Pattern Recognition Letters 133 (2020) 33–40

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Association between work-related features and coronary artery disease: A heterogeneous hybrid feature selection integrated with balancing approach

Elham Nasarian[a], Moloud Abdar[b,1,*], Mohammad Amin Fahami[c,1], Roohallah Alizadehsani[b], Sadiq Hussain[d], Mohammad Ehsan Basiri[e], Mariam Zomorodi-Moghadam[f], Xujuan Zhou[g], Paweł Pławiak[h,i], U. Rajendra Acharya[j,k,l], Ru-San Tan[m], Nizal Sarrafzadegan[n,o]

[a] Department of Industrial Engineering, Islamic Azad University, Najafabad Branch, Najafabad, Iran
[b] Institute for Intelligent Systems Research and Innovation (IISRI)Locked Bag 20000, Geelong, VIC 3220, Australia
[c] Electrical and Computer Engineering, Isfahan University of Technology, Isfahan 841583111, Iran
[d] System Administrator, Dibrugarh University, Dibrugarh 786004, India
[e] Department of Computer Engineering, Shahrekord University, Shahrekord 64165478, Iran
[f] Department of Computer Engineering, Ferdowsi University of Mashhad, Mashhad 9177948974, Iran
[g] School of Management & Enterprise, University of Southern Queensland, QLD 4300 Australia
[h] Department of Information and Communications Technology, Faculty of Computer Science and Telecommunications, Cracow University of Technology, Warszawska 24 st., F-3, Krakow 31-155, Poland
[i] Institute of Theoretical and Applied Informatics, Polish Academy of Sciences, Bałtycka 5, Gliwice 44-100, Poland
[j] Department of Electronics and Computer Engineering, Ngee Ann Polytechnic, Singapore 599489, Singapore
[k] Department of Biomedical Informatics and Medical Engineering, Asia University, Taichung, Taiwan
[l] Department of Biomedical Engineering, School of Science and Technology, Singapore University of Social Sciences, Singapore
[m] Department of Cardiology, National Heart Centre Singapore, Singapore 169609, Singapore
[n] Isfahan Cardiovascular Research Center, Cardiovascular Research Institute, Isfahan University of Medical Sciences, Isfahan 8174673461, Iran
[o] Faculty of Medicine, School of Population and Public Health, The University of British Columbia, Vancouver, Canada

ARTICLE INFO

Article history:
Received 29 October 2019
Revised 23 January 2020
Accepted 6 February 2020
Available online 11 February 2020

MSC:
41A05
41A10
65D05
65D17

Keywords:
Machine learning
Data mining
Heart disease
Coronary artery disease
Feature selection

ABSTRACT

Coronary artery disease (CAD) is a leading cause of death worldwide and is associated with high healthcare expenditure. Researchers are motivated to apply machine learning (ML) for quick and accurate detection of CAD. The performance of the automated systems depends on the quality of features used. Clinical CAD datasets contain different features with varying degrees of association with CAD. To extract such features, we developed a novel hybrid feature selection algorithm called heterogeneous hybrid feature selection (2HFS). In this work, we used Nasarian CAD dataset, in which work place and environmental features are also considered, in addition to other clinical features. Synthetic minority over-sampling technique (SMOTE) and Adaptive synthetic (ADASYN) are used to handle the imbalance in the dataset. Decision tree (DT), Gaussian Naive Bayes (GNB), Random Forest (RF), and XGBoost classifiers are used. 2HFS-selected features are then input into these classifier algorithms. Our results show that, the proposed feature selection method has yielded the classification accuracy of 81.23% with SMOTE and XGBoost classifier. We have also tested our approach with other well-known CAD datasets: Hungarian dataset, Long-beach-va dataset, and Z-Alizadeh Sani dataset. We have obtained 83.94%, 81.58% and 92.58% for Hungarian dataset, Long-beach-va dataset, and Z-Alizadeh Sani dataset, respectively. Hence, our experimental results confirm the effectiveness of our proposed feature selection algorithm as compared to the existing state-of-the-art techniques which yielded outstanding results for the development of automated CAD systems.

1. Introduction

Cardiovascular disease (CVD), which includes coronary artery disease (CAD), is the leading cause of death globally, and responsible for 17.9 million deaths (31% of all mortality) annually [64]. About 85% of all CVD deaths are attributable to heart attack and

* Corresponding author.
E-mail address: m.abdar1987@gmail.com (M. Abdar).
[1] M. Abdar and M. A. Fahami contributed equally to this work.

https://doi.org/10.1016/j.patrec.2020.02.010
0167-8655/

stroke. High blood pressure, raised blood glucose level, obesity and overweight are major risk factors, alongside lifestyle factors like tobacco use, physical inactivity and unhealthy diet [65]. It is estimated that by 2030, 23.6 million persons will die from stroke, heart attack and other CVD diseases worldwide [65]. Many lives can be saved by detecting CVD and CAD early.

CVD can be asymptomatic; stroke or heart attack is its first manifestation [65]. Hence, early detection can save life and also reduce cost of healthcare. Machine learning and optimized methods have been developed for early detection of CVD [34,41,62]. Such algorithms evolved from classical statistical domain in the1990's. The exponential technological advances have facilitated the shift from relational databases to data science [15]. Different platforms and software libraries were created which could extract useful information from large datasets. By combining database, statistics, machine learning, and data mining strives to find the hidden signature present in the large sea of data [43]. Data mining comprises various steps: data collection, pre-processing of the data (data preparation), choosing a model (machine learning algorithms), training the selected model, evaluate of the selected model, parameter tuning and finally make predictions. Machine learning algorithms are used in many fields such as big data, social networking, internet of things, computer-aided diagnosis of diseases [4,27,48,59], analysis of biomedical data, cybersecurity, and intrusion detection. Computer-aided diagnosis (CADx) systems help the doctors to acquire, manage, store, and report medical images, which include X-rays, computed tomography (CT), ultrasound, and magnetic resonance imaging [46]. In healthcare, machine learning algorithms that extract patterns and hidden relationships in available clinical datasets [23] are used to detect or even predict the development of disease. More importantly, selection of most important features in a database plays a significant role in improving the performance of machine learning and data mining algorithms [9]. There are several studies in the literature which applied different feature selection algorithms for the detection of CAD using various datasets (see Table 1). It should be noted even Alizadehsani et. al[9]. achieved an accuracy of 94.08%, however, the applied feature selection algorithm was applied to the original data without consideration of categorical features. For example, Wosiak et. al[66]. advocated for the unsupervised selection of features from datasets and their grouping to obtain better statistical results and proposed a methodology. For the proposed clustering, non-correlated features were used as attributes. The methodology was evaluated using three CVD datasets. The experimental results demonstrated its advantage over existing feature selection strategies. A summary of the existing automated CAD classification methods is summarized in Table 1.

As discussed earlier, numerous feature selection algorithms have been employed to analyse the most important features among CAD dataset. However, there are few critical problems on both datasets and machine learning methods. Firstly, we introduced a new CAD dataset with five new features: office location of patients, shift work, stress, noise exposure and pollutant, which were missing in all previous datasets. In this regard, we introduced a new CAD dataset called the Nasarian CAD dataset. Moreover, the applied feature selection algorithms suffer from critical issues such as low performance, low improvement, and slow learning speed in our study. To do so, we proposed a new hybrid feature selection algorithm called heterogeneous hybrid feature selection (2HFS). Moreover, we found out the importance of newly added features for CAD detection. The proposed 2HFS is applied to Nasarian CAD dataset and found that office location of patients, stress and shift work features have remarkable impact on the detection of CAD. These findings are consistent with the results found by physicians. To show the effectiveness of the proposed method, 2HFS is applied to other three well-known CAD data sets: Hungarian dataset, Long-beach-va dataset, and Z-Alizadeh Sani dataset. The obtained results demonstrated the effectiveness of our proposed feature selection algorithm for CAD detection using other datasets. The rest of our study is structured as follows. Section 2 provides a detailed description of the proposed feature selection algorithm. The obtained results are presented in Section 3. The discussion on the obtained results as well as the comparison with the existing works are described in Section 4. Finally, the study is concluded in Section 5.

2. Proposed methodology: heterogeneous hybrid feature selection (2HFS) algorithm

Feature selection algorithms play a crucial role in machine learning. Selecting the most important features has significant impact on the medical diagnostic process. It helps to get accurate and quick diagnosis. We rank all features based on scores obtained using different features selection algorithms namely Fisher score algorithm (FSA), F_score algorithm (FA) and extra trees classifier algorithm (ETCA). In this study, we first used the top 30%, 40%, and 50% of the highest scoring features in all CAD datasets. We

Table 1
Summary of existing automated CAD classification methods using various public databases.

Study	Year	Dataset	Method	Accuracy (%)
Alizadehsani et al. [9]	2013	Z-Alizadeh Sani	SMO (Sequential Minimal Optimization)	94.08
Alizadehsani et al. [13]	2016	Z-Alizadeh Sani	SVM (Support Vector Machine)	86.14 (LAD), 83.17 (LCX), 83.50 (RCA)
Qin et al. [51]	2017	Z-Alizadeh Sani	EA-MFS (Ensemble algorithm based on multiple feature selection)	93.70
Gokulnath et al. [28]	2018	Cleveland	genetic algorithm (GA)–SVM	88.34
Ahamed and ZahidHasan[5]	2017	Hungarian	J48	72.10
Subramaniyam et al. [58]	2019	Hungarian	TGD (Taylorgradient descent)-based ACNN (actor critic neural network)	82.55
Saqlain et al. [55]	2019	Hungarian	Fisher score-based feature selection &Forward feature selection & Reverse feature selection & RBF kernel-based SVM	76.40
Lo et al. [42]	2016	Long-beach-va	Technique for Order Preference by Similarity to an Ideal Solution (TOPSIS)	79.10
Alizadehsani et al. [12]	2012	Z-Alizadeh Sani	SMO	92.09
Kolukisaet al. [38]	2019	Z-Alizadeh Sani	Enemble Classifier with FLDA	92.07
Verma et al. [60]	2016	Indira Gandhi Medical data	Particle swam optimization (PSO) & correlation-based feature subset (CFS) & K-means	88.40
Alkeshuosh et al. [14]	2017	Cleveland	PSO	87.00
Abdar [1]	2015	Cleveland	C5.0	85.33
Alizadehsani et al. [11]	2012	Z-Alizadeh Sani	SMO (1-1. 2-1, and 3-1)	92.41 (average)
Aouabed et al. [18]	2019	Cleveland	NE-nu-SVC	98.34

first checked the selected CAD datasets. We have filled the missing values (if any) by using the 'mode' value. Then six classifiers are used for classification: decision tree (DT), Gaussian Naive Bayes (GNB), random forest (RF) XGBoost, K-Nearest Neighbors (KNN) and Bernoulli Naive Bayes (BNB). To choose the most important features, we then applied three well-known features selection algorithms namely Fisher score algorithm (FSA), F_score algorithm (FA) and extra trees classifier algorithm (ETCA). After applying these feature selection algorithms, once again six classical classifiers were used to the CAD datasets. In this step, we calculated the average accuracy of all classical classifiers when they applied to all four CAD datasets for each feature selection algorithm one by one. For clarity, in each time we applied one feature selection algorithm and then selected the top 30%, 40%, and 50% of the most important features for all four CAD datasets. Hence, we will have three sets of data for each CAD dataset (30%, 40%, and 50%), so totally we have 12 sets of data. Afterward, we applied six classifiers to all of these 12 datasets. Finally, we used the obtained average accuracy (for testing step) of each feature selection algorithm as its weight when that feature selection is used in our proposed feature selection algorithm.

The main steps of the proposed approach are summarized below.

1. Find missing values and fill them using a statistical approach (mode).
2. Calculate the rank of each feature using three feature selection algorithms used in this study (FSA), FA and ETCA).
3. Select the top 30%, 40%, and 50% of the most important features for each CAD dataset after applying feature selection algorithms.
4. Apply different classification algorithms (DT, GNB, RF, XGBoost, KNN and BNB classifiers) to the original CAD datasets without missing values.
5. Calculate the average accuracy of all classifiers when each feature selection algorithm used as a weight for that feature selection algorithm when it is used in our proposed feature selection algorithm. For example, apply the F_score algorithm as a feature selection algorithm to select the top 30%, 40%, and 50% of the most important features in all four CAD datasets. Then, apply all six classification algorithms to the top 30%, 40%, and 50% of the most important features in each dataset separately. Finally, we calculated the average accuracy of all classifiers applied to the top 30%, 40%, and 50% of the most important features in all datasets.
6. Apply the proposed feature selection algorithm using FSA, FA and ETCA feature selection algorithms taking into consideration the obtained weight for each feature selection algorithm in the previous step (step 5).
7. Apply the best four classification algorithms (DT, GNB, RF and XGBoost) in the step NO. 4 to the CAD datasets for the selected features in step 6.

In this study, a heterogeneous hybrid feature selection (2HFS) algorithm is proposed which combines three well-known feature selection algorithms: a) FSA, b) FA, and c) ETCA. Along with these

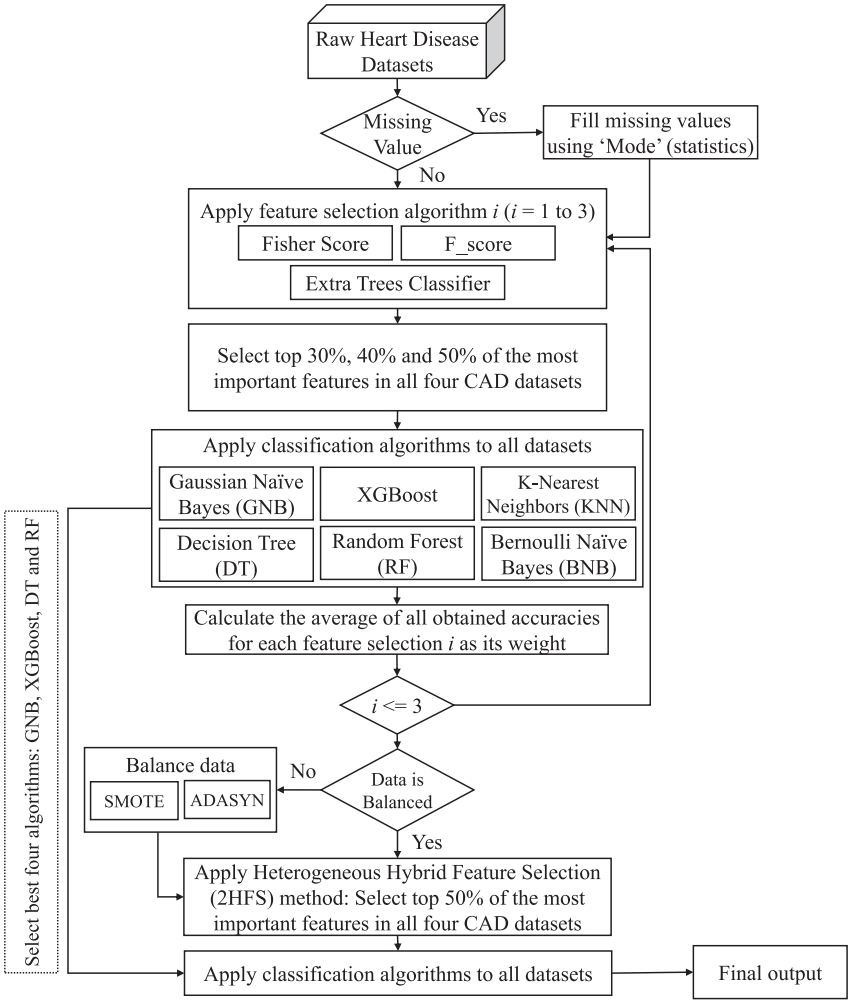

Fig. 1. Schematic block diagram of the proposed model.

Table 2
Details of benchmarking datasets from UCI machine learning repository.

Dataset	Number of features	Number of records	Number of classes
Hungarian dataset	14	294	2
Long-beach-va dataset	14	200	2
Z-Alizadeh Sani dataset	56	303	2

feature selection algorithms, four base classification algorithms: DT, GNB, RT and XGBoost are used (we used six classification algorithms but we selected the best four algorithms in this step). Also, two well-known balancing approach are applied to enhance the performance of the proposed method. A schematic block diagram of the proposed model in this study is illustrated in Fig. 1. We first apply the classification algorithms separately. Then, each feature selection algorithm is separately used to find the most important features. It should be noted that, we choose the top 50% of features. Once again, the selected features are given to the base classification algorithms. The average accuracy of all classification methods applied after using each feature selection algorithm chosen as a weight in our proposed 2HFS approach as indicated in Eq. (1).

$$W_i = \frac{1}{n}\left(\sum_{d=1}^{m} \sum_{j\in\ \{0.3,\ \ 0.4,\ \ 0.5\}} \sum_{k\ =\ 1}^{n} ACC(C_{k_j}^{d})\right), \tag{1}$$

where W_i is the weight of feature selection i, m is the number of datasets, and$n$ is the total number of classifiers. $ACC(C_{k_j})$ denotes the accuracies of all classifiers k for 30%, 40%, and 50% of the most important features by feature selection algorithms F_i.

Moreover, we selected four well-known classification algorithms among six classification algorithms (Decision Trees [54], Gaussian Naive Bayes (GNB) [47], Random Forest (RF) [40], and Extreme Gradient Boosting (XGBoost) [67]) and two balancing techniques (Synthetic Minority Over-sampling Technique (SMOTE) [21] and Adaptive synthetic (ADASYN) [30]). More information for each method is giving in Supplementary Material.

3. Results

We tested 2HFS in the "Nasarian dataset," which comprises records of 150 subjects (all male employees in Iran have visited the Abadan Occupational (Industrial) Medicine Clinic) and 52 features stratified in *four* main clusters: a) work place and environment, b) demographic information, c) symptom and examination, and d) laboratory findings. The detailed description of Nasarian dataset is given in Tables 3.1–3.4 in Supplementary Material. It may be noted that in even though there are female employees in the Abadan Occupational (Industrial) Medicine Clinic, only male employees had CAD. Additionally, all applicants were exposed to pollutants. Therefore, these features (sex, pollutant exposure) are removed from the dataset. We applied our proposed methodology to other three benchmarking datasets collected from the UCI machine learning repository (see Table 2).

As mentioned earlier, in this paper, we first applied six base classifiers: DT, GNB, RF, XGBoost, KNN and BNB. But, we selected the best four base classifiers (DT, GNB, RF and XGBoost) with our proposed feature selection. The main parts of the proposed methodology include hybrid feature selection and balancing techniques. As mentioned earlier, the proposed 2HFS approach is applied with three tiers of percentage thresholds to select the most important features (30%, 40% and 50% of the whole dataset). Detailed information on the most important features selected by 2HFS using Nasarian dataset is given in Table 3 (top 50% of important features). It should be noted that we noticed after running the algorithms for several times, different results were obtained. To deal with issue, we decided to run our model several times and then report the average. Thus, we applied our 2HFS feature selection 10 times and then calculated the average score for each feature as shown in Table 3. In addition, we applied the four base classifiers (DT, GNB, RF and XGBoost) 10 times and then calculated the average evaluation metrics presented in Tables 6 and 7.

Table 3
Ranked features using Heterogeneous Hybrid Feature Selection (2HFS).

No.	Feature	Score
1.	Chest Pain_no	190.4499
2.	BP (Blood Pressure)	183.8219
3.	Exercise test_Positive	182.9989
4.	abdominal obesity_no (less than 90 cm)	171.4174
5.	Chest Pain_yes	170.5389
6.	EX-Smoker_no	169.7306
7.	HTN (Hypertension)_yes	162.2944
8.	abdominal obesity_yes (bigger than 90 cm)	160.6239
9.	**Office**	**159.8481**
10.	HTN (Hypertension)_no	159.8102
11.	**Shift work_fixed**	**157.3491**
12.	FH (Family History)_no	157.3261
13.	EX-Smoker_yes	156.4584
14.	Exercise test_Negative	156.4462
15.	CVA (Cerebrovascular Accident)_no	155.6354
16.	FAMILY HTN_no	144.8905
17.	LDL (Low density lipoprotein)	132.5278
18.	CVA (Cerebrovascular Accident)_yes	132.4020
19.	FBS (Fasting Blood Sugar)	130.8411
20.	FAMILY HTN_yes	130.7856
21.	**Shift work_not fixed**	**124.1576**
22.	FH (Family History)_yes	123.3075
23.	**STRESS_yes**	**114.3345**
24.	Cr (Creatine)	112.5951
25.	HLP (Dyslipidemia)_no	111.7476
26.	TG (Triglyceride)	109.2703
27.	Systolic Murmur _yes	108.4066
28.	DM (Diabetes Mellitus)_no	106.8092
29.	DM (Diabetes Mellitus)_yes	104.3222
30.	eos	103.5046
31.	**STRESS_no**	**102.7140**
32.	RBC	101.8989
33.	Current Smoker_no	101.7746
34.	POLY(Neutrophil)	101.0096
35.	Systolic Murmur _no	100.9353
36.	Function _yes	100.1581
37.	HLP (Dyslipidemia)_yes	100.1271
38.	Age	98.5741

Table 3 shows that Chest Pain, BP, Exercise test, Abdominal obesity, EX-Smoker, HTN, Office, Shift work, FH, Exercise test, CVA, Family HTN, LDL, FBS, Stress, Cr, HLP, TG, Systolic Murmur, DM, eos, RBC, Current Smoker, POLY, Function and Age (totally 26 features) have the highest scores (highly ranked selected features). In Table 3, five work-related features are listed among top selected features for the Nasarian dataset (highlighted in Table 3). The first work-related feature is "office location" which is ranked 9 among 38 selected features in Table 3. The second most important feature among work-related features is "Shift work_fixed" which is ranked 11 of the most important features in the dataset. This demonstrates that having fixed shift work is an important factor for CAD disease dataset and we think that it should be considered for further investigation. Other three important work-related features are "Shift work_not fixed", "Stress-yes" and "Stress-no" which are ranked 21, 23 and 31, respectively. Moreover, we tested the efficiency of our

Table 4
Average accuracies for all classifiers.

Dataset	Accuracy
DT	0.7729
GNB	0.7876
RF	0.8170
XGBoost	0.8302

Table 5
Weights for feature selection algorithms.

Dataset	Weight
Fisher Score algorithm	0.8258
F_score algorithm	0.8284
ExtraTreesClassifier algorithm	0.8298

approach by applying the same method to other three public (UCI) datasets: Hungarian, Z-alizadeh Sani and Long_beach-va datasets. The top 50% of all features in each dataset are chosen for the rest of study. For example, top 38 most important features are selected for the classification step. It should be noted that one hot coding is used to deal with all categorical features all datasets. For example, Chest Pain has two values: yes and no. After applying one hot coding, Chest Pain converted to "Chest Pain_yes" and "Chest Pain_no".

Average accuracies for DT, GNB, RF and XGBoost classifiers are shown in Table 4 and weights for feature selection algorithms are presented in Table 5.

Table 6 shows the results of various algorithms and selected feature sets for Nasarian and Hungarian datasets (called set 1) and Table 7 shows the results of various algorithms and selected feature sets (using different feature selection algorithms and the proposed method) with Long_beach and Z-Alizadeh Sani datasets (called set 2) using different feature selection algorithms and the proposed method. The proposed method increased the accuracy of classification using the feature selection algorithms. The best accuracy for Nasarian dataset is obtained using hybrid feature selection algorithm, with ADASYN balancing method, and XGBoost algorithm. Using this method, we achieved an accuracy of 81.23%. Tables 4 and 5 demonstrate that the proposed feature selection can improve the performance of almost all algorithms used in the study (please see improvement charts in Supplementary Material). As shown in Table 4, the proposed 2HFS approach improves the performance of all four methods applied (DT, GNB, RF, and XGBoost classifiers) with Nasarian CAD dataset in terms of all metrics except recall of random forest. It should be noted that we used original datasets (in Tables 6 and 7) without any missing values. In other words, we filled missing values using a statistical approach.

4. Discussion

It can be shown from Table 3 that the proposed method selected top 26 features (top 50% of most important features) in the Nasarian CAD dataset. The influence of work place (also called office or location) conditions on CAD cannot be ignored. According to Price [50], the type of job is an important factor and is associated with heart disease. In this study, data include job-related features such as office (location of job), stress and shift work. Virtanen et al. [63] demonstrated strong association between job insecurity and incidence of CAD. Kobayashi [37] indicated that long work hours (combined with lack of sleep), shift work, working in a cold or noisy place, irregular working hours, jet lag, frequent work-related trips, and psychological job strain can be considered as risk factors for stroke and CAD. Similarly, Sara et al., [56] observed strong association between work-related stress and CAD. These findings and ours underscore the impact of psychosocial risk factors on heart disease. Janczura et al., [33] found that having high stress increases the prevalence of CAD and metabolic syndrome. In addition to internal factors (e. g., sex, age, BMI, blood pressure, etc.), external factors (e. g., stress, shift work, etc.) play a significant role in the development or aggravation of heart disease. Moran et al. [45] observed that financial stress is linked to the incidence of CAD. Among patients with stable CAD, Hagström et al. [29] found various psychosocial stress such as loss of interest, depressive symptoms, financial stress and living alone were associated with increased cardiovascular mortality. Moreover, Kivimäki and Steptoe [36] showed that stress in childhood can damage the health of children and increase the risk of multiple chronic conditions whereas stressful experiences in adulthood had relatively less influence. In our study, we have 52 real features in the beginning

Table 6
Comparison of results using various algorithms and datasets (set 1) for 50% features.

Feature selection algorithm		Nasarian Dataset				Hungarian Dataset			
Original data	Classification algorithm	Accuracy	Precision	Recall	F-score	Accuracy	Precision	Recall	F-score
	DT	0.5999	**0.7198**	0.6723	0.6873	0.7398	0.6635	0.6355	0.6418
	GNB	0.6266	0.7036	0.7502	0.7193	**0.8086**	0.7512	**0.7632**	**0.7472**
	RF	0.6133	0.6845	0.7582	0.7107	0.7909	**0.7972**	0.6500	0.7048
	XGBoost	**0.6533**	0.7102	**0.8122**	**0.7523**	0.7743	0.7248	0.6690	0.6760
Fisher Score algorithm	DT	0.6300	0.7202	0.6649	0.6838	0.8023	**0.8148**	0.7026	0.7218
	GNB	0.7066	0.7462	0.8501	0.7886	0.8137	0.7880	0.7086	**0.7730**
	RF	0.7000	0.7498	0.7962	0.7666	**0.8047**	0.7701	0.6688	0.7112
	XGBoost	**0.7200**	**0.7641**	**0.8660**	**0.7960**	0.7967	0.7257	**0.7346**	0.7161
F_score algorithm	DT	0.6266	0.7135	0.6869	0.6856	0.8012	0.7627	0.6299	0.6728
	GNB	**0.7333**	**0.7632**	**0.8861**	**0.8139**	**0.8133**	**0.8021**	**0.7518**	**0.7599**
	RF	0.6866	0.7697	0.7738	0.7600	0.8039	0.8014	0.6864	0.7281
	XGBoost	0.7236	0.7561	0.8255	0.7813	0.8022	07588	0.7510	0.7400
Extra Trees Classifier	DT	0.6399	0.7214	0.7149	0.7084	0.7658	0.7163	0.6322	0.6580
algorithm	GNB	**0.7400**	**0.7697**	**0.9137**	**0.8319**	**0.8180**	**0.8033**	**0.7760**	**0.7634**
	RF	0.6547	0.7128	0.8485	0.7649	0.8042	0.7602	0.7178	0.7228
	XGBoost	0.6666	0.7343	0.8141	0.7639	0.7735	0.7203	0.6467	0.6739
Proposed+ADASYN	DT	0.7086	**0.7087**	**0.7254**	**0.7015**	**0.8171**	**0.8706**	**0.7347**	**0.7871**
	GNB	0.7628	0.6608	0.8683	0.7449	0.8227	0.8487	0.7828	0.8032
	RF	0.7844	0.7640	0.7908	0.7744	0.8196	0.8130	0.8233	0.8118
	XGBoost	**0.7994**	0.7968	0.8234	0.7955	0.8384	0.8284	0.8308	0.8259
Proposed+SMOTE	DT	0.7550	0.7890	0.7195	0.7385	0.8193	0.8713	0.7475	0.8005
	GNB	0.7750	0.7507	0.7917	0.7619	**0.8394**	0.8680	0.7544	0.8049
	RF	0.7749	0.7840	0.7340	0.7547	0.8286	**0.8717**	0.7769	0.8198
	XGBoost	**0.8123**	**0.8024**	**0.8522**	**0.8207**	0.8283	0.8279	**0.8399**	**0.8297**

Table 7
Comparison of results using various algorithms and datasets (set 2) for 50% features.

Feature selection algorithm		Long_beach Dataset				Z-Alizadeh Sani Dataset			
Original data	Classification algorithm	Accuracy	Precision	Recall	F-score	Accuracy	Precision	Recall	F-score
	DT	0.6722	0.8011	0.7619	0.7756	0.7400	0.5574	0.6720	0.5836
	GNB	0.6300	**0.8850**	0.5985	0.7064	0.7262	0.5114	**0.9030**	0.6470
	RF	0.7211	0.7420	0.7281	0.7214	0.8215	0.7388	0.5989	0.6544
	XGBoost	**0.7244**	0.8041	**0.8100**	**0.8359**	**0.8445**	**0.7932**	0.8117	**0.7292**
Fisher Score algorithm	DT	**0.7188**	0.8320	0.7972	0.8055	0.7983	0.6728	0.6766	0.6529
	GNB	0.7311	**0.8708**	0.7791	0.8086	0.7786	0.5582	**0.9548**	0.6923
	RF	0.7432	0.8024	0.8771	0.8284	0.8546	**0.8222**	0.6278	0.6963
	XGBoost	**0.7544**	0.8000	**0.9280**	**0.8504**	**0.8579**	0.7742	0.6859	**0.7130**
F_score algorithm	DT	**0.7200**	0.7941	0.8864	0.8264	0.8216	0.6850	0.6909	0.6762
	GNB	0.7344	**0.8988**	0.7552	0.8116	0.7855	0.5373	**0.9203**	0.6612
	RF	0.7511	0.8380	0.8585	0.8402	0.8584	0.8172	0.6504	0.7195
	XGBoost	**0.7544**	0.8058	**0.9133**	**0.8442**	**0.8615**	**0.8057**	0.7119	**0.7457**
Extra Trees Classifier algorithm	DT	**0.7500**	**0.8277**	0.8485	0.8315	0.7989	0.6522	0.6139	0.6217
	GNB	0.7490	0.8054	0.8170	0.8479	0.8511	0.7461	**0.7316**	0.7327
	RF	**0.7733**	0.8222	0.9091	**0.8580**	0.8518	**0.8116**	0.6191	0.6827
	XGBoost	0.7611	0.7947	**0.9278**	0.8518	**0.8611**	0.7869	0.7298	**0.7474**
Proposed+ADASYN	DT	**0.7753**	0.7827	0.8099	**0.8700**	0.8812	0.9045	0.8750	0.8836
	GNB	0.7944	0.8738	0.8777	0.8688	0.8872	0.8913	0.8807	0.8827
	RF	**0.8045**	0.8778	**0.8830**	0.8626	0.9156	0.9156	0.9101	0.9127
	XGBoost	0.8041	**0.8944**	0.7608	0.8680	**0.9235**	**0.9374**	**0.9118**	**0.9214**
Proposed+SMOTE	DT	**0.7841**	0.7543	0.8410	0.8406	0.8702	0.8565	0.8917	0.8730
	GNB	0.7787	0.8030	0.7406	0.7504	0.8958	0.9083	0.8861	0.8935
	RF	0.8150	**0.9191**	**0.9287**	**0.9212**	0.9190	0.9226	0.9178	**0.9180**
	XGBoost	**0.8158**	0.8552	0.7947	0.8913	**0.9258**	**0.9259**	**0.9299**	0.9062

and after using one hot encoding, the number of the most important features increased to 77. The importance of each sub-feature after applying one hot encoding is presented in Table 3. For example, the Shift work_fixed feature has a greater impact (higher importance score) than the Shift work_not fixed feature. Our findings showed that, both STRESS_yes and STRESS_no have close importance scores but having stress (STRESS_yes feature in Table 3) is more important.

In summary, we studied a dataset comprising of both external and internal factors for CAD among male employees visited the Abadan Occupational (Industrial) Medicine Clinic. According to Table 3, the office location of CAD patients is an important feature for the identification of CAD. Other extracted important work-related features are shift work and stress. Tables 4 and 5 show that, ADASYN technique performed better than SMOTE in all four different heart disease datasets. Amin et al. [16] investigated the selection of most important features among heart disease datasets. They found that, chest pain feature is the most important feature in all 12 studies used to identify the heart disease [16]. This is consistent with our observation in Table 3, where chest pain has the highest score among the selected features. Table 6 compares the results of our proposed approach with other works that have been done on the UCI datasets. As shown, our proposed method yields better results compared with other studies in the literature.

In Table 8 (for Hungarian dataset), Akay [6] obtained the highest classification accuracy of 84% and we obtained the accuracy of 83.94%. For Long-beach-va dataset in Table 6 it can be seen that our proposed method obtained the best performance of 81.58% followed by Lo et al. [42] and Arabasadi et al. [19] with the accuracies of 79.10% and 78%, respectively. However, for Z-Alizadeh Sani dataset), Abdar et al. [2] and Abdar et al. [3] achieved the better accuracies of 94.64% and 93.08%, respectively followed by our study with accuracy of 92.58%. It is worth mentioning that in this we have not modified any classification algorithms whereas most of previous studies improved the performance of their methods using optimization techniques such as ensemble techniques. In other words, we proposed a simple but efficient pre-processing method which can improve the performance of classification algorithms. Overall, the performance of the proposed method is among the best performances in the literature, suggesting that selection of the most important features can significantly enhance the performance of machine learning algorithms. In summary, the main advantages of our proposed method are as follows:

- Improved the performance of classical machine learning algorithms.
- Extracted the most important features from the CAD dataset.
- Increased the speed of training and testing as we used only 50% of the total number of features.

Table 8
Comparison of performance with other existing methods using different datasets, **Accuracy (%)**.

Study	Hungarian	Long-beach-va	Z-Alizadeh Sani
Proposed	**83.94**	**81.58**	**92.58**
[52]	78.50	–	–
[17]	46.93	–	–
[19]	82.90	78.00	–
[25]	80.50	–	–
[6]	**84.00**		
[44]	59.68	–	–
[49]	80.00	–	–
[20]	83.33	–	–
[57]	72.00	–	–
[26]	78.57	71.50	–
[53]	66.67	–	–
[5]	72.10	–	–
[58]	82.55	–	–
[55]	76.40	–	–
[42]	–	79.10	–
[32]	–	71.50	–
[31]	–	57.36	–
[12]	–	–	92.09
[38]	–	–	92.07
[24]	–	–	90.91
[8]	–	–	87.22
[7]	–	–	74.20
[10]	–	–	82.16
[39]	–	–	88.16
[22]	–	–	86.49
[3]	–	–	**93.08**
[35]	–	–	89.40
[61]	–	–	88.22
[2]	–	–	**94.66**

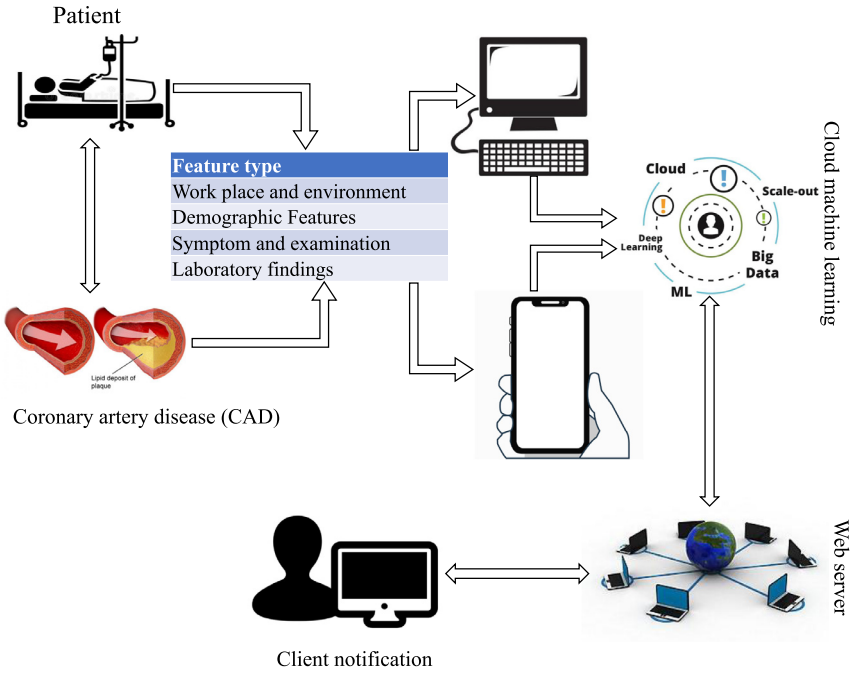

Fig. 2. Interpretation of the proposed machine learning cloud-based system to diagnose the CAD.

- Reduced the data complexity.
- Preserved the data characteristics for interpretation.
- Proposed method can be used as clinical decision support system.

Disadvantages of the proposed method are given below:

- Used small dataset.
- Relationship between factors should be significantly checked when incorporated in some combination approaches (e. g., ensemble learning) with new features.
- Collect a bigger dataset including female patients.

We hope to tackle these disadvantages in the future by presenting new methods. Firstly, we would collect CAD dataset with a greater number of patients with more clinical and external features. Moreover, it will be interesting to check the importance of our newly added features (office location of penitents, shift work, stress, noise exposure and pollutant) for female patients. In addition, we plan to apply our proposed methodology for different evolutionary-based feature selection algorithms. The proposed system can be used in cloud-based CAD recognition system as illustrated in Fig. 2.

5. Conclusion

CAD is the leading cause of death worldwide. It is important to extract relevant features from the CAD subjects in order to obtain highest detection performance. Different features in datasets are associated with various degrees of CAD. A novel 2HFS feature selection algorithm has been proposed and applied to the new Nasarian CAD dataset. This dataset contains work-related features in addition to other clinical features. Also, we have used SMOTE and ADASYN balancing techniques and four classifiers (DT, GNB, RT, and XGBoost). The proposed methodology is also applied to three established UCI CAD datasets namely Hungarian, Long_beach-va and Z-Alizadeh Sani datasets. Our experiments show that the proposed combination of 2HSF technique and SMOTE balancing approach yielded high accuracies. In future, we intend to test our proposed algorithm with huge datasets with more features. Also, we intend to apply various evolutionary algorithms with our proposed 2HFS method. In addition, we aim to apply different ensemble learning methods with other new features.

Declaration of Competing Interest

The authors declare that there is no conflict of interest regarding the publication of this article.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.patrec.2020.02.010.

References

[1] M. Abdar, Using decision trees in data mining for predicting factors influencing of heart disease., Carpathian J. Electron. Comput. Eng. 8 (2) (2015).
[2] M. Abdar, U.R. Acharya, N. Sarrafzadegan, V. Makarenkov, Ne-nu-svc: a new nested ensemble clinical decision support system for effective diagnosis of coronary artery disease, IEEE Access 7 (2019) 167605–167620.
[3] M. Abdar, W. Ksiazek, U.R. Acharya, R.-S. Tan, V. Makarenkov, P. Plawiak, A new machine learning technique for an accurate diagnosis of coronary artery disease, Comput. Methods Programs Biomed. 179 (2019) 104992.
[4] U.R. Acharya, Y. Hagiwara, J.E.W. Koh, S.L. Oh, J.H. Tan, M. Adam, R. San Tan, Entropies for automated detection of coronary artery disease using ecg signals: a review, 2018.
[5] M.S. Ahamed, N.M. ZahidHasan, A decision support system for classification of heart disease using data mining algorithms, Int. J. Comput. Sci. Inf. Secur. 15 (1) (2017) 573.
[6] M. Akay, Noninvasive diagnosis of coronary artery disease using a neural network algorithm, Biol. Cybern. 67 (4) (1992) 361–367.
[7] R. Alizadehsani, J. Habibi, B. Bahadorian, H. Mashayekhi, A. Ghandeharioun, R. Boghrati, Z.A. Sani, Diagnosis of coronary arteries stenosis using data mining, J. Med. Signals Sens. 2 (3) (2012) 153.
[8] R. Alizadehsani, J. Habibi, M.J. Hosseini, R. Boghrati, A. Ghandeharioun, B. Bahadorian, Z.A. Sani, Diagnosis of coronary artery disease using data mining techniques based on symptoms and ecg features, Eur. J. Scient. Res. 82 (4) (2012) 542–553.
[9] R. Alizadehsani, J. Habibi, M.J. Hosseini, H. Mashayekhi, R. Boghrati, A. Ghandeharioun, B. Bahadorian, Z.A. Sani, A data mining approach for diagnosis of coronary artery disease, Comput. Methods Programs Biomed. 111 (1) (2013) 52–61.
[10] R. Alizadehsani, J. Habibi, Z.A. Sani, H. Mashayekhi, R. Boghrati, A. Ghandeharioun, B. Bahadorian, Diagnosis of coronary artery disease using data mining based on lab data and echo features, J. Med. Bioeng. 1 (1) (2012).
[11] R. Alizadehsani, M.J. Hosseini, R. Boghrati, A. Ghandeharioun, F. Khozeimeh, Z.A. Sani, Exerting cost-sensitive and feature creation algorithms for coronary

artery disease diagnosis, Int. J. Knowl. DiscoveryBioinf. (IJKDB) 3 (1) (2012) 59–79.
[12] R. Alizadehsani, M.J. Hosseini, Z.A. Sani, A. Ghandeharioun, R. Boghrati, Diagnosis of coronary artery disease using cost-sensitive algorithms, in: 2012 IEEE 12th International Conference on Data Mining Workshops, IEEE, 2012, pp. 9–16.
[13] R. Alizadehsani, M.H. Zangooei, M.J. Hosseini, J. Habibi, A. Khosravi, M. Roshanzamir, F. Khozeimeh, N. Sarrafzadegan, S. Nahavandi, Coronary artery disease detection using computational intelligence methods, Knowl. Based Syst. 109 (2016) 187–197.
[14] A.H. Alkeshuosh, M.Z. Moghadam, I. Al Mansoori, M. Abdar, Using pso algorithm for producing best rules in diagnosis of heart disease, in: 2017 international conference on computer and applications (ICCA), IEEE, 2017, pp. 306–311.
[15] G. Amato, L. Candela, D. Castelli, A. Esuli, F. Falchi, C. Gennaro, F. Giannotti, A. Monreale, M. Nanni, P. Pagano, et al., How Data Mining and Machine Learning Evolved from Relational Data Base to Data Science, in: A Comprehensive Guide Through the Italian Database Research Over the Last 25 Years, 2018, pp. 287–306.
[16] M.S. Amin, Y.K. Chiam, K.D. Varathan, Identification of significant features and data mining techniques in predicting heart disease, Telemat. Inf. 36 (2019) 82–93.
[17] P. Anooj, Clinical decision support system: risk level prediction of heart disease using weighted fuzzy rules, J. King Saud Univ.-Comput. Inf. Sci. 24 (1) (2012) 27–40.
[18] Z. Aouabed, M. Abdar, N. Tahiri, J.C. Gareau, V. Makarenkov, A novel effective ensemble model for early detection of coronary artery disease, in: International Conference Europe Middle East & North Africa Information Systems and Technologies to Support Learning, Springer, 2019, pp. 480–489.
[19] Z. Arabasadi, R. Alizadehsani, M. Roshanzamir, H. Moosaei, A.A. Yarifard, Computer aided decision making for heart disease detection using hybrid neural network-genetic algorithm, Comput. Methods Programs Biomed. 141 (2017) 19–26.
[20] A. Bhatt, S.K. Dubey, A.K. Bhatt, Sudden cardiac arrest prediction using predictive analytics, International Journal of Intelligent Engineering & Systems (2017). 10–3.
[21] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, Smote: synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2002) 321–357.
[22] A. Cüvitoğlu, Z. Işik, Classification of cad dataset by using principal component analysis and machine learning approaches, in: 2018 5th International Conference on Electrical and Electronic Engineering (ICEEE), IEEE, 2018, pp. 340–343.
[23] S.K. Dehkordi, H. Sajedi, Prediction of disease based on prescription using data mining methods, Health. Technol. 9 (1) (2019) 37–44.
[24] A. Dekamin, A. Sheibatolhamdi, A data mining approach for coronary artery disease prediction in iran, J. Adv. Med. Sci. Appl. Technol. 3 (1) (2017) 29–38.
[25] H. Dubey, V. Pudi, Class based weighted k-nearest neighbor over imbalance dataset, in: Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, 2013, pp. 305–316.
[26] R. El-Bialy, M.A. Salamay, O.H. Karam, M.E. Khalifa, Feature analysis of coronary artery heart disease data sets, Procedia Comput. Sci. 65 (2015) 459–468.
[27] H. Fujita, D. Cimr, Computer aided detection for fibrillations and flutters using deep convolutional neural network, 2019.
[28] C.B. Gokulnath, S. Shantharajah, An optimized feature selection based on genetic approach and support vector machine for heart disease, Cluster Comput. (2018) 1–11.
[29] E. Hagström, F. Norlund, A. Stebbins, P. Armstrong, K. Chiswell, C. Granger, J. López-Sendón, D. Pella, J. Soffer, R. Sy, et al., Psychosocial stress and major cardiovascular events in patients with stable coronary heart disease, J. Intern. Med. 283 (1) (2018) 83–92.
[30] H. He, Y. Bai, E.A. Garcia, S. Li, Adasyn: Adaptive synthetic sampling approach for imbalanced learning, in: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), IEEE, 2008, pp. 1322–1328.
[31] S. Iwata, Y. Ohsawa, S. Tsumoto, N. Zhong, Y. Shi, L. Magnani, Communications and discoveries from multidisciplinary data, 123, Springer, 2008.
[32] M. Jalal, et al., Performance Evaluation of Machine Learning Algorithms for Coronary Artery Disease Features, United International University, 2019 Ph.D. thesis.
[33] M. Janczura, G. Bochenek, R. Nowobilski, J. Dropinski, K. Kotula-Horowitz, B. Laskowicz, A. Stanisz, J. Lelakowski, T. Domagala, The relationship of metabolic syndrome with stress, coronary heart disease and pulmonary function-an occupational cohort-based study, PLoS One 10 (8) (2015) e0133750.
[34] V. Jayaraman, H.P. Sultana, Artificial gravitational cuckoo search algorithm along with particle bee optimized associative memory neural network for feature selection in heart disease classification, J. Ambient Intell. Humaniz. Comput. (2019) 1–10.
[35] Ü. Kiliç, M.K. Keleş, Feature selection with artificial bee colony algorithm on z-alizadeh sani dataset, in: 2018 Innovations in Intelligent Systems and Applications Conference (ASYU), IEEE, 2018, pp. 1–3.
[36] M. Kivimäki, A. Steptoe, Effects of stress on the development and progression of cardiovascular disease, Nat. Rev. Cardiol. 15 (4) (2018) 215.
[37] F. Kobayashi, Job stress and stroke and coronary heart disease, Jpn. Med. Assoc. J. 47 (5) (2004) 222–226.
[38] B. Kolukısa, H. Hacılar, M. Kuş, B. Bakır-Güngör, A. Aral, V. Güngör, Diagnosis of coronary heart disease via classification algorithms and a new feature selection methodology, Int. J. Data Mining Sci. 1 (1) (2019) 8–15.
[39] H. Li, X. Wang, Y. Li, C. Qin, C. Liu, Comparison between medical knowledge based and computer automated feature selection for detection of coronary artery disease using imbalanced data, in: BIBE 2018; International Conference on Biological Information and Biomedical Engineering, VDE, 2018, pp. 1–4.
[40] A. Liaw, M. Wiener, et al., Classification and regression by randomforest, R News 2 (3) (2002) 18–22.
[41] M. Liu, Y. Kim, 40th annual international conference of the ieee engineering in medicine and biology society (embc) (2018) 2707–2710.
[42] Y.-T. Lo, H. Fujita, T.-W. Pai, Prediction of coronary artery disease based on ensemble learning approaches and co-expressed observations, J. Mech. Med. Biol. 16 (1) (2016) 1640010.
[43] H. Mannila, Data mining: machine learning, statistics, and databases, in: Proceedings of 8th International Conference on Scientific and Statistical Data Base Management, IEEE, 1996, pp. 2–9.
[44] H. Mohebbi, Y. Mu, W. Ding, Learning weighted distance metric from group level information and its parallel implementation, Appl. Intell. 46 (1) (2017) 180–196.
[45] K.E. Moran, M.J. Ommerborn, C.T. Blackshear, M. Sims, C.R. Clark, Financial stress and risk of coronary heart disease in the jackson heart study, Am. J. Prev. Med. 56 (2) (2019) 224–231.
[46] R.M. Nishikawa, Computer-aided Detection and Diagnosis, in: Digital Mammography, 2010, pp. 85–106.
[47] M. Ontivero-Ortega, A. Lage-Castellanos, G. Valente, R. Goebel, M. Valdes-Sosa, Fast gaussian naïve bayes for searchlight classification analysis, Neuroimage 163 (2017) 471–479.
[48] H. Özkan, O. Osman, S. Şahin, A.F. Boz, A novel method for pulmonary embolism detection in cta images, Comput. Methods Programs Biomed. 113 (3) (2014) 757–766.
[49] A.K. Paul, P.C. Shill, M.R.I. Rabin, M. Akhand, Genetic algorithm based fuzzy decision support system for the diagnosis of heart disease, in: 2016 5th International Conference on Informatics, Electronics and Vision (ICIEV), IEEE, 2016, pp. 145–150.
[50] A.E. Price, Heart disease and work, Heart 90 (9) (2004) 1077–1084.
[51] C.-J. Qin, Q. Guan, X.-P. Wang, Application of ensemble algorithm integrating multiple criteria feature selection in coronary heart disease detection, Biomed. Eng. 29 (6) (2017) 1750043.
[52] J.R. Quinlan, Improved use of continuous attributes in c4. 5, J. Artif. Intell. Research 4 (1996) 77–90.
[53] G.T. Reddy, N. Khare, An efficient system for heart disease prediction using hybrid ofbat with rule-based fuzzy logic model, J. Circuits Syst. Comput. 26 (4) (2017) 1750061.
[54] S.R. Safavian, D. Landgrebe, A survey of decision tree classifier methodology, IEEE Trans. Syst. Man Cybern. 21 (3) (1991) 660–674.
[55] S.M. Saqlain, M. Sher, F.A. Shah, I. Khan, M.U. Ashraf, M. Awais, A. Ghani, Fisher score and matthews correlation coefficient-based feature subset selection for heart disease diagnosis using support vector machines, Knowl. Inf. Syst. 58 (1) (2019) 139–167.
[56] J.D. Sara, M. Prasad, M.F. Eleid, M. Zhang, R.J. Widmer, A. Lerman, Association between work-related stress and coronary heart disease: a review of prospective studies through the job strain, effort-reward balance, and organizational justice models, J. Am. Heart Assoc. 7 (9) (2018) e008073.
[57] R. Saravana Kumar, G. Tholkappia Arasu, Rough set theory and fuzzy logic based warehousing of heterogeneous clinical databases, Int. J. Uncertainty Fuzziness Knowl. Based Syst. 25 (3) (2017) 385–408.
[58] A. Subramaniyam, R.P. Mahapatra, P. Singh, Taylor and gradient descent-based actor critic neural network for the classification of privacy preserved medical data, Big Data 7 (3) (2019) 176–191.
[59] F. Tang, D.K. Ng, D.H. Chow, An image feature approach for computer-aided detection of ischemic stroke, Comput. Biol. Med. 41 (7) (2011) 529–536.
[60] L. Verma, S. Srivastava, P. Negi, A hybrid data mining model to predict coronary artery disease cases using non-invasive clinical data, J. Med. Syst. 40 (7) (2016) 178.
[61] J. Vijayashree, H.P. Sultana, A machine learning framework for feature selection in heart disease classification using improved particle swarm optimization with support vector machine classifier, Programm. Comput. Softw. 44 (6) (2018) 388–397.
[62] J. Vijayashree, H.P. Sultana, Heart disease classification using hybridized ruzzo-tompa memetic based deep trained neocognitron neural network, Health. Technol. (2019) 1–10.
[63] M. Virtanen, S.T. Nyberg, G.D. Batty, M. Jokela, K. Heikkilä, E.I. Fransson, L. Alfredsson, J.B. Bjorner, M. Borritz, H. Burr, et al., Perceived job insecurity as a risk factor for incident coronary heart disease: systematic review and meta–analysis, BMJ 347 (2013) f4746.
[64] WHO, Cardiovascular diseases (cvds), 2019a, (https://www.who.int/cardiovascular_diseases/en/a). Accessed: 14.3.2019.
[65] WHO, Cardiovascular diseases (cvds), 2019b, (https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)/b). Accessed on 14.3. 2019.
[66] A. Wosiak, D. Zakrzewska, Integrating correlation-based feature selection and clustering for improved cardiovascular disease diagnosis, Complexity 2018 (2018) 1–11.
[67] H. Zhang, D. Qiu, R. Wu, Y. Deng, D. Ji, T. Li, Novel framework for image attribute annotation with gene selection xgboost algorithm and relative attribute model, Appl. Soft Comput. 80 (2019) 57–79.