J Med Syst (2016) 40: 178
DOI 10.1007/s10916-016-0536-z

A Hybrid Data Mining Model to Predict Coronary Artery Disease Cases Using Non-Invasive Clinical Data

Luxmi Verma[1] · **Sangeet Srivastava**[2] · **P. C. Negi**[3]

Received: 1 March 2016 / Accepted: 1 June 2016 / Published online: 11 June 2016

Abstract Coronary artery disease (CAD) is caused by atherosclerosis in coronary arteries and results in cardiac arrest and heart attack. For diagnosis of CAD, angiography is used which is a costly time consuming and highly technical invasive method. Researchers are, therefore, prompted for alternative methods such as machine learning algorithms that could use noninvasive clinical data for the disease diagnosis and assessing its severity. In this study, we present a novel hybrid method for CAD diagnosis, including risk factor identification using correlation based feature subset (CFS) selection with particle swam optimization (PSO) search method and K-means clustering algorithms. Supervised learning algorithms such as multi-layer perceptron (MLP), multinomial logistic regression (MLR), fuzzy unordered rule induction algorithm (FURIA) and C4.5 are then used to model CAD cases. We tested this approach on clinical data consisting of 26 features and 335 instances collected at the Department of Cardiology, Indira Gandhi Medical College, Shimla, India. MLR achieves highest prediction accuracy of 88.4 %.We tested this approach on benchmarked Cleaveland heart disease data as well. In this case also, MLR, outperforms other techniques. Proposed hybridized model improves the accuracy of classification algorithms from 8.3 % to 11.4 % for the Cleaveland data. The proposed method is, therefore, a promising tool for identification of CAD patients with improved prediction accuracy.

Keywords Classification · Particle swarm optimization · Coronary artery disease · Clustering

This article is part of the Topical Collection on *Transactional Processing Systems*

✉ Sangeet Srivastava
sangeetsrivastava@ncuindia.edu

[1] Department of Computer Science and Engineering, The NorthCap University, Gurgaon, India

[2] Department of Applied Sciences, The NorthCap University, Gurgaon, India

[3] Department of Cardiology, Indira Gandhi Medical College, Shimla, India

Introduction

Cardiovascular diseases (CVD) are caused by disorders of the heart and blood vessels and result in coronary heart disease, heart failure, cardiac arrest, ventricular arrhythmias and sudden cardiac death, ischemic stroke, transient ischemic attack, subarachnoid and intracerebral hemorrhage, rheumatic heart disease, abdominal aortic aneurysm, peripheral artery disease and congenital heart disease [1]. According to World Health Organization (WHO), 17.5 million people died from CVD in 2012 amounting to 31 % of all global deaths [2]. CAD is a type of CVD in which presence of atherosclerotic plaques in coronary arteries, leads to myocardial infarction or sudden cardiac death [3]. In order to diagnose positive sign of heart disease and to assess the level of damage of heart muscles, certain tests may be prescribed by a medical practitioner including nuclear scan, angiography, echocardiogram, electrocardiogram (ECG), exercise stress testing [4]. ECG is a noninvasive technique used to identify CAD cases [5, 6], though it could lead to undiagnosed symptoms of CAD [7]. This limitation leads to angiography which is an invasive diagnosis to confirm CAD cases and is considered as the gold standard for disease detection and severity analysis. However, it is costly and requires high level of technical expertise [8]. Researchers are, therefore, seeking less expensive and effective alternatives, say, using

data mining for predicting CAD cases. During the past few decades, image processing, signal processing, statistical and machine learning techniques have been increasingly applied to assist medical diagnosis using ECG and echocardiogram [9–17, 23, 24]. ECG and echocardiogram are specialized processes conducted by trained practioners. Sometimes ECG is not able to confirm CAD cases. This process is complex, costly, involves lot of time and effort.

To overcome these limitations many researchers used other risk factors excluding angiography to predict CAD cases. These methods are noninvasive, less complex, low cost, reproducible and objective diagnoses, can do automated detection of disease and can be used for screening large number of patients based on clinical data easily obtained at hospitals. Alizadehsani et al. [18] applied the data mining techniques with 10 fold cross validation namely Bagging, Sequential Minimal Optimization (SMO), Artificial Neural Network (ANN) and Naive Bayes algorithms for prediction of CAD. They used data collected from Rajaie Cardiovascular Medical and Research Center, Tehran, Iran having 54 features and 303 instances. Bagging and SMO achieved the same accuracy of about 89 %. ANN had prediction accuracy of 85 % and Naive Bayes' accuracy was lowest among the other classifiers. Alizadehani created new features such as Left Anterior Descending (LAD), Left Circumflex (LCX) and Right Coronary Artery (RCA) from the existing features. Higher values of any of these newly created features indicated higher probability of having CAD. Karaolis et al. [19] employed a decision tree (DT) algorithm (C4.5) and analyzed the clinical data using five splitting criteria based on information gain, likelihood ratio, gini index, chi-square statistics and gain ratio. They used distance measures for assessment of risk factors for CAD events. The most important risk factors identified for myocardial infraction were age, history of hypertension and smoking and classification accuracy was 66 % using the data collected from Paphos General Hospital, Cyprus. Ordonez [20] applied association rules for prediction of presence and absence of heart disease based on heart perfusion measurement and proposed an algorithm that uses search constraints to reduce the number of insignificant and redundant rules. He used clinical data set of 655 patients with 25 attributes for this experiment. Srinivas et al. [21] compared performance of supervised learning techniques, namely, DT, ANN and Bayes, for the prediction of heart disease in Singareni coal mining regions in Andhra Pradesh, India. DT performed better than the other methods. Palaniappan et al. [22] developed prototype intelligent heart disease prediction system with Data Mining Extension (DMX) query language to model supervised learning techniques namely, Naïve Bayes, DT and ANN, using clinical dataset of 15 features and 909 instances. Naïve Bayes gave the highest prediction accuracy of 86.5 %. Melillo et al. [23] discusses the importance of data mining techniques: Support Vector Machine (SVM), random forest, ANN for automatic prediction of significant CVD risk factors based on heart rate variability (HRV). In their study, data mining based classifier showed the higher predicted accuracy as compared to echographic parameters. Random forest outperformed the other classifiers. Acharya et al. [24] developed noninvasive computer-aided diagnostic system using image processing and data mining techniques to classify symptomatic and asymptomatic plaques. Discrete wavelet transformation was used to extract features from ultrasound images and SVM was used for classification. They achieved accuracy of 83.7 % using SVM with polynomial kernel of order 2.

Many researchers worked on optimized classifiers to remove irrelevant features and insignificant instances to increase prediction accuracy. Lin et al. [25] proposed hybrid evolutionary algorithm using endocrine based PSO and Artificial Bee Colony algorithm (ABC) for the selection of optimal feature subsets. Hybrid method improves the prediction accuracy of the classification algorithm with reduced number of features. Subanya et al. [26] applied swarm intelligence based ABC for feature selection and found that ABC-SVM performs better than reverse ranking and forward ranking method of feature selection on Cleveland heart disease. Amin et al. [27] employed ANN and genetic algorithms for prediction of heart disease. Genetic algorithm is used to optimize neural network to achieve the prediction accuracy of 89 %. In this study, we use feature subset selection with optimization techniques to improve prediction accuracy. This technique is further combined with clustering and classification techniques to improve accuracy of classification techniques.

Material and methods

Clinical data of 335 suspected CAD patients were collected at Indira Gandhi Medical College (IGMC), Shimla, India and is summarized in Table 1. Treadmill test (TMT) was performed as per BRUCE protocol using GE Healthcare TMT machine. Patients, who had already undergone TMT within the 3 months preceding coronary angiography, were not subjected to repeat TMT. Results of coronary angiography performed with Siemens Artis Zee Cath Lab Equipment through femoral or radial route are used as indicator for CAD cases. This is used as 26th feature used as predictant. Out of 335 instances, 48.9 % were confirmed CAD cases [28].

In this study, we prose a hybrid method consisting of clinical data collection, dimensionality reduction with correlation based feature subset selection with PSO, followed by data clustering for identification of incorrectly assigned cluster data points. Finally, models were constructed with MLP, MLR, FURIA and decision tree (C4.5). Figure 1 shows the block diagram of the proposed system.

Table 1 Features of patients clinical data collected at IGMC, Shimla

Features (encoding)	Description	Range	Mean ± StDev
Age	Age (in years)	30–86	55.47 ± 9.44
Sex	1-male,0-female	0–1	0.68 ± 0.46
Smoking habit	0-Non-smoker, 1- Ex-smoker, 2- Current smoker	0–2	0.8 ± 0.91
HTN	Hyper Tension	0–1	0.472 ± 0.5
DM	0 – No and 1 – Yes Diabetes mellitus	0–1	0.15 ± 0.36
Dyslipidemia	0 – No and 1 – Yes 0 – No and 1 – Yes	0–1	0.78 ± 0.41
Chest pain type	0 – Non – specific chest pain 1 – Atypical chest pain 2 – Typical angina	0–2	1.31 ± 0.75
RBS	Random Blood Sugar (mg/dL)	57–180	99.41 ± 26.83
TC	Total cholesterol (mg/dL)	117–287	182.5 ± 30.66
LDL	Low density lipoprotein (mg/dL)	56–178	112.71 ± 20.49
HDL	High density lipoprotein (mg/dL)	23–56	36.692 ± 6.933
TG	Triglyceride (mg/dL)	103–298	148.97 ± 28.29
SBP	Systolic blood pressure (mmHg)	100–170	124.18 ± 12.43
DBP	Diastolic blood pressure (mmHg)	46–110	77.96 ± 7.09
HT	Height (Cm)	133–188	164.79 ± 9.105
WT	Weight (Kg)	33–110	65.46 ± 10.72
BMI	Body mass index (kg/m2)	13.7–38.3	24.08 ± 3.56
WC	Waist circumference (Cm)	70–110	88.072 ± 6.75
Visceral Obesity	0 = False 1 = True	0–1	0.52 ± 0.5
ABI	Ankle-brachial index test	0.7–1.4	1.22 ± 0.08
Exercise Duration	Exercise Duration (minutes)	1–11	7.85 ± 1.78
METS	Metabolic exercise stress test	2–14	8.974 ± 1.704
RPP	Rate Pressure product	114–412	249.39 ± 41.14
Duration Recovery	Duration of recovery with persistent ST changes	0–7	1.57 ± 1.56
Duke	Duke treadmill score	−25-11	−4.835 ± 6.414
CAD	Coronary Artery Disease	0-No 1-Yes	

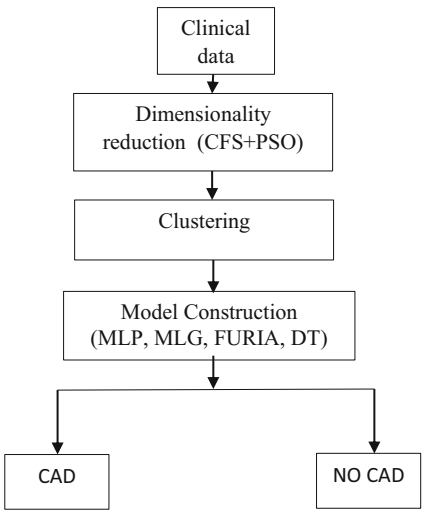

Fig. 1 Block diagram of the proposed hybrid system

Models for CAD identification

Multi layer perceptron (MLP)

ANN is a mathematical model for information processing consisting of a number of highly interconnected elements organized into layers inspired by the human brain. It is trained with a part of the data to explore the association between inputs and outputs, and tested on the rest of the data. Multilayer perceptron is a popular ANN architecture is used to model complex relationship between inputs and outputs [29–31].

Multinomial logistic regression model (MLR)

It is an extension of logistic regression with ridge estimator. MLR is a simple extension of binary logistic regression that

allows for more than two categories of the dependent or outcome variable. Like binary logistic regression, MLR uses maximum likelihood estimation to evaluate the probability of categorical membership [32].

Fuzzy unordered rule induction algorithm (FURIA)

It is an advancement of RIPPER algorithm [33]. It generates fuzzy rules instead of conventional rules to model decision boundaries in a more flexible way. Fuzzy rules are obtained through replacing intervals with fuzzy intervals using trapezoidal membership function combined with the sophisticated rule induction techniques employed by the original RIPPER algorithm [34].

C4.5

is a decision tree induction algorithm developed by the Quinlan [35]. It uses divide-and-conquer approach to construct decision tree, works with top–down approach. It uses gain ratio as splitting criteria. The algorithm uses heuristics for pruning based on the statistical significance of splits [19].

Evaluation parameters

Prediction accuracy of classification model is usually measured in terms of accuracy and misclassification rate [36].

$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn}$$

$$\text{Misclassification rate} = \frac{fp + fn}{tp + fn + fp + fn}$$

where, *tp*- true positive rate, *tn*-true negative rate, *fp*- false positive rate, *tn*- true negative rate.

Results and discussion

IGMC data

The proposed study uses CAD data collected from Department of Cardiology, IGMC, Shimla, India. All the patients were referred for invasive angiography suspected for coronary artery disease. The result of angiography is included as a class with yes and no values for presence and absence of disease with patient's laboratory and demographic factors. We excluded chest pain type from the data set because most cases are oriented towards one type of chest pain and could give the biased results. We constructed predictive models for CAD case identification using classification techniques and considered all the twenty-six features. The results showed that MLR achieved highest prediction accuracy of 83.5 % among the other techniques, namely, MLP, FURIA and C4.5. MLP showed the lowest accuracy of 77.0 % (Table 2).

Table 2 Performance of MLP, MLG, FURIA and C4.5 using all the features of CAD data

Algorithm	Accuracy (%)	Incorrectly classified instances (%)
MLR	83.5	16.4
MLP	77.0	22.9
FURIA	77.9	22.8
C4.5	77.3	22.6

Table 3 Risk factors identified with correlation based feature selection with PSO search

S. No.	Risk factors for CAD
1	Smoking
2	Diabetes Mellitus
3	High Density Lipoprotein
4	Duke Tread mill Score
5	Duration recovery

However, all the features are not always significant, some feature are irrelevant and redundant, and do not contribute significantly to prediction. A feature selection technique usually reduces the dimensionality of feature space and removes redundant, irrelevant, or noisy data. Feature reduction immediately affects the modeling framework in terms of speeding up data mining algorithm, improving the data quality, performance of data mining and increasing the understanding of the mining results [37]. The best feature subset contains the least number of features that contribute most to accuracy and efficiency. Merits of feature selection include: facilitating data visualization and data understanding, reducing the measurement and storage requirements, reducing training and utilization times [38, 39].

We propose hybrid method for prediction of coronary heart disease where data preprocessing is done by correlation-based feature subset selection [40] with PSO search method [41]. PSO is a population-based stochastic optimization technique developed by Kennedy and Eberhart in 1995 [42]. PSO is motivated by social behaviors such as bird flocking and fish

Table 4 CFS + PSO + classification (5 predictors with 335 instances)

Algorithm	Accuracy	Incorrectly classified instances	% improvement in accuracy (all features vs CFS + PSO feature selection
MLR	84.17	15.8	0.67
FURIA	80.29	19.70	2.39
MLP	79.7	20.2	2.7
C4.5	77.9	22.08	0.6

Table 5 Accuracy obtained using hybrid method (CFS + PSO + k-means clustering + calssification)

Algorithm	Accuracy (%)	Incorrectly classified instances (%)	% improvement in accuracy (CFS + PSO vs CFS + PSO + clustering + classification)
MLR	88.4	11.5	4.23
MLP	84.11	15.8	4.41
FURIA	82.8	17.13	2.51
C4.5	80.68	19.3	2.78

Table 7 Performance of prediction models for Cleaveland heart disease data set

Algorithm	All the features Accuracy	CFS + PSO	PSO + clustering	% improvement in accuracy with our approach
MLR	83.16	85.47	91.36	8.2
FURIA	79.86	79.20	87.05	8.3
MLP	78.87	83.49	90.28	11.4
C4.5	77.22	79.86	85.6	8.3

schooling [43]. After applying CFS & PSO we found only five significant features contributing most to CAD (Table 3). The first three are well known risk factors for CAD and the other two are considered to confirm CAD cases based on clinical data. Clustering and diagnostic models are constructed using supervised learning algorithms namely MLP, MLG, FURIA and C4.5 and validated using tenfold cross validation method. The experiments are conducted using Weikato Environment for knowledge analysis toolkit [44]. The performance of the models was evaluated using confusion matrix, accuracy. Misclassified rate are also calculated and compared.

When we used these five features to model CAD cases, accuracy of the classification methods was improved by 2.7 % in case of MLP, 0.67 % in case of MLG, 2.39 % in case of FURIA and 0.6 % in case of C4.5 (Table 4).

After feature selection, k-means clustering algorithm [45] is applied and incorrectly classified instances with wrong cluster membership are removed before the predictive model construction step. The results show that the proposed approach increased the accuracy of prediction algorithm between 2.51 % to 4.41 %, as shown in Table 5. The k-fold cross validation results are shown in Table 6.

Cleveland data

We also investigated the framework on Cleaveland heart disease data set with 14 features and 303 instances [https://archive.ics.uci.edu/ml/datasets/Heart+Disease]. The attributes of Cleaveland data set are age, sex, cp - chest pain type (typical angina, atypical angina, non-angina pain, asymptomatic), trestbps - resting blood pressure on admission, chol - serum cholesterol, fbs - fasting blood sugar, restecg - resting ECG outcome, thalch - maximum heart rate achieved, old peak - ST depression induced by exercise related to rest, slope - slope of the peak exercise ST Segment, ca - number of fluoroscopy colored vessels, thal - reversible defect and class (sick/healthy). After feature reduction step we got only seven risk factors: cp, thalch, exang, old peak, slope, ca, thal. With this novel hybridization method, prediction accuracy of classification models are increased by 11.4 % in case of MLP, 8.2 % in case of MLG, 8.3 % in case of FURIA and 8.3 % in case of C4.5 (Table 7).

We compared accuracy achieved by earlier used models for Cleveland dataset with our hybridized model. Kahramanli et al. [46] employed a hybrid model including ANN and fuzzy neural network (FNN) for coronary heart disease. Hybridization achieves the prediction accuracy of 86.8 % for CAD prediction. Peter et al. [47] employed Naïve Bayes K-Nearest Neighbor and Decision Tree with CFS, Chi square, consistency subset, filtered attribute, filtered subset and gain ratio methods for dimensionality reduction. NB classifier gave better accuracy for CAD prediction after applying the CFS feature selection method. MLP and Decision tree obtained a classification accuracy of 82.22 % and 81.11 %. Bouali et al. [48] analyzed classification methods such as decisions trees, MLP, Bayesian network, SVM and Fuzzy Pattern tree in order to predict heart disease. SVM outperformed the rest using train and test method as compared to 10 fold cross validation with 85.7 % accuracy.

Table 6 Performance of proposed method with number of k-folds

No. of folds	MLR		MLP		FURIA		C4.5	
	Acc.	ICI	Acc.	ICI	Acc.	ICI	Acc.	ICI
2	87.5	12.4	81.9	18.0	83.4	16.5	81.6	18.3
4	88.7	11.21	83.4	16.5	81.9	18.0	80.9	19
6	87.5	12.46	85.0	14.9	81.3	18.6	82.5	17.4
8	87.8	12.14	85.0	14.9	82.2	17.7	81.9	18.0
10	88.4	11.5	84.11	15.8	82.8	17.1	80.6	19.3

Acc: Accuracy ICI: Incorrectly classified instances

Table 8 Accuracy obtained by researchers using Cleaveland Heart Disease data set

Reference	Accuracy	Method	Authors
[46]	86.8 %	MLP + Fuzzy	Kahramanli et al. (2008)
[47]	82.22	CFS + MLP	Peter et al. (2012)
[48]	58.44	MLP + train and test method	Bouali et al. (2014)
Hybrid Method (used in this study)	90.28	CFS + PSO + Clustering + MLP	

Our model performed better with an accuracy of 92.8 % (Table 8). This improvement could be due to noise and feature reduction in the dataset.

Conclusion

The paper presents a novel hybrid model to identify and confirm CAD cases at low cost by using clinical data that can be easily collected at hospitals. Complexity of the system is decreased by reducing the dimensionality of the data set with PSO. It provides reproducible and objective diagnosis, and hence can be a valuable adjunct tool in clinical practices. Results are comparably, promising and therefore the proposed hybrid method will be helpful in disease diagnostics. Experiment results demonstrate the superiority of the proposed hybrid method with regard to prediction accuracy of CAD with the features selected by CFS & PSO, we need only a few clinical data to apply this model. The accuracy can be further increased with more data instances.

References

1. Wong, N.D., Epidemiological studies of CHD and the evolution of preventive cardiology. *Nat. Rev. Cardiol.* 11(5):276–289, 2014.
2. http://www.who.int/mediacentre/factsheets/fs317/en/ (Accessed on January 2016).
3. Tsipouras, M.G., Exarchos, T.P., Fotiadis, D.I., Kotsia, A.P., Vakalis, K.V., Naka, K.K., and Michalis, L.K., Automated diagnosis of coronary artery disease based on data mining and fuzzy modeling. *IEEE Trans. Inf. Technol. Biomed.* 12(4):447–458, 2008.
4. http://heartdiseaseonline.com (Accessed on November 2015).
5. Acharya, U.R., Faust, O., Sree, V., Swapna, G., Martis, R.J., Kadri, N.A., and Suri, J.S., Linear and nonlinear analysis of normal and CAD-affected heart rate signals. *Comput. Methods Prog. Biomed.* 113(1):55–68, 2014.
6. Giri, D., Acharya, U.R., Martis, R.J., Sree, S.V., Lim, T.C., Ahamed, T., and Suri, J.S., Automated diagnosis of coronary artery disease affected patients using LDA, PCA, ICA and discrete wavelet transform. *Knowl.-Based Syst.* 37:274–282, 2013.
7. http://www.nhlbi.nih.gov/health/health-topics/topics/cad (Accessed on February 2016).
8. Alizadehsani, R., Hosseini, M. J., Sani, Z. A., Ghandeharioun, A., & Boghrati, R., Diagnosis of coronary artery disease using cost-sensitive algorithms. In Data Mining Workshops (ICDMW), 2012 I.E. 12th International Conference on (pp. 9–16). IEEE, 2012.
9. Arafat, S., Dohrmann, M., & Skubic, M., Classification of coronary artery disease stress ECGs using uncertainty modeling. In Computational Intelligence Methods and Applications, 2005 ICSC Congress on (pp. 4-pp). IEEE, 2005.
10. Lee, H. G., Noh, K. Y., & Ryu, K. H., A data mining approach for coronary heart disease prediction using HRV features and carotid arterial wall thickness. In BioMedical Engineering and Informatics, 2008. BMEI 2008. International Conference on (Vol. 1, pp. 200–206). IEEE, 2008.
11. Acharya, U.R., Sree, S.V., Krishnan, M.M.R., Molinari, F., Saba, L., Ho, S.Y.S., and Suri, J.S., Atherosclerotic risk stratification strategy for carotid arteries using texture-based features. *Ultrasound Med. Biol.* 38(6):899–915, 2012.
12. Acharya, U.R., Mookiah, M.R.K., Sree, S.V., Afonso, D., Sanches, J., Shafique, S., and Suri, J.S., Atherosclerotic plaque tissue characterization in 2D ultrasound longitudinal carotid scans for automated classification: a paradigm for stroke risk assessment. *Med. Biol. Eng. Comput.* 51(5):513–523, 2013.
13. Zhao, Z., & Ma, C., An intelligent system for noninvasive diagnosis of coronary artery disease with EMD-TEO and BP neural network. In Education Technology and Training, 2008. and 2008 International Workshop on Geoscience and Remote Sensing. ETT and GRS 2008. International Workshop on (Vol. 2, pp. 631–635). IEEE, 2008.
14. Acharya, U.R., Sree, S.V., Krishnan, M.M.R., Krishnananda, N., Ranjan, S., Umesh, P., and Suri, J.S., Automated classification of patients with coronary artery disease using grayscale features from left ventricle echocardiographic images. *Comput. Methods Prog. Biomed.* 112(3):624–632, 2013.
15. Kim, W. S., Jin, S. H., Park, Y. K., & Choi, H. M., A study on development of multi-parametric measure of heart rate variability diagnosing cardiovascular disease. In World Congress on Medical Physics and Biomedical Engineering 2006 (pp. 3480–3483). Springer: Berlin Heidelberg, 2007.
16. Patidar, S., Pachori, R.B., and Acharya, U.R., Automated diagnosis of coronary artery disease using tunable-Q wavelet transform applied on heart rate signals. *Knowl.-Based Syst.* 82:1–10, 2015.
17. Xing, Y., Wang, J., Zhao, Z., & Gao, Y., Combination data mining methods with new medical data to predicting outcome of coronary heart disease. In Convergence Information Technology, 2007. International Conference on (pp. 868–872). IEEE, 2007.
18. Alizadehsani, R., Habibi, J., Hosseini, M.J., Mashayekhi, H., Boghrati, R., Ghandeharioun, A., and Sani, Z.A., A data mining approach for diagnosis of coronary artery disease. *Comput. Methods Prog. Biomed.* 111(1):52–61, 2013.
19. Karaolis, M.A., Moutiris, J.A., Hadjipanayi, D., and Pattichis, C.S., Assessment of the risk factors of coronary heart events based on data mining with decision trees. *IEEE Trans. Inf. Technol. Biomed.* 14(3):559–566, 2010.
20. Ordonez, C., Association rule discovery with the train and test approach for heart disease prediction. *IEEE Trans. Inf. Technol. Biomed.* 10(2):334–343, 2006.
21. Srinivas, K., Rao, G. R., & Govardhan, A., Analysis of coronary heart disease and prediction of heart attack in coal mining regions using data mining techniques. In Computer Science and Education (ICCSE), 2010 5th International Conference on (pp. 1344–1349). IEEE, 2010.
22. Palaniappan, S., & Awang, R., Intelligent heart disease prediction system using data mining techniques. In Computer Systems and Applications, 2008. AICCSA 2008. IEEE/ACS International Conference on (pp. 108–115). IEEE, 2008.
23. Melillo, P., Izzo, R., Orrico, A., Scala, P., Attanasio, M., Mirra, M., and Pecchia, L., Automatic prediction of cardiovascular and cerebrovascular events using heart rate variability analysis. *PLoS One.* 10(3):e0118504, 2015.
24. Acharya, U.R., Faust, O., Sree, S.V., Molinari, F., Saba, L., Nicolaides, A., and Suri, J.S., An accurate and generalized approach to plaque characterization in 346 carotid ultrasound scans. *IEEE Trans. Instrum. Meas.* 61(4):1045–1053, 2012.
25. Lin, K.C., and Hsieh, Y.H., Classification of medical datasets using SVMs with hybrid evolutionary algorithms based on endocrine-based particle swarm optimization and artificial bee Colony algorithms. *J. Med. Syst.* 39(10):1–9, 2015.
26. Subanya, B., & Rajalaxmi, R. R., Feature selection using Artificial Bee Colony for cardiovascular disease classification. In Electronics

and Communication Systems (ICECS), 2014 International Conference on (pp. 1–6). IEEE, 2014.
27. Amin, S. U., Agarwal, K., & Beg, R., Genetic neural network based data mining in prediction of heart disease using risk factors. In Information & Communication Technologies (ICT), 2013 I.E. Conference on (pp. 1227–1231). IEEE, 2013.
28. Kumar, R., Negi, P.C., Bhardwaj, R., Kandoria, A., Asotra, S., Ganju, N., and Marwah, R., Clinical and non-invasive predictors of the presence and extent of coronary artery disease. *Indian Heart J.* 66:S28, 2014.
29. Eom, J.H., Kim, S.C., and Zhang, B.T., AptaCDSS-E: a classifier ensemble-based clinical decision support system for cardiovascular disease level prediction. *Expert Syst. Appl.* 34(4):2465–2479, 2008.
30. Yeh, D.Y., Cheng, C.H., and Chen, Y.W., A predictive model for cerebrovascular disease using data mining. *Expert Syst. Appl.* 38(7): 8970–8977, 2011.
31. Kupusinac, A., Stokic, E., and Kovacevic, I., Hybrid EANN-EA system for the primary estimation of Cardiometabolic risk. *J. Med. Syst.* 40(6):1–9, 2016.
32. Le Cessie, S., & Van Houwelingen, J. C., Ridge estimators in logistic regression. Applied statistics, 191–201, 1992.
33. Cohen, W. W., Fast effective rule induction. In Proceedings of the twelfth international conference on machine learning (pp. 115–123), 1995.
34. Hühn, J., and Hüllermeier, E., FURIA: an algorithm for unordered fuzzy rule induction. *Data Min. Knowl. Disc.* 19(3):293–319, 2009.
35. Quinlan, J. R., *C4. 5: Program for machine learning Morgan Kaufmann*. San Mateo, CA, 1993.
36. Melillo, P., De Luca, N., Bracale, M., and Pecchia, L., Classification tree for risk assessment in patients suffering from congestive heart failure via long-term heart rate variability. *IEEE J. Biomed. Health Inform.* 17(3):727–733, 2013.
37. Novaković, J., Štrbac, P., & Bulatović, D., Toward optimal feature selection using ranking methods and classification algorithms. *Yugoslav Journal of Operations Research* ISSN: 0354-0243 EISSN: 2334-6043, 21(1), 2011.
38. Guyon, I., and Elisseeff, A., An introduction to variable and feature selection. *J. Mach. Learn. Res.* 3:1157–1182, 2003.
39. Piramuthu, S., Evaluating feature selection methods for learning in data mining applications. *Eur. J. Oper. Res.* 156(2):483–494, 2004.
40. Hall, M. A., *Correlation-based feature selection for machine learning* (Doctoral dissertation, The University of Waikato), 1999.
41. Babaoglu, İ., Findik, O., and Ülker, E., A comparison of feature selection models utilizing binary particle swarm optimization and genetic algorithm in determining coronary artery disease using support vector machine. *Expert Syst. Appl.* 37(4):3177–3183, 2010.
42. Ebenhart, R., Kennedy. Particle swarm optimization. In Proceeding IEEE Inter Conference on Neural Networks, Perth, Australia, Piscat-away (Vol. 4, pp. 1942–1948), 1995.
43. Xue, B., Zhang, M., and Browne, W.N., Particle swarm optimization for feature selection in classification: a multi-objective approach. *IEEE Trans. Cybern.* 43(6):1656–1671, 2013.
44. http://www.cs.waikato.ac.nz/ml/weka/index.html (Accessed on October 2015).
45. Purwar, A., and Singh, S.K., Hybrid prediction model with missing value imputation for medical data. *Expert Syst. Appl.* 42(13):5621–5631, 2015.
46. Kahramanli, H., and Allahverdi, N., Design of a hybrid system for the diabetes and heart diseases. *Expert Syst. Appl.* 35(1):82–89, 2008.
47. Peter, T. J., & Somasundaram, K., An empirical study on prediction of heart disease using classification data mining techniques. InAdvances in Engineering, Science and Management (ICAESM), 2012 International Conference on (pp. 514–518). IEEE, 2012.
48. Bouali, H., & Akaichi, J., Comparative Study of Different Classification Techniques: Heart Disease Use Case. In Machine Learning and Applications (ICMLA), 2014 13th International Conference on (pp. 482–486). IEEE, 2014.