and are given to the imperialist n. The imperialist competitive algorithm starts after obtaining the initial form of all empires. The developmental process is in a cycle that will continue until meet a suspension condition. The power of the empires is calculated using the following equation:

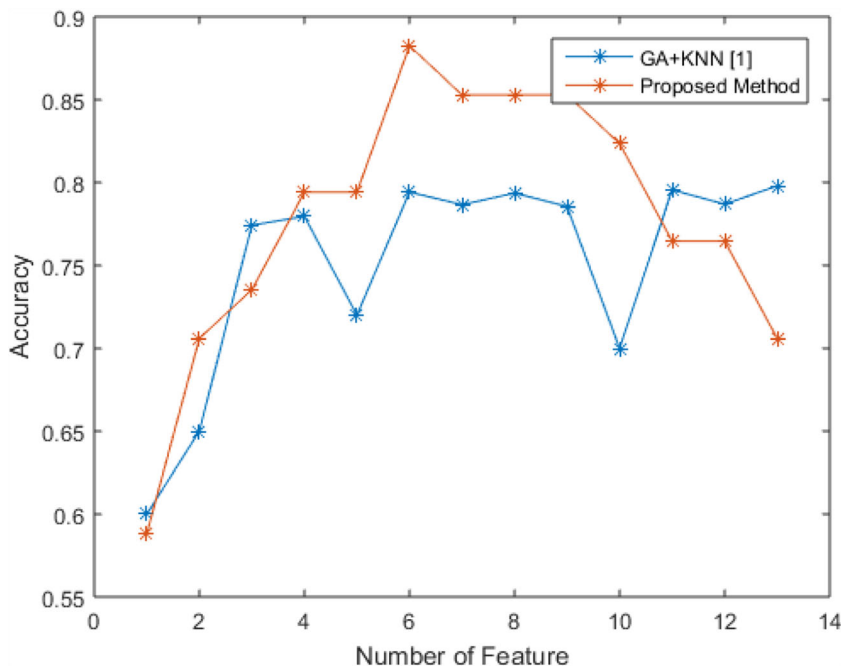
$$T.C_n = \text{Cost}(\text{imperialist}_n) \text{ mean}\{\text{Cost}(\text{colonies of mpire}_n)\} \tag{6}$$

During colonial competitive the weak empires lose their power gradually and then they are eliminated, and only one empire will be left that will rule the world and manage it. When this happens, the imperialist competitive algorithm stops as it reaches to the optimized point of the goal function. In this algorithm, sometimes the power of an imperialist becomes less than one of its colonies, the power of an empire is the set of the power of the imperialist plus some percentage of the total power of its colonies. The cost of an emperor is calculated using the following equation:

Table 4 Results of applying different algorithms on datasets

Methods	Accuracy	Sensitivity	Specificity
Naïve Bayes	85.7	88	83
Decision tree	76	80	74
Neural Network	85.7	88	83
SVM	85.7	93.3	80.8
Proposed Method	88.25	94.2	83.4

Fig. 3 Comparison of the accuracy of diagnosis with different number of features of the suggested method with method used in [1]



$$T.C_n = |\text{Cost}(\text{imperialist}_n) + \xi \cdot \text{mean}\{\text{cost}(\text{colonies of empire}_n)\}| \tag{7}$$

In this equation, the total cost of empire n is shown by $T.C_n$ and ξ is a positive magnitude between 0 and 1. If ξ is a small magnitude, the total cost of an empire goes toward its own costs. If ξ increases, the effect of the costs of other colonies becomes higher than the cost of the empire. Generally $\xi = 0.05$ leads to appropriate results. In this study ξ is equal to 0.1.

During the ongoing competition in the algorithm, if an empire cannot increase its power, it will go toward non-optimized answers and it will lose its colonies, therefore gradually will be eliminated from the competition. Then, other empires will take control of the eliminated empire’s colonies; hence, they become more powerful. Each optimization algorithm will end after its run. Imperialist competitive algorithm also follows the condition of convergence and ending the total number of repetition. This algorithm will achieve convergence to reach general optimization. This also happens when all of the empires are eliminated and turn into the colonies of a single empire. This single empire is called the only optimized answer.

Feature selection in this method is used for prediction and diagnosis of heart disease, prevention from implementation of

difficult procedures with high side-effects for diagnosis of heart disease, such as coronary angiography, as well as decreasing the number of implementation of useless tests, and therefore decreases the cost. In the next section, it will be explained that how the selected features are used for classification.

3.2 K-nearest neighbor algorithm for classification

K-nearest neighbor algorithm is a learning algorithm with observer that is used by imperialist competitive algorithm to classify

Table 6 Demographic features of the Shahid Rajaei hospital’s data [23]

Feature Name
Age
Weight
Sex
BMI (Body Mass Index Kg/m2)
DM (Diabetes Mellitus)
HTN (Hypertension)
Current Smoker
Ex-Smoker
FH (Family History)
Obesity
CRF (Chronic Renal Failure)
CVA (Cerebrovascular Accident)
Airway Disease
Thyroid Disease
CHF (Congestive Heart Failure)
DLP (Dyslipidemia)

Table 5 The comparison of diagnosis of heart disease to the latest studies

Methods	Accuracy
Reference 2014 [16]	81.51
Reference 2011 [17]	75.8
Reference 2008 [18]	72.55
Proposed Method	88.25

Table 7 Physical assessment features of the Shahid Rajaei hospital's data [23]

Features Name
BP (Blood Pressure)
PR (Pulse Rate)
Edema
Weak peripheral pulse
Lung rales
Systolic murmur
Diastolic murmur
Typical Chest Pain
Dyspnea
Function Class
Atypical
Nonanginal CP
Exertional CP (Exertional Chest Pain)
Low Th Ang (low Threshold angina)
FBS (Fasting Blood Sugar)
Cr (creatine)
TG (Triglyceride)
LDL (Low density lipoprotein)
HDL (High density lipoprotein)
BUN (Blood Urea Nitrogen)
ESR (Erythrocyte Sedimentation rate)
Hb (Hemoglobin)
K (Potassium)
Na (Sodium)
WBC (White Blood Cell)
Lymph (Lymphocyte)
Neut (Neutrophil)
PLT (Platelet)
EF (Ejection Fraction)
Region with RWMA (Regional Wall Motion Abnormality)
VHD (Valvular Heart Disease)
Rhythm
Q Wave

Table 8 Information profit of the selected features using feature selection method

Feature	Information Rate
LAD recognizer	1
LCX recognizer	0.921
RCA recognizer	0.750
Typical Chest Pain	0.622
Region RWMA2	0.270
Age	0.217
EF2	0.204
HTN	0.158
DM	0.138
Tinversion	0.119
ESR	0.097
Q wave	0.070
ST elevation	0.060
PR	0.059
BMI	0.041
Lymph	0.041
BP2	0.037
Dyspnea	0.028
HDL	0.022
CR2	0.021
WBC2	0.018
Weight	0.018
VHD	0.017
Function Class	0.017
Airway disease	0.014
HB	0.013
TG2	0.013
BBB	0.011
Na2	0.007
Sex	0.006
LVH	0.003
Hb2	0.001
FH	0.001

the selected features in this research. This algorithm is utilized for two objectives; to estimate the density function of learning data distribution, and to classify data based on the learning patterns. The second objective is used in this study.

4 Evaluation of the results

The dataset used in this study belongs to the UCI Machine Learning Repository and heart disease data of Tehran Shahid Rajaei hospital. Chosen UCI dataset included 303 samples and 13 features [7]. These features are the criteria for diagnosis of heart disease that have been determined by WHO. Dataset

of Shahid Rajaei hospital included 303 random patients with chest pain who were referred to the hospital. Data are selected by random selection of people. This data includes 59 features. The results are divided into two datasets in order to ease the evaluation process. These datasets are explained in the following sections.

4.1 Performance measure

Accuracy, Sensitivity, and Specificity are used to make comparison the performance of algorithms, as the most common metrics for assessment in medical field. In Table 1, Confusion

Table 9 Features for magnitudes with high reliability

Features	Number	Reliability
Q Wave	16	1
ST Elevation	14	1
Poor R Progression	9	1
CRF	6	1
Week Peripheral Pulse	5	1
Region with RWMA2	86	0.953
Typical Chest Pain	164	0.9390
Airway Disease	11	0.9090
DM	90	0.8888
Tinversion	90	0.8777
FBS2	84	0.8690
CR2	22	0.8636
BP2	48	0.8541
WBC	27	0.8518
Edema	12	0.8333
ST Depression	71	0.8309
EF2	197	0.8223
HTN	179	0.8212
Age2	177	0.8192
Lung Rales	11	0.8181
LVH	20	0.8
Ex-Smoker	10	0.8
CVA	5	0.8
TG2	62	0.7903
ESR	46	0.7826
Age2	238	0.7815
Current Smoker	63	0.7777
Function Class	92	0.7717
Na2	34	0.7647
Neut2	89	0.7640
FH	48	0.75
PLT2	12	0.75
HB	157	0.7324
LDL2	62	0.7258
HDL2	87	0.7241
Systolic Murmur	41	0.7073
Obesity	211	0.7061
DLP	112	0.7053
Dyspnea	134	0.6492
K2	37	0.6486
Thyroid Disease	7	0.5714
Atypical	93	0.4301
Diastolic Murmur	9	0.3333

Matrix is shown to describe component of three criteria which requires Table 2 information.

According to Table 1, Accuracy, Specificity and Sensitivity are calculated:

Table 10 Comparison of the efficacy of different algorithms except the three created features

Methods	Accuracy	Sensitivity	Specificity
Bagging SMO	89.43% ± 6.78%	91.67%	83.91%
Naïve Bayes	47.84% ± 6.35%	28.70%	95.40%
SMO	89.76% ± 7.31%	92.13%	83.91%
Neural Network	85.43% ± 7.02%	90.28%	73.560%
Proposed Method	91.53% ± 6.48%	93.13	81.40%

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (8)$$

$$Specificity = \frac{TN}{TN + FP} \quad (9)$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (10)$$

4.2 UCI dataset

Datasets for diagnosis of heart disease of Cleveland from data resource of California University are used in this study. Heart disease has many signs and symptoms. Finding some patterns of heart disease will help to recognize the reasons of cases. The database of heart disease is created by Long Beach medical center, Cleveland clinic foundation, and V.A in 1998 [7].

This database consists of 303 samples including 297 complete samples and 6 samples, which are with missed data. This database has 76 raw features however; all the tests are implemented just on 13 features. This database has 13 signs of the disease and 1 diagnosis feature. The goal field alludes to the presence of heart disease based on the signs of the patients, and it varies from 0 (no disease) to 1 (presence of disease). Table 3 shows the concept of each sign and symptom.

In the preprocess phase, the missed data are substituted by the mean scores of the same features in other samples. To train and test on this dataset, the dataset is divided into training and test subsets. So that, 80% of data are randomly put into training subset, and the rest are used to examine accuracy and precision of the model. All of these techniques were applied on the intended dataset in a similar situation. The advantages

Table 11 Comparison of efficacy of different algorithms with all features and three created features

Methods	Accuracy	Sensitivity	Specificity
Bagging SMO	90.10% ± 6.96%	91.67%	86.21%
Naïve Bayes	63.31% ± 8.01%	50%	96.550%
SMO	90.09% ± 6.49%	91.67%	86.21%
Neural Network	87.11% ± 6.05%	91.67%	75.86%
Proposed Method	91.03% ± 6.45%	91.67%	75.86%

Table 12 Comparison of algorithms efficacy and methods after selection of a subset of features without three created features

Methods	Accuracy	Sensitivity	Specificity
Bagging SMO	92.74% ± 6.43%	95.37%	86.21%
Naïve Bayes	55.37% ± 9.62%	38.89%	96.550%
SMO	93.39% ± 5.14%	95.37%	88.51%
Neural Network	87.13% ± 5.84%	90.28%	79.31%
Proposed Method	94.03% ± 6.85%	96.27%	90.36%

of the suggested imperialist competitive algorithm is its ability to optimize at the same level or even higher than other optimization algorithms in the time of facing different kinds of optimization issues, and to quick find the optimized answer toward other feature selection methods.

Table 4 shows that Bayesian algorithm is simple in accuracy, sensitivity, and negative prediction compare to other three algorithms. Neural network algorithm is better in positive prediction and feature, and although it possesses the least sensitivity, functions better than decision tree algorithm in negative prediction. The suggested method is more efficient in many aspects than other algorithms, especially in accuracy that leads to better diagnosis of heart disease. Figure 3 shows the results of comparing the suggested method to the method used in [1], which is a basic article. As shown in the Table 4, the suggested method has used better and smarter features for classification using imperialist competitive algorithm. This has led to more accuracy of classification with fewer features, and consequently less costs spent for tests and more accuracy of tests in diagnosing diseases.

The comparison of the accuracy of the suggested method to the latest works done for diagnosis of heart disease on the UCI dataset, which Table 5 shows that the suggested method is more efficient.

4.3 Datasets of Shahid Rajaei hospital

This dataset consisted of 303 random patients with chest pain who were referred to Tehran Shahid Rajaei hospital. This dataset also consisted of 59 features. Some of these features are not considered for diagnosis of coronary disease in the

Table 13 Comparison of the efficacy of the algorithms and methods after selection of a subset of the features with three created features

Methods	Accuracy	Sensitivity	Specificity
Bagging SMO	93.40% ± 5.53%	95.83%	87.36%
Naïve Bayes	75.51% ± 10.32%	67.59%	95.40%
SMO	94.08% ± 5.48%	96.30%	88.51%
Neural Network	88.11% ± 6.17%	91.20%	80.46%
Proposed Method	94.43% ± 6.25%	96.57%	91.06%

Table 14 Comparison of algorithms efficacy used on the datasets of heart disease obtained from Tehran Shahid Rajaei hospital

Methods	Accuracy (%)	Sensitivity (%)	Specificity (%)
Bagging SMO	89.43	91.67	83.91
Naïve Bayes	47.84	70.28	95.40
SMO	89.76	92.13	83.91
Neural Network	85.43	90.28	73.56
Proposed Method	91.34	93.35	85.36

previous studies. The features are divided into four groups consisted of demographic, physical assessment, ECG, and cardiac eco and laboratory tests. These features are shown in the following tables.

In this evaluation, first all the vessels are considered together. Each obstructed vessel is considered as CAD, and the rest are considered as normal vessels. Every obstructed vessel in the present study is considered separately in order to obtain better results. The importance of the features, selection methods, and creation of features will be explained subsequently. The datasets of this study are also assessed for algorithms that are sensitive to cost. As the cost of wrong diagnosis of CAD is higher than the cost of wrong diagnosis of the normal cases, the proposed method in this study is more beneficial than other methods.

4.4 Extraction of feature selection's results

In this section, utilizing feature selection using the suggested algorithm will be explained. The features that are extracted from this study using the feature selection algorithm are listed and explained in the table. The informative profit of every feature is presented in the table. Three of these features have the most information profit and have more power to separate Normal and CAD. Information Rate shows that how much a feature has significant effect on the classification task. Therefore the nearer to 1, the better features or important features.

Table 8 shows that among all the selected features using feature selection method, Typical Chest Pain, Region RWMA2, Age, EF, HTN, DM, Tinversion, ESR, Q wave, ST elevation, PR, BMI, Lymph, BP and dyspnea had the best information usefulness and power to separate normal from CAD samples Table 6 and 7.

In the beginning, the information usefulness of the features with high magnitudes is shown in Table 8, which were features that have higher reliability in diagnosis of a sample with CAD. Table 9 shows features based on reliability as decreasing trend line and also the number of samples which have high information usefulness. Comparing Tables 8 and 9 shows that features with more information profits may not be highly reliable. High reliability of features shows that how much a

Table 15 Evaluation of different methods for diagnosis of heart disease and other diseases presented in previous years

Methods	Results	Disease	Year	Author
Clustering Bayesian classification, Neural Networks Decision Tree, KNN	Decision tree showed dominant results and sometimes its results were similar to Bayesian classifier.	Cardiovascular disease prediction	2011	Jyoti Soni et.al (Soni, Ansari, Sharma, & Soni) [26]
J48 Decision Tree	The accuracy was 89% using Weka software and J48 method.	Blood donor	2012	Arvind Sharma et.al [24]
Neural Network, Association, Naïve Bayes, C4.5 decision tree algorithm	Decision tree obtained the highest accuracy (93.6%).	Breast cancer diagnosis	2012	Shweta Kharya [19]
K-means, Clustering	The accuracy after changing the samples was 97.2%	Medical data prediction	2012	Bushra M. Hussan [20]
Artificial Neural Network	The accuracy of Weka software for prediction of breast cancer was 97.77% and for prediction of the success of IVF is 70%.	Efficient plans in healthcare system	2013	M. Durairaj et.al [21]
association, rule clustering, prediction and Evaluation pattern	Decision tree provided better answers and higher accuracy for raw data and pattern finding	Efficient plans in medical data mining	2014	Aarti Sharma et.al [25]
Random Forest	High accuracy was obtained using Weka software and multiclass method	Prediction of primary tumors	2014	Mehak Naib et.al [22]

person may have CAD, and it will not make the assessment and check for other magnitude of features. While, features with more information profit show that an individual has CAD based on some magnitudes, and an individual is normal based on other magnitudes. Number caption in Table 9 means number of people which has that features and CAD with probability and reliability score. For example 16 people has Q wave which have CAD with reliability 1. Tables 8 and 9 are just shown that via using impact of the feature by getting especial metrics. However it has been shown that the hybrid method finds these features by looking into features and comparison with others in the dataset.

4.5 Assessment of available algorithm efficiency

The efficacy of different algorithms and methods are compared in Table 10. Three created features are not mentioned in this table.

The results of the comparisons show that SMO and Bagging methods have almost same magnitudes and both of them have 89% accuracy. The accuracy of neural network is 85%. Naïve Bayes has the lowest accuracy. The suggested method with 91% accuracy is the most accurate method compare to other methods. The sensitivity of Naïve Bayes method is more than its accuracy; therefore, it tends to diagnose normal people. But the suggested algorithm and other algorithms tend to diagnose patients. The results of all the methods and algorithms with three created features are shown in Table 11.

Table 11 shows that the accuracy and features of all algorithms have increased due to the three created features. Increment in accuracy and sensitivity is more in

Naïve Bayes method. The sensitivity of SMO method is partially decreased, but the amount of feature is increased. Table 12 compares the efficacy of algorithms and methods after selection of features' subset without three created features.

The accuracy of feature selection of all methods except for Naïve Bayes method is increased. These results show that the selection of specific features that affects the calculation of the accuracy of an algorithm have priority. Non-related features decrease the accuracy of the algorithms. The suggested method has more accuracy than other with 94% accuracy. The rest of the results obtained from the applied methods after selection of features' subset and the use of three created features are shown in Table 13.

Table 13 shows that the highest accuracy, feature, and sensitivity of algorithms occur when features are selected and created simultaneously. Comparing Tables 10 and 11 shows that the accuracy and sensitivity of all algorithms increased when three new features are created. The accuracy of Naïve Bayes method increased by 20%, which is the most increment in accuracy compared to other methods. The accuracy of the suggested algorithm (with 94.08% accuracy) is more than SMO.

Table 14 shows that the accuracy of SMO and Bagging methods is almost similar (89% accuracy). Neural network had 85% accuracy, and Naïve Bayes had lower accuracy. In all algorithms except Naïve Bayes method, the sensitivity is higher than the feature. Therefore, Naïve Bayes algorithm tends to diagnose normal people, and the suggested algorithm and other algorithms tend to diagnose patients. The methods in different years are routinely assessed in Table 15, and

it can be concluded whether other methods are needed in diagnosis of heart disease or not.

5 Conclusion

This paper studied data classification of heart disease and selection of the features. The aim of this study was making an automatic system to diagnose heart disease, classify the patients, to be able to use it in clinics. A combined method using imperialist competitive algorithm with meta-heuristic approach used to optimize feature selection, and K-nearest neighbor algorithm used to classify and diagnose heart disease. The aim of the suggested algorithm was to present a better classification with fewer features of data. The required tests implemented to assess the effectiveness of the suggested algorithm in the data classification. Data of these tests obtained from heart disease data of the UCI Machine Learning Repository and from Tehran Shahid Rajaei hospital. Different comparing criteria considered, and the role of every method in the improvement of efficacy of the combined algorithm determined. The future works in this field can focus on the use of the feature selection method for incomplete and missed data. The combination and the use of meta-heuristic algorithm for data classification and utilizing the proposed meta-heuristic algorithm in this study for medical uses are highly sensitive.

Compliance with ethical standards

Conflict of interest The authors declare no conflict of interest.

Ethical approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee in Iran. The dataset has been collected by Roohallah Alizadehsani [23] for his PhD dissertation.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

- Deekshatulu B, Chandra P. Classification of heart disease using K-nearest neighbor and genetic algorithm. *Procedia Technology*. 2013;10:85–94.
- Longe JL. *The Gale encyclopedia of diets: a guide to health and nutrition*. Two Volume Set: Gale; 2008.
- World Health Organization (2000). *The world health report 2000: health systems: improving performance*: world health Organization.
- Hosseini AA, Sobhani-Rad D, Ghandehari K, Benamer HT. Frequency and clinical patterns of stroke in Iran-systematic and critical review. *BMC Neurol*. 2010;10(1):72.
- Palaniappan, S., & Awang, R. (2008). *Intelligent heart disease prediction system using data mining techniques*. Paper presented at the 2008 IEEE/ACS international conference on computer systems and applications.
- Kieser R, Mulligan T. Analysis of echo counting data: a model. *Can J Fish Aquat Sci*. 1984;41(3):451–8.
- Blake C.L., Merz, C.J. (2014). UCI repository of machine learning databases, Available from: <http://www.ics.uci.edu/~mlern/MLRepository.html>.
- Feshki, M. G., & Shijani, O. S. (2016). *Improving the heart disease diagnosis by evolutionary algorithm of PSO and Feed Forward Neural Network*. Paper presented at the artificial intelligence and robotics (IRANOPEN), 2016.
- Orphanou, K., Dagliati, A., Sacchi, L., Stassopoulou, A., Keravnou, E., & Bellazzi, R. (2016). *Combining Naive Bayes Classifiers with Temporal Association Rules for Coronary Heart Disease Diagnosis*. Paper presented at the healthcare informatics (ICHI), 2016 IEEE international conference on.
- Tayefi M, Tajfard M, Saffar S, Hanachi P, Amirabadizadeh AR, Esmaeily H, et al. Hs-CRP is strongly associated with coronary heart disease (CHD): a data mining approach using decision tree algorithm. *Comput Methods Prog Biomed*. 2017;141:105–9.
- Shah SMS, Batool S, Khan I, Ashraf MU, Abbas SH, Hussain SA. Feature extraction through parallel probabilistic principal component analysis for heart disease diagnosis. *Physica A: Statistical Mechanics and its Applications*. 2017.
- Kangwanariyakul, Y., Naenna, T., Nantasenamat, C., & Tantimongcolwat, T. (2010). Data mining of magnetocardiograms for prediction of ischemic heart disease.
- Rajkumar A, Reena GS. Diagnosis of heart disease using datamining algorithm. *Global journal of computer science and technology*. 2010;10(10):38–43.
- Chubb H, Simpson JM. The use of Z-scores in paediatric cardiology. *Annals of pediatric cardiology*. 2012;5(2):179–84. <https://doi.org/10.4103/0974-2069.99622>.
- Atashpaz-Gargari, E., & Lucas, C. (2007). *Imperialist competitive algorithm: an algorithm for optimization inspired by imperialistic competition*. Paper presented at the Evolutionary computation, 2007. CEC 2007. IEEE Congress on.
- Gnanasoundhari, S., Visalatchi, G., & Balamurugan, M. (2014). A survey on heart disease prediction system using data mining techniques.
- Soni J, Ansari U, Sharma D, Soni S. Intelligent and effective heart disease prediction system using weighted associative classifiers. *International Journal on Computer Science and Engineering*. 2011;3(6):2385–92.
- Bhatia, S., Prakash, P., & Pillai, G. (2008). SVM based decision support system for heart disease classification with integer-coded genetic algorithm to select critical features. Paper presented at the Proceedings of the World Congress on Engineering and Computer Science, WCECS.
- Kharya, S. (2012). Using data mining techniques for diagnosis and prognosis of cancer disease. *arXiv preprint arXiv:1205.1923*.
- Hussan DBM. Data mining based prediction of medical data using K-means algorithm. *Basrah Journal of Science (A)*. 2012;30(1):46–56.
- Durairaj M, Ranjani V. Data mining applications in healthcare sector a study. *Int J Sci Technol Res*. 2013;2(10):29–35.
- Naib M, Chhabra A. Predicting primary tumors using multiclass classifier approach of data mining. *International Journal of Computer Applications*. 2014;96(8).
- Alizadehsani R, Habibi J, Bahadorian B, Mashayekhi H, Ghandeharioun A, Boghrati R, et al. Diagnosis of coronary arteries stenosis using data mining. 2012;2(3):153.

24. Sharma, A., & Gupta, P. C. (2012) Predicting the number of blood donors through their age and blood group by using data mining tool. *International Journal of communication and computer Technologies* 1.6, (2012): 6-10.
25. Sharma A, Sharma R, Sharma VK, Shrivatava V. Application of data mining—a survey paper. *International Journal of Computer Science and Information Technologies*. 2014;5(2):2023–5.
26. Soni, Jyoti, et al. "Predictive data mining for medical diagnosis: An overview of heart disease prediction." *International Journal of Computer Applications* 17.8 (2011): 43-48.
27. Zhang, Harry. "The optimality of naive Bayes." *AA* 1.2 (2004): 3.

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