Vol 4, Issue 3, Jul 2017
Journal of Health Management & Informatics
jhmi.sums.ac.ir
Original Article

Real-data comparison of data mining methods in prediction of coronary artery disease in Iran

Azam Dekamin[1*], Ahmad Shaibatalhamdi[2]

Received 3 Dec 2016 ; Accepted 27 Jan 2017

ABSTRACT

Introduction: Cardiovascular diseases are currently of broad prevalence and constitute one of the major causes of mortality in different societies. Angiography is one of the most accurate methods to diagnose heart diseases; it incurs high expenses and comes with side effects. Data mining is intended to enable timely prognosis of diseases with the least expenses possible, making use of the patients' information. The present study aims to provide replies for the question whether it is possible to predict coronary artery diseases with higher efficiency and fewer errors and identify the factors impacting the disease using data mining techniques.

Method: In this study, the data under investigation was collected from a number of 303 persons referring to the heart unit in Shahid Rajaie hospital (Iranian hospital) from 2011 to 2013. It included 54 features. Attempts are made to take advantage of a higher number of characteristics which are helpful for diagnosis of diseases. In addition, Information Gain, Gini, and SVM methods were applied to select influential features, and variables with higher weights were chosen for modeling purposes. In the modeling phase, a combination of classification algorithms and ensemble methods was applied to develop a prediction with fewer errors. Rapid Miner Software was adopted to conduct this study.

Results: Findings of this research indicated that the suggested model, if weighted by SVM index, had the highest efficiency, i.e. 95.83%. This model, moreover, was able to accurately predict all patients with coronary artery disease in Iran. According to the proposed model and obtained accuracies, weighting with SVM was found to be the most effective filtering method, and age as well as typical and atypical chest pain were identified to be the most effective features of coronary artery disease. (Graph 3)

Conclusion: This study can contribute to the diagnosis of influential factors which lead to cardiovascular disease in Iran. Comparison of influential variables showed that chest pain (in its two typical and atypical modes) and patient's age had the highest weight in this study. It demonstrates that coronary artery disease is more likely to happen in older ages. High blood pressure is also an important factor in outbreak of this disease. That is why measures have to be taken to prevent such occurrence. Diabetes constitutes another influential factor in the outbreak of coronary artery disease to which attention should be paid in primary tests.

Keywords: Cardiovascular disease, Coronary Artery Disease, Angiography

►*Please cite this paper as:*
Dekamin A, Shaibatalhamdi A. Real-data comparison of data mining methods in prediction of coronary artery disease in Iran. J Health Man & Info. 2017;4(3):87-94.

Introduction

Coronary Artery Disease (CAD) is a chief cause of mortality in industrial countries. During the past decades, however, CAD outbreak has experienced a downturn as an outcome of improved CAD diagnosis, pursuance, and treatment methods. However, almost one half of CAD-caused mortalities happen in industrial countries and 25% of them in underdevelopment states. While angiography is presently applied to diagnose clogged arteries, it is sought by scholars to be replaced by alternatives due to the complications it poses to the human body and the costs it incurs. Hence, diagnosis of this disease by means of non-invasive methods is both important and helpful (1).

Data mining is the process for finding knowledge out of a huge amount of data saved in databases. Applying some algorithms, data mining intends to discover the relationships and patterns among data. Some applications of data mining lie in banking, discovery of cheats and crimes, marketing, and medication that would lead to reduced risk and cost levels (2).

Physicians are seeking to perform studies on the factors which exacerbate heart diseases, cause them to occur, and ways to diagnose them using methods which present higher accuracy levels and lower complications. To do so, application of data mining methods in medication and healthcare services has experienced a sharp rise, and several studies have been carried out on

[1] *Department of information technology management, Faculty of Management and Economics, Tehran Science and Research Branch, Islamic Azad university, Tehran, Iran*

[2] *Department of industrial management, Firoozkooh branch, Islamic Azad university, Firoozkooh, Iran*

*Corresponding Author: ***A Dekamin,*** *Department of information technology management, Faculty of Management and Economics, Tehran Science and Research Branch, Islamic Azad university, Tehran, Iran, Email: dekamin@gmail.com.*

application of data mining in medical sciences all around the globe. Thus, one of the most popular areas in which data mining has found a broad application is medical fields and healthcare services. In doing so, a wide array of studies has been conducted in different countries regarding diagnosis of heart diseases that have made remarkable contributions to healthcare managers in improving their services (3).

In 2016, Manimekalai proposed "Prediction of Heart Diseases using Data Mining Techniques". The objective of the proposed work was to find the best method of prediction to predict the heart disease and their main objective was to provide a study of Heart Diseases using various Data Mining Techniques. When the data mining technique was used separately, accuracy was low. Then, some data mining techniques (SVM Classifier + Genetic Algorithm) were combined and achieved 95% accuracy (4).

Newton Cheung made use of C4.5 algorithms and a Naive Bayes classifier to categorize the diseases associated with heart and blood vessels and achieved 81.11% and 81.48% accuracy, respectively (5).

Kemal Polat proposed a method called artificial immune system (AIS), achieving an accuracy of 84.5% in his classifications. Applying a method similar to that adopted by Kemal Polat, SalihGunes achieved an accuracy of 87%. Afterwards, other different results were obtained by Weka and RA software, the highest of their accuracy being 77% (6).

In 2015, Ram Bilas Pachori and his colleagues studied and diagnosed heart diseases using Tunable-Q wavelet obtained from heart rate signals. They used LS-SVM method and achieved the accuracy of 96.8%, sensitivity equal to 100%, and specificity of 93.7% (7).

In 2016, Bhalerao and Gunjal proposed "Survey of Heart Disease Prediction Based on Data Mining Algorithms". They compared single data mining techniques and hybrid models in order to select better methods for predicting heart diseases. (8). Also, in 2014, Nihat Yilmaz and Onur Inan proposed a new data preparation method based on clustering algorithms for diagnosis of heart and diabetes diseases (9).

In 2015, Ankur Makwana and Jaymin Patel used "Decision Support System to diagnosis of heart diseases". In the proposed work, they designed a predictive model for heart disease detection using Machine Learning and Data Mining techniques. They showed that fuzzy model had the capacity to evaluate the connections between the input of predicted patients and predicted outcomes of patients' results. This method can stabilize framework variables and the admission of a patient (10).

Also, in 2015, Kemal Akyol and Elif Calik used random forest classification method and they obtained the ratio of the correct classification to be 97.72% (11).

In 2015, Georgeena used Apriori algorithm in order to diagnose heart diseases. The main purpose of his work was to reduce the number of attributes that would result in the number of tests which patients should take. (12).

Moreover, in 2015, Parchi Paliwal and Mahesh Malviya showed that the percentage of classifying correct records by the proposed method is more as compared to the previous methods (13).

In 2015, Moloud Abdar, Sharareh compared the performance of data mining algorithms in UCI dataset. They realized that C5.0 decision tree was able to build a model with greatest accuracy since the model prediction accuracy was 93.02% (14).

In 2016, Gunsai Pooja and Lolita Singh's conducted a study entitled "A Review on Data Mining for Heart Disease Prediction". They used data mining techniques in order to discover knowledge. The system used medical terms such as sex, blood pressure, cholesterol, etc. like attributes to predict the likelihood of patients getting a heart disease. The performance of these techniques was compared, based on accuracy. The main objective of this research was to develop a prototype Intelligent Heart Disease Prediction System (IHDPS) using data mining modeling technique, namely Clustering. It could discover and extract the hidden

knowledge (patterns and relationships) associated with heart disease from a historical heart disease database (15).

In 2016, Verma, Luxmi and Sangeet Srivastava proposed a hybrid model to predict a coronary artery disease (16).

In the same line in 2012, ZamanPour examined and compared traditional data mining techniques in prediction of heart diseases (17). Also, in 2014, Safdari and Ghazi Saeidi made an attempt to compare the performance of decision tree and neural network in prediction of infection by myocardial infarction, using K-means clustering technique (18).

KeyvanPour and Khalatbari compared classification algorithms in diagnosis of diabetes and heart failure. As cited above, nosocomial large data usually includes useful information on demographic characteristics of patients, on the one hand, and features relevant to the manner by which they are treated, on the other. Studies under investigation indicated that some variables such as EF, Region RWMA, Q Wave, and Twave inversion applied here intending to diagnose CAD have been either left unheeded or less noticed (19).

Methods

In this study, the data under investigation were collected from 303 persons referring to the heart unit in Shahid Rajaie hospital from 2011 to 2013. It includes 54 features containing 53 features of disease symptoms and 1 attribute of disease diagnosis called class field. The value zero indicates absence of the disease and the value one shows existence of disease. Features of disease symptoms are categorized into four classes: demographic information of patients, symptoms and impacts investigable by physicians, electrocardiogram features, and laboratorial post-echo features. Features of this study are rarely adopted by prior research.

Pre-processing and preparation of data

This phase is one of the most important and time-consuming stages of research. Since information achieved from this phase is served as inputs for further phases and low-quality data would result in low-grade outcomes, pre-processing stage was undertaken in order to guarantee the accuracy of data (20). The data applied in this study

were devoid of lost values. Therefore, outlying data was detected using the two methods: Local outlier factors detection and Distanced base outlier detection. Parameter adjustment was applied in this phase to make sure of omission of outlying data.

Feature Subset Selection

One of the important factors in development of a high-accuracy model is proper selection of the features. To reach this purpose, it is mandatory to select a subset of desirable features and finding suitable characteristics. In this study, filtering methods such as weighting by Information Gain, Gini, and SVM were applied, and influential features were selected and inserted into the model.

Weighting Based on Gini Index: A subset of features with the highest influence was selected to participate in modeling. After parameter adjustment, 10 features were selected.

$$GINI(t) = 1 - \sum_{i} [p(j|t)]^2 \; GINI_{split} = \sum_{i-1}^{k} \frac{n_i}{n} GINI(i)$$

Weighting Based on SVM Index: Using coefficients of normal vector of SVM, one line as a feature weight is responsible for weighting procedure. After parameter adjustment, a number of 17 features were selected to take part in the modeling process.

Weighting Based on Gain Index: This criterion is one of the most well-known criteria that is applied for weighting, which makes use of another criterion called entropy. After parameter adjustment, a number of 10 features were selected.

$$Entropy(t) = -\sum_{j} p(j|t) \log p(j|t) \; GAIN_{split} = Entropy(p) - \left(\sum_{i-1}^{\kappa} \frac{n_i}{n} Entropy(i) \right)$$

Testing Stage

The present methodology is of practical-quantitative type. To categorize the educational and experimental data, we applied the Hold Out method: 80% of data was used for education stage and the rest 20% for test phase. Culled data was investigated using Rapid Miner Software. After primary examinations, it was revealed that the data was unbalanced, and a number of 87 records had the objective variable 0 and 216 of them had the objective variable 1. Therefore, sub-sampling method was adopted for balancing data in order to achieve better results. In classification methods, one field is held as output field. Here, the variable Cath was regarded as output and 53 other variables as inputs. After cleaning and selecting the package features, 10 and 17 variables were introduced into the model, and modeling was completed using stacking method. Stacking is composed of two stages: basic learners in level 0 and stacking model learners in level 1. In level 0, the algorithms decision tree, K-nearest neighbor, and Naive Bayes were applied. Outputs of each model were taken for construction of new dataset. Afterwards, in level 1, this new dataset was applied by stacking model learners that was a random forest model.

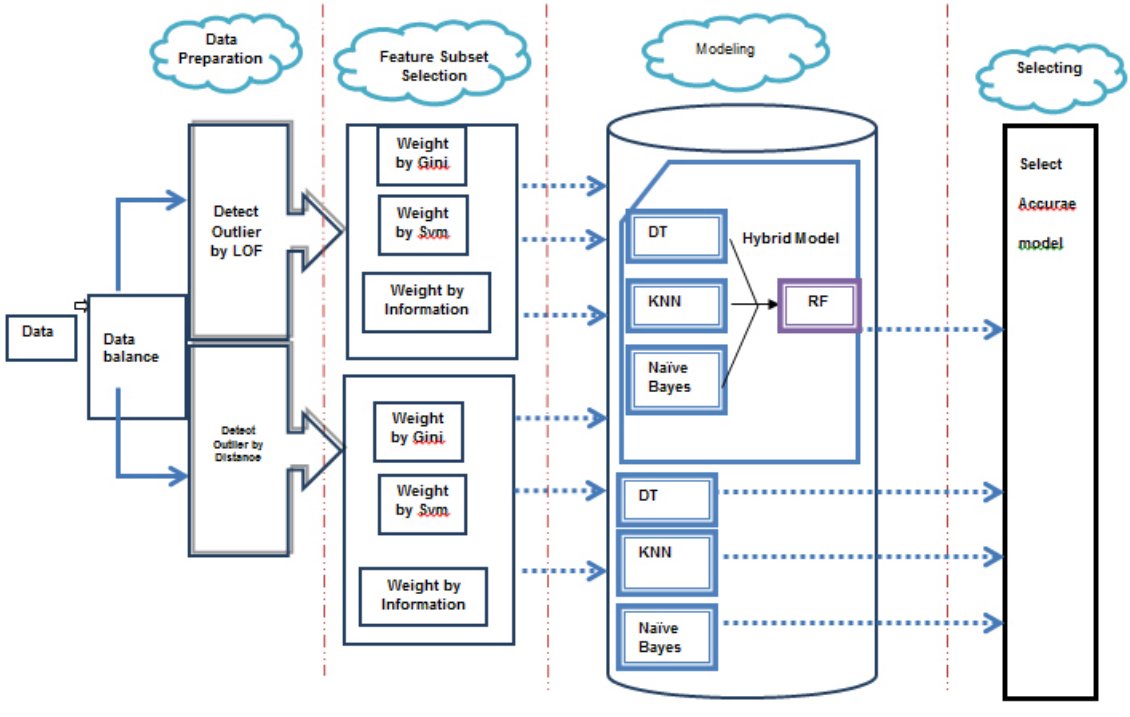
Evaluation

Evaluation criteria F-measure, sensitivity, specificity, recall, precision, and accuracy relevant to the research model were measured and cited in Tables 2 and 3.

Confusion Matrix

The Confusion Matrix illustrates the manner the classification technique performs with respect to input dataset as separated by different types of classification issue. The concepts TN, FP, FN, and TP in this matrix are defined as follows:

Figure 1. Puroposed Model of This study



TN: This value indicates the number of records whose real class was negative, and classification algorithm could accurately diagnose their class to be negative.
FP: This value indicates the number of records whose real class was negative, and classification algorithm wrongly diagnosed their class as positive.
FN: This value indicates the number of records whose real class was positive, and classification algorithm wrongly diagnosed their class as negative.
TP: This value indicates the number of records whose real class was positive, and classification algorithm could accurately diagnose their class to be positive.

Table 1. Cofusion Matrix

	PREDICTED CLASS		
ACTUAL CLASS		Class=Yes	Class=No
	Class=Yes	a (TP)	b (FN)
	Class=No	c (FP)	d (TN)

Now, different types of important evaluation criteria pertinent to classification technique with regard to confusion matrix are expounded. Accuracy of the model shows that all items that have to be observed by a classification in order to possess a suitable performance are considered in the criterion Accuracy. Clearly, TP and TN are the most important values that have to be maximized in a problem. Since both TP and TN are located in nominator, it is safe to assert that all classes existing in classification problem are considered in the above relation. Thus, the criterion Accuracy of Classification is the commonest and most renowned criterion for calculation of efficiency of classification algorithms. The criterion Recall(x) expresses the accuracy of classification of class x with respect to all records with the label x. The criterion Precision(x) articulates the accuracy of classification x with regard to the whole items for which the label x has been proposed by classifier for the examined record. Remember that the criterion Recall(x) expresses the efficiency of classifier with regard to the number of events of class x. However, the criterion Precision(x) is basically grounded upon prediction accuracy of classification, demonstrating the degree by which the outputs of classification are reliable. Another important point is that denominator of Recall(x) relations should be held as equal to the total number of records with label x in classifications where the label of some records is determined to be unknown. The criterion F-Measure(x) exemplifies a combination of the criteria Precision(x) and Recall(x) that is utilized when special importance could be attached neither to Precision(x)nor Recall(x).

$$Accuracy = \frac{a+d}{a+b+c+d} = \frac{TP+TN}{TP+TN+FP+FN}$$

$$Precision\ (p) = \frac{a}{a+c}$$

$$Recall\ (r) = \frac{a}{a+b}(F-measure) = \frac{2rp}{r+p} = \frac{2a}{2a+b+c}$$

$$\text{Sensitivity} = \frac{a}{a+d} = \frac{TP}{TP+FN} \quad \text{Specificity} = \frac{d}{b+c} = \frac{TN}{TN+FP}$$

Results

This section presents experimental results using our model on the Dataset which was randomly selected from among medical cases of 303 patients referring to Tehran Shahid Rajaee Heart Center, from which 216 subjects were infected with CAD and the rest were not; this was described in the previous section. As we mentioned in the Methods section, features used in this study have rarely been adopted by prior research.
Our findings are shown in several tables and graphs. The performance of the proposed hybrid model was compared with some algorithms. These algorithms were also used in the hybrid model as learners of the model and after that the results were recorded. One of the goals of the research was to prove that single separated algorithms were not as accurate as hybrid models.
In this research, primary studies showed that a number of 87 records had the objective variable 0 and 216 of them had the objective variable 1. It is obvious from these observations that data were unbalanced, and improved results could be achieved by balancing methods. Hence, data were balanced using a sub-sampling method, followed by outlier detection using the two methods, Distance and LOF.
In this research, due to the reduction in the number of features and the computational time and also the cost of the experiments, the most effective features which increase the risk of heart disease were extracted and selected using Feature subset selection, as cited above. The three methods, weight by Information Gain, weight by Gini, and weight by SVM, were applied to select the features influential in the model. In each method, a number of 10 features with higher weights were selected for modeling process after the parameters were regulated.
Afterwards, data were prepared for introduction into modeling section where training and testing data were divided with a ratio of 80 to 20. In this phase, firstly Naive Bayes algorithm, decision tree, and K-nearest neighbor were applied for modeling in a separated manner, as shown in Fig. 1. Relevant results are displayed in Table 3. From these results, it is shown that the accuracy of Naïve Bayes is 92.00 which is the most accurate algorithms between Naïve Bayes, decision tree and K-nearest neighbor. The recall index is 77.78, meaning that we can predict 77.78% of the patients with coronary artery disease in that hospital. One of the most important goals of this research is to predict the highest rate of patients.
Secondly, the hybrid stacking model with the algorithms Naive Bayes, decision tree, and K-nearest neighbor was utilized as basic learner in level 0 and random forest algorithm in level 1. Modeling outcomes are cited in Table 3. It is certainly true that the hybrid model can diagnose the cardiovascular disease with an accuracy of 95.83 which can increase the accuracy in comparison with the separated model mentioned in last paragraph. It is certainly the case that with this hybrid model, we can predict all (100%) the patients with the heart disease (CAD=1). It is a great progress in this field because when we can differentiate between patients and others, we can decrease the cost of experiments and also the rate of

mortality. We tried to cite all results with different feature subset selections and also outlier detection methods which have a profound impact on the results. It means that we tried to test a lot of possibilities which may increase the accuracy.

Graph 1 represents a comparison of accuracies of the selected algorithms in this research. As shown, the suggested method is ranked first among other algorithms.

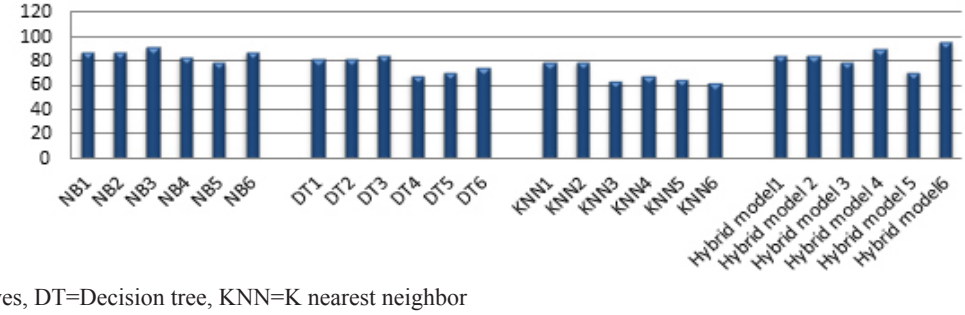
Table 2. The performance of Hybrid Model

Numbers	Modeling Algorithms	Outlier Detection	Feature Subset Selection	Accuracy	Precision	Recall	F-measure	Sensitivity	Specificity
1	Hybrid Model	LOF	Gini	85.29	84.62	78.57	81.48	78.57	90.00
2	Hybrid Model	LOF	Info Gain	85.29	84.62	78.57	81.48	78.57	90.00
3	Hybrid Model	LOF	Svm	79.41	70.59	85.71	77.42	85.71	75.00
4	Hybrid Model	Distance	Gini	90.62	83.33	87.50	90.32	87.50	93.75
5	Hybrid Model	Distance	Info Gain	71.88	88.89	50.00	64.00	50.00	93.75
6	Hybrid Model	Distance	Svm	95.83	92.31	100.00	96.00	100.00	91.67

Table 3. The performance of Naïve Bayes, Decision tree and K nearest neighbor algorithm

Numbers	Modeling Algorithms	Outlier Detection	Feature Subset Selection	Accuracy	Precision	Recall	F-measure	Sensitivity	Specificity
1	Naïve Bayes	LOF	Gini	88.00	87.50	77.78	82.35	77.78	93.75
2	Naïve Bayes	LOF	Info Gain	88.00	87.50	77.78	82.35	77.78	93.75
3	Naïve Bayes	LOF	Svm	92.00	100.00	77.78	87.50	77.78	100.00
4	Naïve Bayes	Distance	Gini	83.33	83.33	83.33	83.33	83.33	83.33
5	Naïve Bayes	Distance	Info Gain	79.17	76.92	83.33	80.00	83.33	75.00
6	Naïve Bayes	Distance	Svm	87.50	80.00	100.00	88.89	100.00	75.00
7	Decision tree	LOF	Gini	82.35	83.33	71.43	76.92	71.43	90.00
8	Decision tree	LOF	Info Gain	82.35	83.33	71.43	76.92	71.43	90.00
9	Decision tree	LOF	Svm	85.29	90.91	71.43	80.00	71.43	95.00
10	Decision tree	Distance	Gini	68.75	75.00	56.25	64.29	56.25	81.25
11	Decision tree	Distance	Info Gain	71.88	81.82	56.25	66.67	56.25	87.50
12	Decision tree	Distance	Svm	75.00	75.00	75.00	75.00	75.00	75.00
13	KNN	LOF	Gini	79.41	73.33	78.57	75.86	78.57	80.00
14	KNN	LOF	Info Gain	79.41	73.33	78.57	75.86	78.57	80.00
15	KNN	LOF	Svm	64.71	57.14	57.14	57.14	57.14	70.00
16	KNN	Distance	Gini	68.75	71.43	62.50	66.67	62.50	75.00
17	KNN	Distance	Info Gain	65.62	66.67	62.50	64.52	62.50	68.75
18	KNN	Distance	Svm	62.50	58.82	83.33	68.97	83.33	41.67

Graph 1. Comparison of precision index of algorithms



NB=Naïve Bayes, DT=Decision tree, KNN=K nearest neighbor

Graph 2 compares the criteria Recall and F-Measure for evaluation of algorithms. As Table 3 shows, our index Recall in this model is equal to 100, meaning that we have managed to rightly predict all individuals with CAD.

The relative importance of each variable in evaluating the model is associated with the importance of each feature in making a prediction, and it is not related to the model accuracy. In this research, three methods were used to determine the influential factors. The three methods of weight by Information Gain, weight by Gini, and weight by SVM were examined, and SVM index was found to present better results in weighting variables. For assurance, after these methods were used, the researcher made an attempt to find overlaid elements which were mutual in all methods. As shown in the Graphs, the common agents found as influential factors in all feature selection methods are typical and atypical chest pain and patient's age. Physicians are, thus, recommended to include these items in their examination checklist. Getting these factors from patients may decrease the cost of experiments as patients are not required to do angiography to be diagnosed as a heart disease patient.

According to the Graph 3, influential variables in this research are obtained in weight by SVM. As it is observed, chest pain (in its two typical and atypical modes) and patient's age had the highest weight in this study, showing that CAD is more likely to happen in senility.

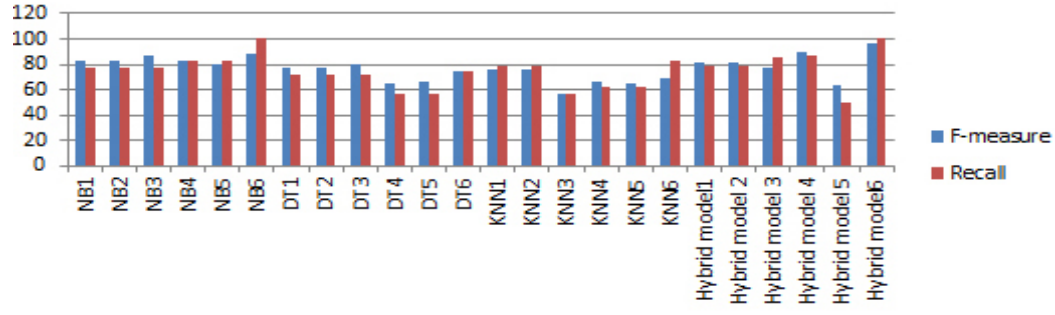
High blood pressure is also an important factor in outbreak of this disease. Electrocardiogram features, such as Region with RWMA, are also among the influential features. It is important to note that diabetes (DM) is also among significant features of this disease.

For further assurance, other filtering methods such as weight by Information Gain and Gini were also applied in this study, and the results are shown in the following graphs. According to Graph 4, typical chest pain is on top of the list. Atypical chest pain and age are the two other important features which have mostly to do with the disease. Electrocardiogram features such as EF are also among the influential features. Hypertension has also a profound effect on this disease.

According to Graph 5, typical chest pain is on top of the list. Atypical chest pain and age are the two other important features which have mostly to do with the disease. Electrocardiogram features such as EF are also among the influential features. Hypertension has also a deep effect on this disease. According to the Graphs 4 and 5, the results of these two methods (weight by Gini, weight by Info Gain) are similar.

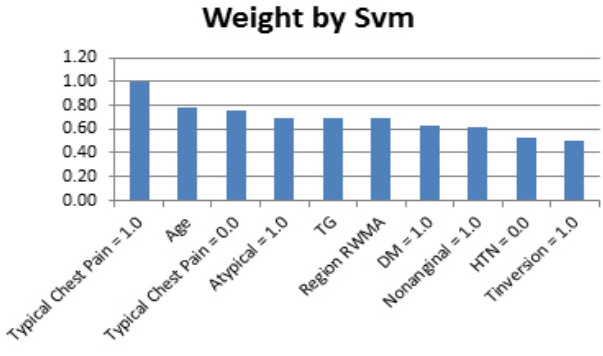
It is obvious from these observations that all these methods can help to sort the important factors, but when we used selected factors in the hybrid model, it was revealed that with weight by svm technique, the results were better than other techniques.

Graph 2. Comparison of the criteria Recall, and F-Measure for evaluation of algorithms



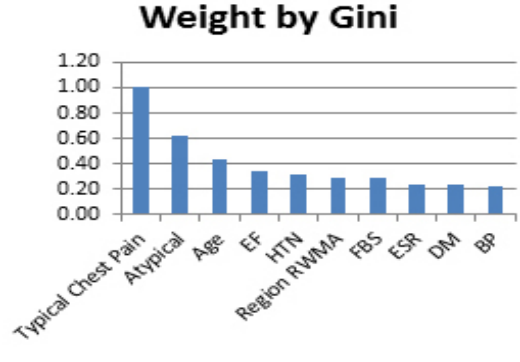
NB=Naïve Bayes, DT=Decision tree, KNN=K nearest neighbor

Graph 3. Comparison of variables influential in model using weighting with SVM index



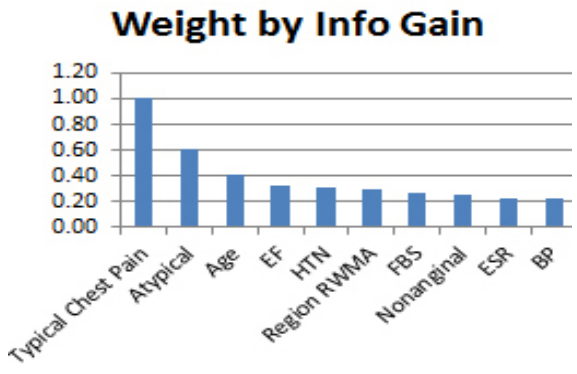
TG=triglyceride, DM=Diabetes, HTN=hyper tension, FBS=fasting blood sugar, Region RWMA=regional wall motion abnormality, T inversion=ECG feature

Graph 4. Comparison of variables influential in model using weighting with Gini index



DM=Diabetes, HTN=hyper tension, FBS=fasting blood sugar, ESR=erythrocyte sedimentation rate, Region RWMA=regional wall motion abnormality, BP=blood pressure, EF =ejection fraction

Graph 5. Comparison of variables influential in model using weighting with Gain index



TG=triglyceride, DM=Diabetes, HTN=hyper tension, BS=fasting blood sugar, ESR=erythrocyte sedimentation rate, FH=family history, Region RWMA=regional wall motion abnormality, BP=blood pressure, EF =ejection fraction

Discussion

In this study, the likelihood of CAD affliction was attempted to be predicted using data mining algorithms. According to Table 3, hybrid model with Naive Bayes learners, decision tree, and K-nearest neighbor at level 0 and random forest learner at level 1 had the highest accuracy and highest Recall index value. With an accuracy of 95.83% and Recall value of 100, the above method could well predict patients with CAD. In this study, attempts were made to take advantage of ensemble algorithms that were less applied beforehand. For level-0 and level-1 learning, the algorithms were adopted that were alone able to arrive at better results in previous research. An aggregation of above algorithms was arranged aiming at reaching more efficient outcomes and fewer errors.

Zaman Pour and Shamsi (17) examined and compared the accuracy level of data mining algorithms in prediction of heart diseases and achieved considerable outcomes thereof. Efficiency of Naive Bayes algorithm and decision tree was, as observations corroborate, higher than traditional methods under different conditions. Analysis of efficiency of these algorithms was performed based on their accuracy.

Existence of suitable data and proper pre-processing as well as application of ensemble data mining algorithms would provide good outcomes on medical data. In pre-processing phase, it was made clear that data were unbalanced. To acquire better results in pre-processing phase, therefore, we performed data balancing—the point which was mostly left unnoticed in prior studies. After selection of Disease Diagnosis as the target field and 53 features from the four categories of demographical information of patients, symptoms and impacts investigable by physicians, electrocardiogram features, and laboratorial post-eco features, it was observed that variables such as typical and atypical chest pain and patient's age were influential variables and attention to them would result in reduced degrees of affliction to this disease.

Conclusion

One of the most difficult parts of the research was collecting, preparing and categorizing data from the patient's medical files. The features included in this dataset were possible indicators of CAD, according to our medical knowledge. In this study, several algorithms were applied on the dataset and the results were discussed above. One of the goals of this study was to prove that our hybrid model was more accurate than single separated algorithms. It is evident that the accuracy of the hybrid model is 95.83 which cannot be achieved by other single separated algorithms. It is certainly the case that we can predict the status of all (100%) those with cardiovascular disease (CAD=1). It is a big step toward decreasing the cost, time and side effects of diagnosis of heart disease. In addition, data mining techniques including feature selection are used to improve the accuracy. It can contribute to diagnosis of influential factors which lead to cardiovascular disease in Iran. Finally, larger datasets, more features and also broader data mining approaches could be used to achieve better and more interesting results.

Conflict of Interest

None declared.

References

1. Nahar J, Imam T, Tickle KS, Chen Y-PP. Association rule mining to detect factors which contribute to heart disease in males and females. Expert Systems with Applications. 2013;40(4):1086-93.
2. Alizadehsani R, Habibi J, Hosseini MJ, Mashayekhi H, Boghrati R, Ghandeharioun A, et al. A data mining approach for diagnosis of coronary artery disease. Comput Methods Programs Biomed. 2013 Jul;111(1):52-61.
3. Han J, Pei J, Kamber M. Data mining: concepts and techniques: Elsevier; 2011.
4. Manimekalai K. Prediction of Heart Diseases using Data Mining Techniques. International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) Vol. 2016;4.
5. Cheung N. Machine learning techniques for medical analysis: University of Queenland.
6. Polat K, Gunes S. A hybrid approach to medical decision support systems: combining feature selection, fuzzy weighted pre-processing and AIRS. Comput Methods Programs Biomed. 2007 Nov;88(2):164-74.
7. Padmavathi K, Ramakrishna KS. Detection of Atrial Fibrillation using Autoregressive modeling. International Journal of Electrical and Computer Engineering (IJECE). 2015;5(1):64-70.
8. Bhalerao S, Gunjal DB. Survey Of Heart Disease Prediction Based On Data Mining Algorithms. Vol-2 Issue-2. 2016.
9. Yilmaz N, Inan O, Uzer MS. A new data preparation method based on clustering algorithms for diagnosis systems of heart and diabetes diseases. J Med Syst. 2014 May;38(5):48.
10. Patel J, Makwana A. Decision Support System for Heart Disease Prediction using Data Mining Techniques. International Journal of Computer Applications. 2015;117(22):1-5.
11. Akyol K, Çalik E, Bayir Ş, Şen B, Çavuşoğlu A. Analysis of Demographic Characteristics Creating Coronary Artery Disease Susceptibility Using Random Forests Classifier. Procedia Computer Science. 2015;62:39-46.
12. Georgeena T, Thomas S, Siddhesh S, Budhkar, Siddhesh K, Cheulkar, et al. Heart Disease Diagnosis System Using Apriori Algorithm. 2015;5(2).
13. Paliwal P, Malviya M. An efficient method for predicting heart disease problem using fitness value. International Journal of Computer Science and Information Technologies. 2015;6(2):1290-3.
14. Abdar M, Kalhori SRN, Sutikno T, Subroto IMI, Arji G. Comparing Performance of Data Mining Algorithms in Prediction Heart Diseases. International Journal of Electrical and Computer Engineering (IJECE). 2015;5(6):1569-76.

15. Dineshgar GP. A Review on Data Mining For Heart Disease Prediction ISSN: 2278 - 909X. International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE). 2016;5(2).
16. Verma L, Srivastava S, Negi PC. A Hybrid Data Mining Model to Predict Coronary Artery Disease Cases Using Non-Invasive Clinical Data. J Med Syst. 2016 Jul;40(7):178.
17. Zamanpoor S, Shamsi M. Comparing datamining's algorithms validity in predicting heart disease. 4th Iranian Conference on Electrical and Electronics Engineering (ICEEE2012)2012.
18. Safdari R, Ghazi Saeedi M, Gharooni M, Nasiri M, Arji G. Comparing performance of decision tree and neural network in predicting myocardial infarction. Journal of Paramedical Science and Rehabilitation (JPSR). 2014;2(3).
19. Keyvanpour M, Khalatbari L. Comparing classification algorithms in diagnosing diabetes and heart disease. 3rd datamining conference; Tehran2010.
20. Rao VS, Kumar MN. A new intelligence-based approach for computer-aided diagnosis of Dengue fever. IEEE Trans Inf Technol Biomed. 2012 Jan;16(1):112-8.