Computers in Biology and Medicine 113 (2019) 103400

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

A Hybrid Risk Assessment Model for Cardiovascular Disease Using Cox Regression Analysis and a 2-means clustering algorithm

T. Vivekanandan[a,b,*], Swathi Jamjala Narayanan[a]

[a] School of Computing Science and Engineering, VIT University, Vellore, 632014, Tamil Nadu, India
[b] Department of Computer Science and Engineering, SITAMS, Chittoor, 517127, Andhra Pradesh, India

ARTICLE INFO

Keywords:
Cardiovascular disease
Risk assessment
Modified differential evolution
Cox regression
K-means clustering

ABSTRACT

Cardiovascular disease (CVD) refers to a state that indicates narrowed or blocked blood vessels, and it can lead to cardiac arrest, chest pain (angina) or stroke. CVD is a leading cause of silent massive heart attacks and is a major threat to life. The mere prediction of the presence or absence of CVD alone is inefficient in current scenarios. Rather, a major need has arisen for the prediction of CVD, the acquisition of knowledge about CVD and the assessment of the likelihood that an individual will experience cardiac arrest. The objective of establishing an individual CVD risk assessment has been attained in this paper using a hybrid model. The CVD of an individual is due to various controllable and uncontrollable factors. The computation and analysis of all these factors are difficult and time consuming. Only a few attributes are identified to be the most critical. This optimization of the critical features is performed using a modified Differential Evolution (DE) algorithm. The identified critical factors are sufficient to predict the presence/absence of CVD. In this paper, these identified critical features of individuals are considered using Cox regression analysis that evaluates the prevalence rates of the critical attributes. These individual prevalence rates together predict the cumulative prevalence ratios of the respective individuals. This cumulative prevalence ratio of an individual, along with the class attribute, is processed using the 2-means clustering technique to determine the risk of a particular individual developing CVD. The evaluation of the risk assessment model is carried out in this paper by calculating the prediction accuracy of the Cox regression analysis and the Davies–Bouldin (DB) index for 2-means clustering. The Cox regression analysis results in a 91% CVD prediction accuracy using the critical attributes and is comparatively higher than that of other models. The DB index of 2-means clustering with specific initial means for clusters of individuals with CVD is 0.282 and that for clusters of individuals without CVD is 0.2836, which are comparatively lower than those of the traditional k-means clustering algorithm.

1. Introduction

CVD is the leading cause of death worldwide according to the World Health Organization (WHO) [3]. It has been estimated that every year, 17.9 million deaths are due to CVD, which is 32% of the total deaths, and this number is projected to increase to 24 million by 2030 [7,8,12,21]. The number of deaths due to CVD in developing and underdeveloped countries is very high [10].

The exponential growth of medical data, compared to other domains, has resulted in the highly challenging task of handling and processing the data. The adoption of advanced technology in the medical field has been highly beneficial in several ways. In the present Information Technology era, the healthcare systems in developing and underdeveloped countries have developed models based on extensive treatments administered to only limited populations. Nearly three-fourths of the CVD deaths occur in these countries. These people have less access to effective healthcare systems that respond to their needs [41,42,59].

Most CVD events can be prevented by identifying the behavioural risk factors. Therefore, the prevention of CVD through disease awareness and the risk assessment of CVD events are vitally important. Prevention is based on individuals' lifestyles, such as physical activity, not smoking, healthy eating habits, etc. Risk assessment tools, such as the Framingham risk score [32,33,39,43], QRISK [33], European study [30], etc., have been used to prevent long term CVD. To further analyse the existence of CVD, the risk assessment of individuals based on their associated attributes is carried out. Risk assessment helps to assess the treatment decisions that reduce the incidence of CVD or increased the

* Corresponding author. SITAMS, D.K.Audikesavalu Marg, Tripati-Banglore Bye pass Road, Chittoor, 517127, AP, India.
E-mail addresses: stvanand@gmail.com, stvanand@sitams.org (T. Vivekanandan), swathi.jns@gmail.com (S.J. Narayanan).

https://doi.org/10.1016/j.compbiomed.2019.103400
Received 7 April 2019; Received in revised form 21 August 2019; Accepted 22 August 2019
Available online 27 August 2019
0010-4825/

rate of survival.

Treatment is then instituted based on the level of risk, which is determined by the estimated risk of any such event. In the conventional method, screening and treatment are primarily focused on single risk factors [9]. In recent times, the focus has moved towards considering moderate levels of numerous risk factors that can estimate the total CVD risk for an individual rather than the prominence of one risk factor [11].

Few risk factors that are identified as contributing to the disease are uncontrollable and modifiable factors. In adults, the risk of developing CVD is higher. The adoption of a healthy lifestyle can greatly help to postpone the onset or progression of disease. In addition, a healthy lifestyle will greatly reduce the risk of experiencing diseases [44].

This paper aims to identify the CVD risk levels of individuals based on the optimization of nine critical attributes. These critical attributes are optimized using the modified DE algorithm [5] from our previous work, and they are used as covariates for the Cox regression algorithm. The output of this process is used in the Fuzzy Analytical Hierarchy process (AHP) to determine the CVD risk, and the Artificial Neural Network (ANN) from our earlier work has been used in this research to evaluate the performance of the Cox regression. The individuals are further clustered in to four groups based on their estimated CVD risk.

2. Background study

The purpose of this work is to develop a risk assessment model for CVD. Some of the techniques that significantly contribute in achieving this goal are discussed below.

2.1. Modified DE, fuzzy AHP and ANN

The DE algorithm is a kind of evolutionary computing algorithm for optimization problems. Rainer Stron and Kenneth Price [1,2,4] designed 10 strategies for traditional DE, which are known for their own simplified structure, efficiency, speed and robustness. Out of these ten strategies, the seventh strategy has been proved to be the best [8].

This strategy has been modified by choosing four vectors from the current generation population vector to perform the 2-weighted difference and form a mutant vector. This process will be followed by a crossover function of the target vector and mutant vector to generate the trial vector. The trial vector will be compared with the objective function to process the next generation. This process will be repeated until the current generation is the same as the next generation. The output of this kind of modified DE algorithm provides the critical attributes that are responsible for causing CVD. Feature selection is the most widely used technique to form the subset of relevant features. This kind of modified DE has been adapted for optimal feature selection [5].

Multi-criterion decision making using the fuzzy AHP is one of the most powerful techniques that has been used in different fields, such as healthcare, business, industry, education, government, etc. The fuzzy AHP uses structured approaches for complex decisions using fuzzy set theory and the AHP. The pair-wise comparison matrix is formed for the multi-criterion decision making using derived weights. Further, the decision weights are processed for the decision alternatives to achieve the goal [48,61].

The ANN is a computing system that is inspired by a biological nervous system that is composed of neurons. The network connections are processed and interconnected as weights. The modelled network is trained using learning algorithms to classify the unknown labels that are the inputs of the network. ANNs are most often used for classification problems [48,60].

2.2. Risk analysis using Cox regression model

Cox regression is also known as the Cox proportional hazard model, which was first introduced in the 1970s by Cox, and it is still receiving great attention in survival analysis research [22,23,29]. Cox regression is one of the most popular and widely used statistical methods in medical research to predict the risk factors for various diseases [24–26]. The basic purpose of the Cox regression model is to assess the effects of various factors on survival. This model, which begins with multivariate analysis, allows for observing how the effects of several variables influence the occurrence of a particular event at a particular instant (for example, death, infection, etc.). This is commonly referred to as the hazard rate (i.e., the risk probability of an event) [27,28] in longitudinal studies and is referred as the prevalence rate in cross sectional studies [62].

2.3. K-means clustering

The process of grouping data into more similar objects is known as clustering. Different clustering methods that have been used in practice include partitioning clustering, pattern based clustering, density based clustering and constraint based clustering. Among the available clustering methods, the most widely used is k-means clustering [19]. This method is a kind of partitioning-based unsupervised learning algorithm that is used for data objects that are not assigned to a specific group or cluster. K-means clustering works faster when the value of k is smaller, and it produces tighter clusters compared to other clustering techniques. The objective of this algorithm is to find k number of clusters based on the input variable k. It is the most widely used concept in broad application areas, such as data mining, statistics, biology, knowledge discovery, machine learning, data compression, vector quantization, pattern recognition, Image processing, computer graphics, etc. [6,17].

The algorithm works by choosing k initial objects, which form the means of the initial clusters. Furthermore, the remaining data objects are assigned to most similar clusters based on the Euclidian distance. The new mean will be computed for each cluster, and the process continues until there is no change in the cluster means [18].

3. Related work

Risk assessment in healthcare has been practised for several decades. Risk identification and assessment in the current healthcare sector is one of the most pertinent approaches to perform a complete analysis of the causes of adverse incidents [13,14,31].

Risk assessment for CVD is indispensable for its prevention, and it also helps the experts to implement appropriate treatments. Risk assessment tools, such as the Framingham Risk Score (FRS) [33,39], QRISK score [34], Systematic COronary Risk Evaluation (SCORE) [47], WHO/ISH Risk prediction charts [35], ACC/AHA [37,38], JBS3 risk calculator [36], JBS3 [36], etc., are primarily used for risk assessments. The individual's lifestyle, physical activity, tobacco use, Body Mass Index (BMI), etc. greatly influence the primary stage of prevention. The individuals who are already diagnosed, diagnosed in later stages or at the verge of developing CVD are greatly limited in the primary stage of prevention [40,46]. This paper mainly focuses on the risk assessment of individuals who are diagnosed with CVD in the secondary stage of prevention. This secondary stage of prevention mainly focuses on the risk assessment of a particular disease, which enables an individual to assess the risk of developing the disease in the near future. In the previous literature, several risk assessment models have been developed to predict the mortality rates of different heart diseases [32].

In the last two decades, several data mining techniques and machine learning algorithms have been used for the risk assessment and prediction of CVD. Algorithms, such as support vector machines, decision trees, the k-nearest neighbours, the Bayesian classifier, the Bayesian decision tree, the ANN, fuzzy logic, etc., have been proposed for CVD risk prediction [48]. The most common method in these studies is that a binary output class was adopted to predict the presence or absence of heart failure. In previous works, several attempts have been made to

progress and develop complex risk assessment methods to improve CVD risk assessment and prediction [5,15,20,30]. The assessment of CVD risk is the key for the prevention of CVD in the primary and secondary stages. Furthermore, to identify and greatly enhance the implementation of effective treatments, the existing literature has generated additional interest in developing CVD risk assessment and prediction algorithms.

Cox regression [22,23,29], although invented in the 1970s, is still receiving significant interest in many domains and is extensively used for survival analysis [45]. Long term risk prediction models for CVD were developed using the Cox proportional hazard regression [27,44]. Traditional risk factors, including age, sex, CVD subtype, hypertension, diabetes, total cholesterol, body mass index, current smoking status, current drinking status and physical activity, were used for the analysis. A multi-level risk assessment [15] technique has been proposed to predict heart failure at five risk levels, including no risk, low risk, moderate risk, high risk and extremely high risk. The C4.5 decision tree classifier was used to predict heart failure. The traditional methods for diagnosing CVD risks have analysed the patients' medical histories and have utilized all the attributes [16].

The proposed approach focuses on risk assessment by considering the hazard rate for each and every value of a particular attribute using Cox regression analysis. Thus, the nine critical attribute values of an individual are replaced by the corresponding hazard rates using the Cox regression. The sum of these values results in the cumulative hazard function of the individual, which forms the basis of risk assessment. Therefore, the total CVD risk of individuals should be considered as the prime measure for the risk assessment.

4. Materials and methods

4.1. Prediction of cumulative prevalence ratio

CVD is associated with various factors, and the critical attributes have been optimized using a modified DE algorithm. The output from our previous work [5] has been used in this work. The effect of each critical attribute upon the occurrence of CVD has been investigated using a Cox regression, which is also known as the Cox proportional hazard regression. In this work, the Cox proportional hazard regression is used to assess several risk factors and their impacts on CVD [26,27].

The covariates of the Cox regression algorithm are the optimized critical attributes, as shown in Fig. 1. If the difference between the values of a covariate for different individuals changes with time, then the covariate is time dependent. If the values of a covariate do not change with time, then the covariate is fixed [13]. Here, the values of the critical attributes at a fixed time have been given to evaluate the hazard function, which is the cumulative prevalence ratio in a cross sectional study [62].

In this paper, the cumulative prevalence ratio represents the probability that an individual is diagnosed with CVD. The mathematical representation for the hazard function [49], which is equivalent to the cumulative prevalence ratio in our work, is given as

$$\mathrm{h(t)} = \lim_{\Delta T \to 0}\left(\frac{P(T \le t < (T + \Delta T) | T \le t)}{\Delta T}\right) \tag{1}$$

$$\mathrm{h(t)} = \lim_{\Delta T \to 0}\left(\frac{F(T + \Delta T) - F(T)}{\Delta T}\right) \tag{2}$$

$$h(t) = \frac{f(t)}{s(t)} \tag{3}$$

$$s(t) = 1 - F(T) \tag{4}$$

Here, s(t), which is the survivor function, is the probability that CVD is absent in an individual.

The cumulative prevalence ratio, H (t), is determined by summing up the prevalence rates of all critical attributes of a particular individual, and it is given as

$$H(t) = \int_0^T h(u)du \tag{5}$$

Here, time T is fixed and, hence, H (t) = H, which represents the cross sectional analysis.

4.2. Risk based clustering model

The cumulative prevalence ratio, H, is determined using the Cox regression algorithm and forms the basis for the further clustering of individuals. The K-means clustering technique has been used for this purpose, since it has good efficiency and performance. In the traditional K-means clustering technique, the initial mean is randomly chosen. Due to this random selection of the initial mean, the quality and accuracy of the final clusters might differ. To eliminate this issue, an improvement has been made in the traditional k-means clustering technique for the selection of the initial mean.

In this paper, the 2-means clustering technique has been separately applied to the dataset of individuals with CVD and the dataset of individuals without CVD. Instead of the random selection of the initial mean, the actual mean of the data is taken as the initial mean, as given below:

$$M_{A1} = \sum_{i=1}^{n} H_{Ai} \tag{6}$$

$$M_{U1} = \sum_{i=1}^{k} H_{ui} \tag{7}$$

where

M_{A1} = Initial mean of CVD affected individuals,
n = No. of CVD affected individuals
H_{Ai} = Cumulative prevalence ratio of the individuals with CVD
M_{U1} = Initial mean of the individuals without CVD, k = No. of individuals who are without CVD, and
H_{Ui} = Cumulative prevalence ratio of the individuals without CVD.

Then, each dataset is divided into two clusters based on the distance between the cluster elements and the calculated initial means, D_{Ai} and D_{Ui}.

$$D_{Ai} = |H_{Ai} - M_{A1}| \tag{8}$$

$$D_{Ui} = |H_{Ui} - M_{U1}| \tag{9}$$

D_{Ai} = Distance between the cumulative prevalence ratio of an individual with CVD and its initial mean, and
D_{ui} = Distance between the cumulative prevalence ratio of an individual without CVD and its initial mean.

$\mathrm{H_{Ai}}$ ε C_1 if $D_{Ai} \le M_{A1}$; otherwise, H_{Ai} ε C_2
Similarly.
H_{ui} ε C_3 if $D_{Ui} \le M_{U1}$; otherwise, H_{Ui} ε $C_4$where

$C1$ = Cluster with high risk patients,
$C2$ = Cluster with critical risk patients,
$C3$ = Cluster with low risk patients, and
$C4$ = Cluster with medium risk patients.

The separate means for each cluster are calculated by using the following expressions.

$$M_{A2} = \sum_{i=1}^{n} H_{Ai} \ of \ C1 \tag{10}$$

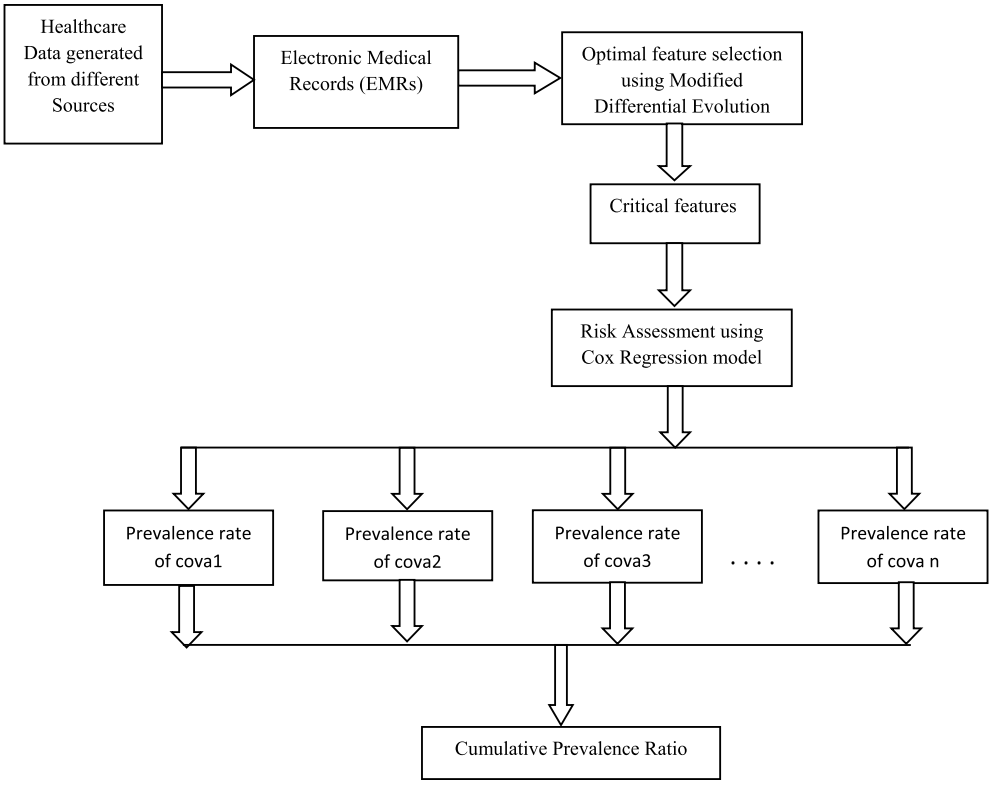

Fig. 1. Risk Assessment model using Cox regression model.

$$M_{A3} = \sum_{i=1}^{k} H_{Ai} \; of \; C2 \tag{11}$$

$$M_{U2} = \sum_{i=1}^{n} H_{Ui} \; of \; C3 \tag{12}$$

$$M_{U3} = \sum_{i=1}^{k} H_{Ui} \; of \; C4 \tag{13}$$

M_{A2} and M_{A3} are the respective means of the cumulative prevalence ratios of *C1* and *C2*, which are the individuals with CVD. M_{U2} and M_{U3} are the respective means of the cumulative prevalence ratios of *C3* and *C4*, which are the individuals without CVD.

Later, the Euclidean distance is calculated as

$$E1_{Ai} = |H_{Ai} - M_{A2}| \tag{14}$$

where H_{Ai} is the cumulative prevalence ratio of *C1* individuals.

$$E2_{Ai} = |H_{Ai} - M_{A3}| \tag{15}$$

The cumulative prevalence ratio H_{Ai} of the *C1* and *C2* individuals is calculated as

H_{Ai} ε *C1* if $E1_{Ai} \leq E2_{A1}$; otherwise, H_{Ai} ε *C2*.

In the same way, the clusters are reshuffled for the individuals without CVD.

$$E1_{Ui} = |H_{Ui} - M_{U2}| \tag{16}$$

and

$$E2_{Ui} = |H_{Ai} - M_{U3}| \tag{17}$$

Here, H_{Ui} is the cumulative prevalence ratio of all individuals in *C3* and *C4*.

H_{Ui} ε *C3* if $E1_{Ui} \leq E2_{Ui}$; otherwise, H_{Ui} ε *C4*.

Again, the 2-means are calculated for the newly formed clusters and the iterations continue until the means of the current iteration are equal to the calculated 2-means of the previous iteration. When the above condition is satisfied, the 2-means clustering algorithm is halted, and the following two clusters are formed for the individuals with CVD:

i. Individuals in the high risk category, and
ii. Individuals in the critical risk category.

In a similar way, the individuals without CVD are further grouped into the following two clusters:

i. Individuals in the low risk category, and
ii. Individuals in the medium risk category.

As shown in Fig. 2, the cumulative prevalence ratios of individuals with CVD and without CVD are calculated in parallel to form the above clusters.

5. Datasets and implementation

In this paper, a dataset from a population living near Cleveland, which is available in the online repository of the University of California, Irvine (UCI), is used [16]. The dataset contains 303 data samples of individuals with some missing values. For this study, 300 individuals' data samples with 13 diagnostic attributes are extracted and used. The details of the dataset are given in Table 1.

Along with the above set of attributes, the class attribute (diagnosis of CVD) that is available in the dataset is also taken from the UCI repository. To predict CVD and evaluate the results, the dataset that is described above has been used.

Initially, the prevalence rate of each value of an attribute is calculated using the Cox regression algorithm. The prevalence rates corresponding to the values of all critical attributes of an individual are summed up to arrive at the cumulative prevalence ratio.

The 2-means clustering technique is applied to the cumulative prevalence ratio of individuals, and it results in four clusters, namely, low risk individuals, medium risk individuals, high risk individuals and critical risk individuals.

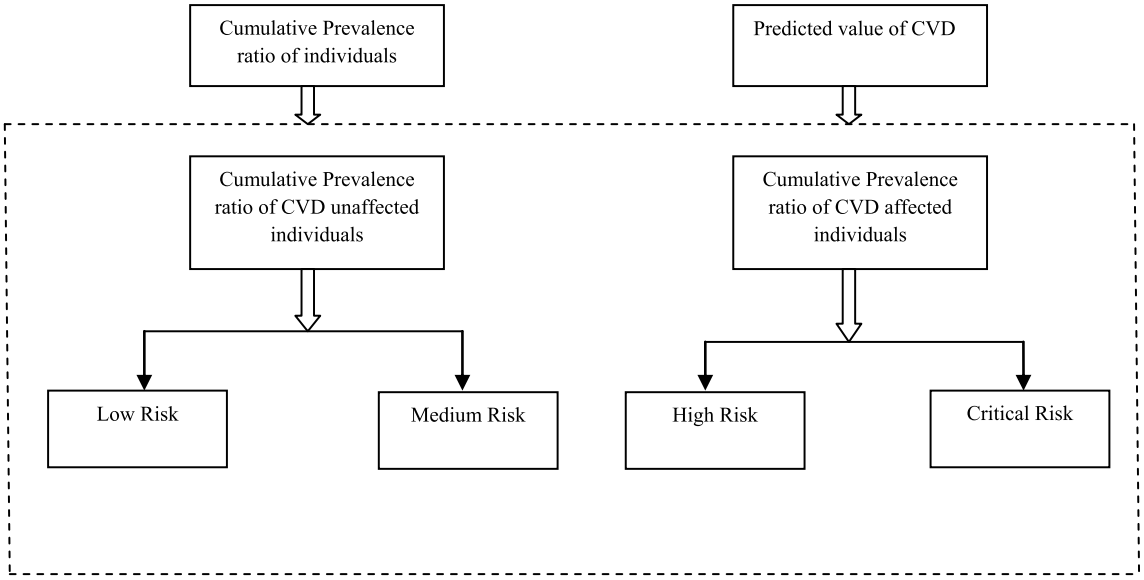

Fig. 2. 2-means clustering.

Table 1
CVD dataset description.

S No	Features	Notation	Description	Attribute Type
1.	Age	AGE	age in years	Young Medium Old Very Old
2.	Sex	SEX	Male/Female	Male Female
3.	Chest Pain Type	CPT	Chest Pain Type induced by deoxygenized blood in the heart	Typical Angina Atypical Non-Angina Asymptomatic
4.	Resting Blood Pressure (in mm Hg)	RBP	Blood Pressure taken when a patient is in a relaxed state	Low Medium High Very High
5.	Serum cholesterol in mg/dl	SER	Blood cholesterol level	Low Medium High Very High
6.	Fasting Blood Sugar	FBS	Blood glucose level diagnosed with empty stomach	True False
7.	Resting Electrocardiographic Results	ECG	ECG results taken at a resting state	Normal ST-T abnormal Hypertrophy
8.	Maximum heart rate achieved	MHR	Maximum number of heart beats calculated per minute	Low Medium High
9.	Exercise-induced angina	EIA	Reaction induced in a part of the heart due to the absence of oxygen in the blood while a patient is exposed to severe physical work	True False
10.	Old peak	OPK	ECG variation obtained between the tests are taken while a patient is resting and then exposed to exercise, respectively. ST depression induced by exercise relative to rest	Low Risk Terrible
11.	Peak Exercise slope	PES	ST segment slope that appears in a patient's ECG who is exposed to peak exercise	Upsloping Flat Downsloping
12.	Number of major vessels coloured by fluoroscopy	VCA	X-ray used to view the major vessels of the heart	Fluoroscopy-0 Fluoroscopy-1 Fluoroscopy −2 Fluoroscopy-3
13.	Thallium Scan	THA	A method used to obtain information about the supply of blood to the heart muscle by examining the heart of a patient.	Normal Fixed defect Reversible defect

Table 2
Prevalence rate of each critical attribute.

	B	SE	Wald	Df	Sig.	Exp (B)	95.0% CI for Exp (B)	
							Lower	Upper
AGE	0.032	0.020	2.587	1	0.108	1.032	0.993	1.073
CPT	1.040	0.285	13.355	1	0.000	2.829	1.620	4.940
RBP	−0.004	0.010	0.125	1	0.724	0.996	0.977	1.016
MHR	−0.021	0.007	8.749	1	0.003	0.980	0.966	0.993
EIA	0.894	0.305	8.584	1	0.003	2.444	1.344	4.445
OPK	0.292	0.107	7.494	1	0.006	1.339	1.086	1.650
PES	0.500	0.209	5.741	1	0.017	1.648	1.095	2.481
VCA	0.479	0.139	11.859	1	0.001	1.614	1.229	2.119
THA	0.410	0.094	18.932	1	0.000	1.507	1.253	1.813

6. Results and discussion

6.1. Prediction of prevalence rate of critical attributes

Cox regression analysis has been carried out using the values of the selected critical attributes. These critical attributes' values are analysed for CVD. The inferred results provide us with a hierarchical ranking of these critical attributes based upon their key roles in developing CVD.

In the above Table 2, Exp (B) determines the increase in the prevalence rate of CVD with respect to the rate of change of an attribute value, as per column B. Hence, these Exp (B) values determine the importance of the particular attribute in developing CVD.

Based on the prevalence rates of the critical attributes, the hierarchical rankings of the critical attributes are given in Table 3, and the mean prevalence rates of the critical attributes are given in Fig. 3.

6.2. Determination of cumulative prevalence ratio of an individual

Chest pain type has been found to have the highest weight and is the most critical attribute in determining CVD using Cox regression analysis. In a similar way, the prevalence rates are determined for all the critical attributes using Cox regression analysis.

The cumulative hazard function of the Cox regression analysis determines how the hazard rate depends upon the associated attributes, which is given as

$$H = \sum_{i=1}^{n} h_i \tag{18}$$

where h_i is the hazard rate of the ith critical attribute of an individual. In this study, a constant risk period has been assigned, and hence, the cumulative hazard function is equal to the cumulative prevalence ratio [62]. This cumulative prevalence ratio is used to predict the CVD risk ratio of an individual. Based upon these 'H' values, the individuals are categorized into four groups, namely, low risk, medium risk, high risk and critical risk, depending upon their CVD severity.

Table 3
Ranking of Critical attributes.

Critical Attributes	Mean Prevalence rate	Attribute Ranking
CPT	2.829	1
EIA	2.444	2
PES	1.648	3
VCA	1.614	4
THA	1.507	5
OPK	1.339	6
AGE	1.032	7
RBP	0.996	8
MHR	0.980	9

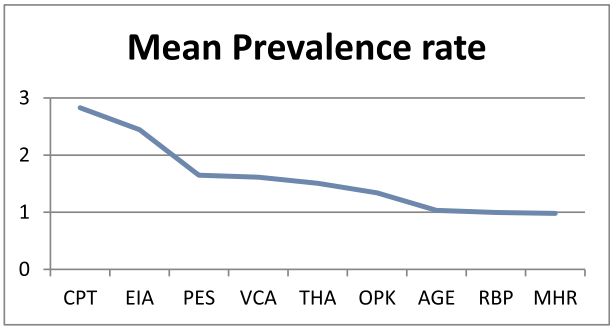

Fig. 3. Mean Prevalence rates of the critical attributes.

6.3. Performance evaluation of Cox regression analysis

A performance evaluation of the Cox regression analysis has been carried out using multiple linear regressions, estimating the AUC using the ROC curve, estimating the prediction accuracy using the confusion matrix and comparing the proposed model with other existing models.

6.3.1. Performance evaluation using multiple linear regressions

The cumulative prevalence ratio of an individual based on the prevalence rate of the critical attributes and all attributes is used to assess the presence of CVD. Multiple linear regressions have been calculated to statistically analyse the significance of the Cox regression model using 9 critical attributes compared with 13 attributes.

It is inferred from Tables 4 and 5 that the multiple linear regression has a statistically good fit. The multiple linear regression shows that five independent variables contribute to CVD, and they are statistically significant at the 1% and 5% levels. The table indicates that the coefficients of respondents' AGE, SEX, CPT, RBP, CHOL, FBS, ECG, MHR, EIA, OPK, PES, VCA, and THA are positively associated with CVD. Further, this result indicates that the contributions of the attributes are statistically significant.

It is inferred from Tables 6 and 7 that the multiple linear regression of 9 critical attributes has a statistically good fit. This result shows that three independent variables contribute to CVD and are statistically significant at the 1% and 5% levels. The table indicates that the critical attribute coefficients of respondents' AGE, CPT, RBP, MHR, EIA, OPK, PES, VCA, and THA are positively associated with CVD. Further, this result indicates that the contributions of these 9 critical attributes are statistically significant and further implies that their influences are stronger than those of the other attributes.

Table 4
Multiple linear regression of the Cox regression analysis using 13 attributes.

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	−7.066	2.604		−2.714	0.008
AGE	0.050	0.302	0.015	0.167	0.868
SEX	0.325	0.196	0.128	1.662	0.100
CPT	0.261	0.095	0.226	2.758	0.007
RBP	2.047	1.265	0.126	1.618	0.109
CHOL	9.187	4.080	0.177	2.252	0.027
FBS	−0.260	1.620	−0.012	−0.160	0.873
ECG	0.394	0.196	0.141	2.006	0.048
MHR	0.131	0.142	0.076	0.917	0.362
EIA	0.049	0.145	0.028	0.338	0.736
OPK	0.356	0.169	0.191	2.106	0.038
PES	0.030	0.187	0.013	0.161	0.872
VCA	0.161	0.122	0.114	1.319	0.191
THA	0.359	0.083	0.384	4.316	0.000

a Dependent Variable: CVD.

Table 5
Summary table for the Cox regression analysis using 13 attributes.

Model	R	R Squared	Adjusted R Squared	Std. Error of the Estimate	Change Statistics				
					R Squared Change	F Change	df1	df2	Sig. F Change
1	0.784(a)	0.614	0.556	0.332	0.614	10.541	13	86	0.000

a Predictors: (Constant), THA, RBP, ECG, FBS, CHOL, PES, VCA, MHR, SEX, CPT, EIA, AGE, and OPK.
b Dependent Variable: CVD.

Table 6
Multiple linear regression of the Cox regression analysis using 9 critical attributes.

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	−1.489	0.781		−1.908	0.060
AGE	0.206	0.290	0.059	0.709	0.480
CPT	0.312	0.096	0.270	3.253	0.002
RBP	1.585	1.253	0.097	1.265	0.209
MHR	0.091	0.143	0.053	0.633	0.528
EIA	0.151	0.145	0.086	1.041	0.301
OPK	0.296	0.170	0.158	1.746	0.084
PES	0.063	0.190	0.028	0.330	0.742
VCA	0.137	0.126	0.097	1.091	0.278
THA	0.349	0.080	0.373	4.340	0.000

a Dependent Variable: CVD.

6.3.2. Estimation of the AUC using the ROC curve for the Cox regression analysis

In this experiment, we obtained the ROC chart and AUC to further validate the effectiveness of the Cox regression analysis using the critical attributes [63].

It is inferred from the above Figs. 4 and 5 and Tables 8 and 9 that the AUC for the Cox regression using the critical attributes is 0.913, which is close to the AUC of the Cox regression analysis using all the attributes of 0.926. Since the AUC values are greater than 0.9, the model's excellence has been proved.

6.3.3. Evaluation of CVD prediction accuracy

The prediction accuracy of the Cox regression analysis is further calculated based on the following confusion matrix (see Table 10).

The confusion matrix elements are as follows:

1. TP, True Positive - Patients diagnosed with CVD who truly have CVD;
2. TN, True Negative - Patients not diagnosed with CVD who truly do not have CVD;
3. FN, False Negative - Patients not diagnosed with CVD who truly have CVD; and
4. FP, False Positive - Patients diagnosed with CVD who truly do not have CVD.

$$Accuracy = \frac{TP + TN}{Total\ No.\ of\ patients} \quad (19)$$

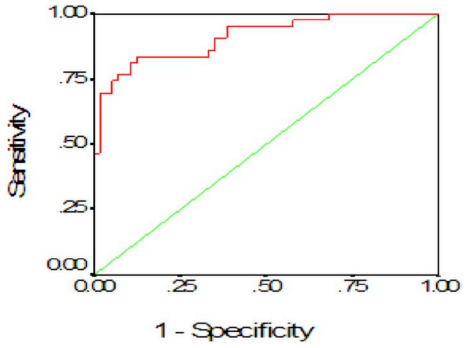

Fig. 4. ROC curve of the Cox regression analysis using the critical attributes.

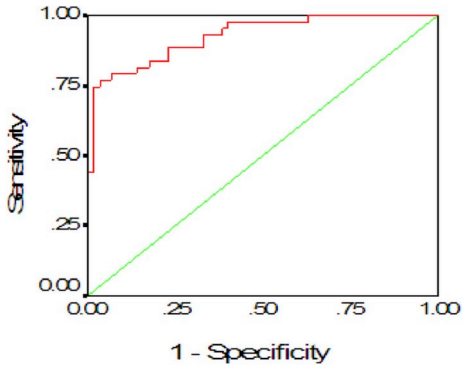

Fig. 5. ROC curve of the Cox regression analysis using all attributes.

Table 8
AUC for the Cox regression analysis using the critical attributes.

Area	Std. Error (a)	Asymptotic Sig. (b)	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.913	0.029	0.000	0.856	0.970

a Under the nonparametric assumption.
b Null hypothesis: true area = 0.5.

FP and FN are incorrect classifications. Hence, the accuracy of the Cox analysis is determined by the correct CVD predictions, as given below.

It is inferred from Table 11 that the prediction accuracy of the Cox regression analysis is 91% when using the 9 critical attributes, and it is

Table 7
Summary table for the Cox regression analysis using 9 critical attributes.

Model	R	R Squared	Adjusted R Squared	Std. Error of the Estimate	Change Statistics				
					R Squared Change	F Change	df1	df2	Sig. F Change
1	0.751(a)	0.564	0.521	0.345	0.564	12.942	9	90	0.000

a Predictors: (Constant), THA, RBP, PES, MHR, AGE, CPT, EIA, VCA, and OPK.
b Dependent Variable: CVD.

Table 9
AUC for the Cox regression analysis using all attributes.

Area	Std. Error (a)	Asymptotic Sig. (b)	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.926	0.026	0.000	0.875	0.976

a Under the nonparametric assumption.
b Null hypothesis: true area = 0.5.

Table 10
Confusion matrix.

Output value	1	0
1	TP	FN
0	FP	TN

Table 11
Accuracy of the Cox regression analysis using various input-attributes.

Performance evaluation	No. of attributes used for the cox analysis	
	9 Critical attributes	13 Attributes
Accuracy	0.91	0.88

Table 12
Comparison of the prediction accuracy with other existing models.

S·NO	Method	Accuracy
1.	Weighted fuzzy rules (2012) [50]	57.85%
2.	Logistic regression (2008) [51]	77%
3.	Binary Partial swarm optimization and rough sets based attribute reduction + Naive Bayes (2013) [52]	79.6%
4.	Binary particle swarm optimization and genetic algorithm + support vector machine (2010) [53]	81.46%
5.	Chaos firefly and Rough set based attribute reduction + ANN (2015) [54]	81.5%
6.	Three phase model + ANN technique (2013) [55]	88.89%
7.	ANN ensemble based model (2009) [57]	89.01%
8.	**Our proposed method**	**91%**

88% when using all 13 attributes for the analysis. It is therefore inferred that the 9 critical attributes are comparatively proved to be the best inputs for the Cox regression to predict CVD, and they are further utilized to cluster individuals based on their risk levels.

6.3.4. Comparative analysis

To compare the effectiveness of the proposed model with other existing models, a comparison of the CVD prediction accuracies is carried out (see Table 12).

6.4. Risk based clustering

The predicted CVD data and the cumulative prevalence ratio of each individual are the inputs of the 2-means clustering algorithm with a specific initial mean. The clustered output of the algorithm gives valuable information for an individual regarding his classification in the four clusters based on his risk level.

It is implied from the above Table 13 and Fig. 6 that the number of iterations is lower for 2-means clustering that uses a specific initial mean compared with that of the standard k-means clustering with k = 2.

Table 14 and Fig. 7 infer that the output clusters are almost the same for the standard k-means clustering with k = 2 and the 2-means clustering with a specific initial mean for the individuals with CVD.

Table 13
Comparison of the number of iterations of K-means using a randomly selected initial mean with K = 2 and 2-means clustering using the mean of the entire cluster as the initial mean.

CVD Data	No. of Iterations	
	K-means clustering with randomly selected initial mean	2-means clustering with specific initial mean
CVD affected dataset	4	3
CVD unaffected dataset	6	5

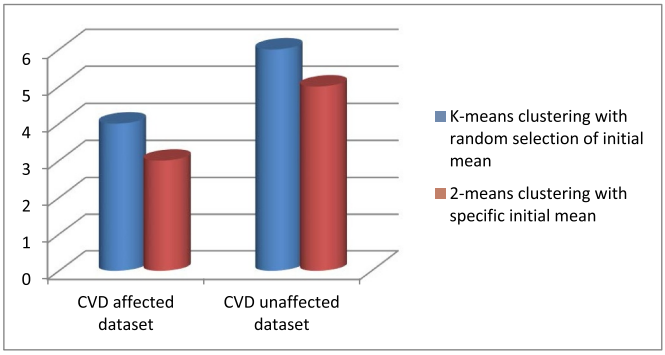

Fig. 6. Comparison of the number of iterations of K-means using a randomly selected initial mean with K = 2 and 2-means clustering using the mean of the entire cluster as the initial mean.

Table 14
Classification of individuals based on assessed CVD risk using standard k-means clustering and 2-means clustering with a specific initial mean.

No. of individuals in each Clusters	Technique	
	Standard K-means clustering	2-means clustering using a specific initial mean
Low Risk	36	30
Medium Risk	21	27
High Risk	16	17
Critical Risk	27	26

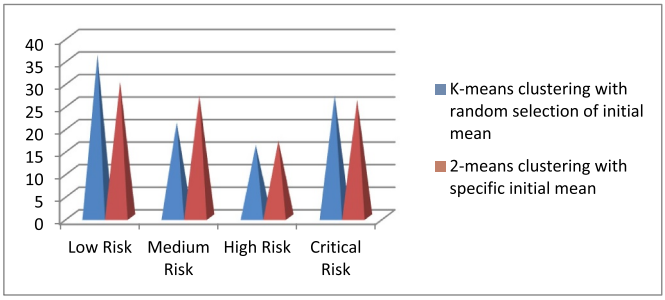

Fig. 7. Classification of individuals based on assessed CVD risk using standard k-means clustering and 2-means clustering with a specific initial mean.

6.5. Performance analysis of clustering algorithm

The clustering algorithm is evaluated using the Davies–Bouldin (DB) index [56,58], which is an internal evaluation method that validates the quality of the clustering.

The DB index is calculated as follows:

Table 15
Performance analysis of standard k-means clustering and 2-means clustering with a specific initial mean.

DB index	Technique	
	Standard K-means clustering	2-means clustering with a specific initial mean
Clustering of CVD affected individuals	0.295	0.282
Clustering of CVD unaffected individuals	0.2912	0.2836

$$DB = \frac{1}{N}\sum_{i=1}^{N}\left(\max_{j\neq i}\left\{\frac{\sigma_i + \sigma_j}{d(C_i,C_j)}\right\}\right) \quad (20)$$

where

N = no. of clusters,
σ_i = intracluster distance of the elements in ith cluster,
σ_j = intracluster distance of the elements in jth cluster, and

$d(C_i,C_j)$ = inter cluster distance.

The lower the value of the DB index is, the better the separation of the clusters is. Table 15 determines the DB index values of 2-means clustering with a specific initial mean and a random initial mean.

It is inferred from the above table that the DB index is lower for the 2 means clustering with a specific initial mean compared to that the of the standard 2 means clustering. Hence, the 2-means clustering technique with a specific initial mean outperforms the standard clustering algorithm.

7. Conclusion

The mortality rate due to CVD has rapidly increased worldwide. The analysis of a large number of factors that tend to cause CVD results in greater confusion among people due to its lower redundancy. To overcome this hurdle, a risk assessment hybrid model has been developed in this work. The output of the modified DE algorithm, which include 9 optimized critical attributes, and the predicted value of CVD, which is the output of the fuzzy AHP and ANN hybrid model, are taken as the inputs in this work. The cumulative prevalence ratio is determined using the Cox regression analysis of the 9 critical attributes. The Cox regression analysis using the 9 critical attributes results in a 91% CVD prediction accuracy compared with 88% accuracy when using all 13 attributes. Later, the risk analysis of each individual is carried out using 2-means clustering. The individuals are clustered into four groups, namely, low, medium, high and critical risk, based on the cumulative prevalence ratio using the 2-means clustering technique with a specific initial mean. This model results in a lower DB index of 0.282 for the clusters of individuals with CVD and 0.2836 for the clusters of individuals without CVD, whereas the traditional k-means clustering results in a DB index value of 0.295 for clusters of individuals with CVD and 0.2912 for the clusters of individuals without CVD. This result further proves the better clustering of individuals based on their assessed risk level. Thus, this hybrid risk assessment model allows individuals to effectively assess their CVD prevalence ratio and estimate their current risk group based on 9 critical attributes, which can further help them to assess their treatment options accordingly.

Conflicts of interest

We, Mr. T. Vivekanandan (corresponding author) and Dr. Swathi Jamjala Narayanan confirm that the content of this article entitled "**A Hybrid Risk Assessment Model for Cardiovascular Disease Using Cox Regression Analysis and 2-Means Clustering Algorithm**" does not contain any conflict of interest and is also not funded by any research organization.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2019.103400.

References

[1] R. Storn, On the Usage of Differential Evolution for Function Optimization, Biennial Conference of the North American Fuzzy Information Processing Society (NAFIPS), 1996, pp. 519–523.
[2] R. Storn, K. Price, Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces, J. Glob. Optim. 11 (4) (1997 Dec 1) 341–359.
[3] World Health Organization, Key facts about cardiovascular diseases (CVDs), Available at: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases- (cvds). Accessed in (Jan 2017).
[4] A. Kai Qin, Vicky Ling Huang, Ponnuthurai N. Suganthan, Differential evolution algorithm with strategy adaptation for global numerical optimization, IEEE Trans. Evol. Comput. 13 (2) (2009) 398–417.
[5] T. Vivekanandan, N. Ch Sriman Narayana Iyengar, Optimal feature selection using a modified differential evolution algorithm and its effectiveness for prediction of heart disease, Comput. Biol. Med. 90 (2017) 125–136.
[6] N. Nidheesh, K.A. Abdul Nazeer, P.M. Ameer, An enhanced deterministic K-Means clustering algorithm for cancer subtype prediction from gene expression data, Comput. Biol. Med. 91 (2017) 213–221.
[7] D. Mendes, S. Paredes, T. Rocha, P. Carvalho, T. Marques, J. Henriques, J. Morais, Short-term cardiovascular disease risk assessment: comparison of two combination approaches, 6th European Conference of the International Federation for Medical and Biological Engineering, Springer, Cham, 2015, pp. 670–673.
[8] R. Storn, K. Price, Differential Evolution - a Simple and Efficient Adaptive Scheme for Global Optimization over Continuous Space, Technical Report TR-95-012, Berkeley, CA, 1995.
[9] Valentin Fuster, Bridget B. Kelly, Rajesh Vedanthan, Promoting global cardiovascular health: moving forward, Circulation 123 (15) (2011) 1671–1678.
[10] Manish Bansal, Ravi R. Kasliwal, Naresh Trehan, Comparative accuracy of different risk scores in assessing cardiovascular risk in Indians: a study in patients with first myocardial infarction, Indian Heart J. 66 (6) (2014) 580–586.
[11] T. Vivekanandan, Swathi J. Narayanan, N. Ch Sriman Narayana Iyengar, Analysis of parallel SVM based classification technique on healthcare using big data management in cloud storage, Recent Pat. Comput. Sci. 11 (3) (2018) 169–178.
[12] Sandra N. Ofori, Osaretin J. Odia, Risk assessment in the prevention of cardiovascular disease in low-resource settings, Indian Heart J. 68 (3) (2016) 391–398.
[13] J. Bohacik, C. Kambhampati, D.N. Davis, G.F. John, Alternating decision tree applied to risk assessment of heart failure patients, Journal of Information Technologies 6 (2) (2013).
[14] P. Melillo, N.D. Luca, M. Bracale, L. Pecchia, Classification tree for risk assessment in patients suffering from congestive heart failure via long-term heart rate variability, IEEE Journal of Biomedical and Health Informatics 17 (3) (2013).
[15] A.J. Aljaaf, D. Al-Jumeily, A.J. Hussain, T. Dawson, P. Fergus, M. Al-Jumaily, Predicting the likelihood of heart failure with a multi level risk assessment using decision tree, IEEE Third International Conference on Technological Advances in Electrical, Electronics and Computer Engineering, TAEECE), 2015, pp. 101–106.
[16] UCI machine learning repository: Cleveland Heart Disease Dataset, https://archive.ics.uci.edu/ml/datasets/Heart + Disease.
[17] Jiawei Han, Micheline Kamber, Data Mining Concepts and Techniques, second ed., Morgan Kaufmann, An imprint of Elsevier, 2009, pp. 383–403.
[18] Ahmed Alsayat, El-Sayed Hoda, Efficient genetic K-means clustering for health care knowledge discovery, IEEE 14th International Conference on Software Engineering Research, Management and Applications (SERA), IEEE, 20162016.
[19] John A. Hartigan, Manchek A. Wong, Algorithm AS 136: a k-means clustering algorithm, Journal of the Royal Statistical Society. Series C (Applied Statistics) 28 (1) (1979) 100–108.
[20] Ahmad Mohawish, et al., Predicting Coronary Heart Disease risk using health risk assessment data, 17th IEEE International Conference on E-Health Networking, Application & Services (HealthCom), 2015.
[21] Naveen Garg, et al., Comparison of different cardiovascular risk score calculators for cardiovascular risk prediction and guideline recommended statin uses, Indian Heart J. 69 (4) (2017) 458–463.
[22] David R. Cox, Regression models and life-tables, J. R. Stat. Soc. Ser. B 34 (2) (1972) 187–202.
[23] D.R. Cox, D. Oakes, Analysis of survival data, boca raton, FL: Chapman & Hall/CRC (1984).
[24] R. Tibshirani, The Lasso method for variable selection in the Cox model, Stat. Med. 16 (1997) 385–395.
[25] J. Gui, H. Li, Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data, Bioinformatics 21 (2005) 3001–3008.
[26] Modelling Survival Data in Medical Research, third ed., By David Collett, 3rd Edition.
[27] Cox proportional hazards regression, https://www.statsdirect.com/help/survival_analysis/cox_regression.htm.

[28] D.Y. Lin, Cox regression analysis of multivariate failure time data: the marginal approach, Stat. Med. 13 (21) (1994) 2233–2247.
[29] Cox regression analysis, https://www.statsdirect.com/.
[30] J. Perk, G. De Backer, H. Gohlke, I. Graham, Z. Reiner, M.W.M. Verschuren, C. Albus, P. Benlian, G. Boysen, R. Cifkova, C. Deaton, M Fisher SEbrahim, G. Germano, R. Hobbs, A. Hoes, S. Karadeniz, A. Mezzani, E. Prescott, L. Ryden, M. Scherer, M. Syvanne, W.J.M. Scholte Op Reimer, C. Vrints, D. Wood, J.-L. Zamorano, F. Zannad, European guidelines on cardiovascular disease prevention in clinical practice (version 2012), Eur. Heart J. 33 (2012) 1635–1701.
[31] J. Santacruz-Varela, F. Hernandez-Torres, G. Fajardo-Dolci, Risk assessment towards patient safety in health care facilities, Cirugia y cirujanos 78 (6) (2010) 515–526.
[32] P.W. Wilson, R.B. D'Agostino, D. Levy, A.M. Belanger, H. Silbershatz, W.B. Kannel, Prediction of coronary heart disease using risk factor categories, Circulation 97 (18) (1998 May 1) 1837–1847.
[33] R.B. D'agostino, R.S. Vasan, M.J. Pencina, P.A. Wolf, M. Cobain, J.M. Massaro, W.B. Kannel, General cardiovascular risk profile for use in primary care: the Framingham Heart Study, Circulation 117 (6) (2008 Feb 12) 743–753.
[34] Performance of the QRISK cardiovascular risk prediction algorithm in an independent UK sample of patients from general practice: a validation study J Hippisley-Cox,1 CCoupland,1 Y Vinogradova,1 J Robson,2 P Brindle3.
[35] WHO/IHR Risk prediction chart, http://www.who.int/cardiovascular_diseases/publications/Chart_predictions/en/.
[36] Joint British societies recommendations on the prevention of cardiovascular disease (JBS3), http://www.jbs3risk.com/.
[37] Sharon Ann Hunt, ACC/AHA 2005 guideline update for the diagnosis and management of chronic heart failure in the adult: a report of the American college of cardiology/American heart association task force on practice guidelines (writing committee to update the 2001 guidelines for the evaluation and management of heart failure), J. Am. Coll. Cardiol. 46 (6) (2005) e1–e82.
[38] E.A. Amsterdam, N.K. Wenger, R.G. Brindis, D.E. Casey, T.G. Ganiats, D.R. Holmes, A.S. Jaffe, H. Jneid, R.F. Kelly, M.C. Kontos, G.N. Levine, AHA/ACC guideline for the management of patients with non–ST-elevation acute coronary syndromes: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines, J. Am. Coll. Cardiol. 64 (24) (2014) e139–228 2014 Dec 23.
[39] C.J. O'Donnell, R. Elosua, Cardiovascular risk factors. Insights from framingham heart study, Revista Espanola de Cardiologia (English Edition) 61 (3) (2008 Jan 1) 299–310.
[40] P. Brindle, A. Beswick, T. Fahey, S. Ebrahim, Accuracy and impact of risk assessment in the primary prevention of cardiovascular disease: a systematic review, Heart 92 (12) (2006 Dec 1) 1752–1759.
[41] M.R. Cowie, J. Bax, N. Bruining, J.G. Cleland, F. Koehler, M. Malik, F. Pinto, E. Vander Velde, P. Vardas, e-health, A position statement of the european society of cardiology, Eur. Heart J. 37 (1) (2016) 63–66.
[42] I. Kadi, A. Idri, J. Fernandez-Aleman, Knowledge discovery in cardiology: a systematic literature review, Int. J. Med. Inf. 97 (2017) 12–32.
[43] Cardiovascular risk charts and risk calculator, http://cvrisk.mvm.ed.ac.uk/help.htm.
[44] Yuanxin Hu, Survival analysis of cardiovascular diseases, All Theses and Dissertations (ETDs) (2013) 1209 http://openscholarship.wustl.edu/etd/1209.
[45] Cox regression analysis, https://www.ibm.com/support/knowledgecenter/de/SSLVMB_24.0.0/spss/advanced/idh_coxr.html.
[46] Bjorn Dahlof, Cardiovascular disease risk factors: epidemiology and risk assessment, Am. J. Cardiol. 105 (1) (2010) 3A–9A.
[47] SCORE Risk Charts, The European cardiovascular disease risk assessment model, https://www.escardio.org/Education/Practice-Tools/CVD-prevention-toolbox/SCORE-Risk-Charts.
[48] Oluwarotimi Williams Samuel, et al., An integrated decision support system based on ANN and Fuzzy_AHP for heart failure risk prediction, Expert Syst. Appl. 68 (2017) 163–172.
[49] Cox regression, https://ncss-wpengine.netdna-ssl.com/wpcontent/themes/ncss/pdf/Procedures/NCSS/Cox_Regression.pdf.
[50] P.K. Anooj, Clinical decision support system: risk level prediction of heart disease using weighted fuzzy rules, Journal of King Saud University-Computer and Information Sciences 24 (1) (2012) 27–40.
[51] S. Sahan, K. Polat, H. Kodaz, S. Güneş, The medical applications of attribute weighted artificial immune system (AWAIS): diagnosis of heart and diabetes diseases, International Conference on Artificial Immune Systems, Springer, Berlin,Heidelberg, 2005, pp. 456–468. Aug 14.
[52] Liam Cervante, Bing Xue, Lin Shang, Mengjie Zhang, Binary particle swarm optimisation and rough set theory for dimension reduction in classification, IEEE Congress on Evolutionary Computation, 2013, pp. 2428–2435.
[53] l Babaoglu, O. Findik, E. Ulker, 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) (2010 Apr 1) 3177–3183.
[54] N.C. Long, P. Meesad, H. Unger, A highly accurate firefly based algorithm for heart disease prediction, Expert Syst. Appl. 42 (21) (2015 Nov 30) 8221–8231.
[55] M.K. Awang, F. Siraj, Utilization of an artificial neural network in the prediction of heart disease, International Journal of Bio-science and Bio-technology 5 (4) (2013 Aug) 159–166.
[56] A.C. Benabdellah, A. Benghabrit, I. Bouhaddou, A survey of clustering algorithms for an industrial context, Procedia computer science 148 (2019 Jan 1) 291–302.
[57] R. Das, I. Turkoglu, A. Sengur, Effective diagnosis of heart disease through neural networks ensembles, Expert Syst. Appl. 36 (4) (2009 May 1) 7675–7680.
[58] Y. Yang, J. Jiang, Bi-weighted ensemble via HMM-based approaches for temporal data clustering, Pattern Recognit. 76 (2018 Apr 1) 391–403.
[59] WHO/ISH cardiovascular risk prediction charts, https://www.who.int/cardiovascular_diseases/guidelines/Chart_predictions/en/.
[60] J.F. Pagel, P. Kirshtein, Machine Dreaming and Consciousness, Academic Press, 2017 Apr 13.
[61] Y.M. Wang, K.S. Chin, Fuzzy analytic hierarchy process: a logarithmic fuzzy preference programming methodology, Int. J. Approx. Reason. 52 (4) (2011 Jun 1) 541–553.
[62] A.J. Barros, V.N. Hirakata, Alternatives for logistic regression in cross-sectional studies: an empirical comparison of models that directly estimate the prevalence ratio, BMC Med. Res. Methodol. 3 (1) (2003 Dec) 21.
[63] X. Jia, M.M. Baig, F. Mirza, H. GholamHosseini, A cox-based risk prediction model for early detection of cardiovascular disease: identification of key risk factors for the development of a 10- year CVD risk prediction, Advances in preventive medicine (2019) 2019.