ACSIJ Advances in Computer Science: an International Journal, Vol. 5, Issue 4, No.22 , July 2016
ISSN : 2322-5157
www.ACSIJ.org

ACSIJ
WWW.ACSIJ.ORG

Timely diagnosis of the heart disease using data mining algorithm

Mohammad Azimimehr[1],Hamid Reza Sahebi[2]

[1]Department of computer science, Islamic Azad University, Ashtian Branch, Ashtian, Iran
M.azimimehr@Yahoo.com

[2]Department of Mathematics, Islamic Azad University, Ashtian Branch, Ashtian, Iran
sahebi@aiau.ac.ir

Abstract

In the current state of medical knowledge, we are facing abundant data collection about various diseases. Investigating these data and obtaining useful results and patterns in conjunction with diseases, is the main propose for using them. In this paper, considering the significant share of the heart disease in human mortality, a variety of data mining methods and algorithms are applied for timely diagnosis of heart diseases. For this purpose, Dezful Hospital standardized data collection for patients with heart disease, have been used. Most important variables of this collection are exercise-induced angina, type of chest pain, age, maximum heart rate and blood pressure at rest. Classification algorithms of K-NN, PSO, Bagging and SVM + PSO hybrid algorithm, are tested and evaluated on the data sets. Based on our evaluations, the proposed hybrid algorithm, with accuracy of 94.74, had the highest accuracy.

Keywords: *Classification algorithms; K-NN; Bagging; SVM-PSO hybrid algorithm*

1. Introduction

Nowadays heart diseases and specially heart attacks are the most common mortality cause in the human communities. Medical environment is a field full of information yet poor in knowledge. Medical diagnoses are made mostly based on doctors' expertise and experience; however, there are still reports about medical errors in the disease diagnosis. Heart disease is one of the most important diseases. The dramatic increase in cardiovascular diseases and their effects and high costs for the community, caused medical community to develop programs for further investigation, prevention, early detection and effective treatment for this dilemma. Therefore, using data mining and knowledge extraction in coronary care units system can provide such a valuable knowledge that could be applied by care center managers to improve service quality used by physicians to predict the future behavior of cardiovascular patients based on previous data. In this way, also, diagnosis of heart disease based on different symptoms and features, and evaluation of risk factors that increase heart attack, are among the most important applications of data mining and knowledge discovery in cardiovascular data systems.

In this paper, we are trying to find a method for detection and prognosis of heart disease by using data mining techniques. To this end, the traditional data mining techniques such as the K-NN, PSO, Bagging classification algorithms are studies and their performance on the existing database, are compared. However, in practice, none of these techniques alone can be considered as the best method. Therefore, we developed and implemented a hybrid algorithm, and tried to present a new and more efficient method to predict and classify patients with heart disease[1].

In this study, due to the numerous applications of data mining in medicine, especially in the diagnosis of diseases, we seek to achieve the following goals:

- Acquiring new database of heart patients' specifications.
- Early detection and prognosis of heart attacks using data mining techniques
- Achieving a hybrid and improved algorithm, with a superior accuracy and precision, compared to other existing algorithms.
- Making these techniques applicable in heart clinics and hospitals.

The remaining sections of this paper, has been organized as follows: in the second section, the project data set is presented. Third section is assigned to the molding and the results of different algorithms are demonstrated and compared. Finally, the article conclusions are presented in the fourth section.

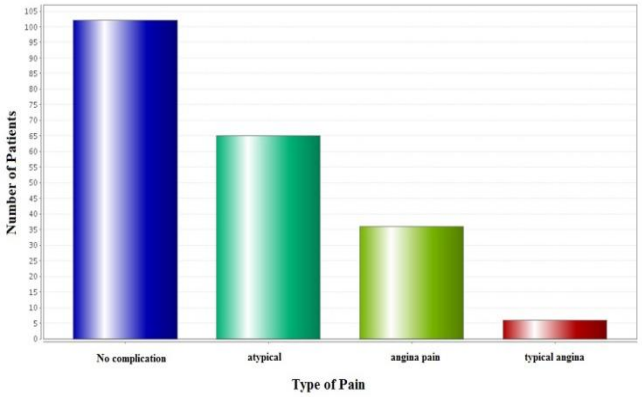
2. The data set

The data used in this study, have been acquired from Dezful hospital's database for patients with heart disease. In this database, the following factors are measured and recorded in each patient. Therefore, our goal will be developing a new method to predict the risk of heart disease positive diagnosis, based on these factors.

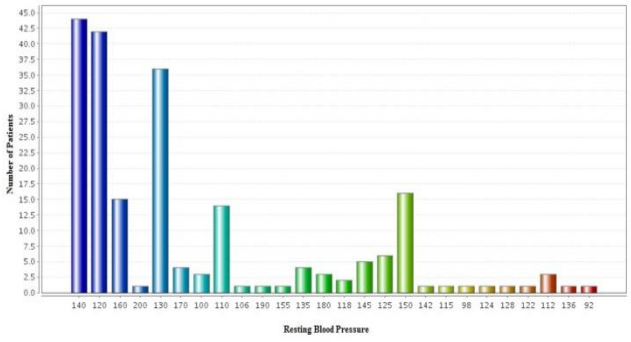
- Age

- Type of chest pain.
- Resting blood pressure
- Blood sugar
- ECG at rest
- Maximum heart rate
- Exercise-induced angina
- Diagnosis

Figure 1 shows the diagrams of two of these recorded parameters. As can be seen, in case of the "Type of chest pain", without complication, atypical angina, angina pain and typical angina conditions, respectively, are the most common conditions in patients. In the blood pressure case, 140 mm Hg pressure has the highest frequency. Figure 2 shows the maximum heart rate for these patients, which the maximum recorded heart rate, were 150 and 140 beats per minute, respectively.



(a)



(b)

Fig. 1 Recorded parameters for patients with heart disease. a) Type of chest pain and b) blood pressure.

Now, we further explore the dataset to examine the relationships between fields and use them in the model structure. As these are the input data of the project, the higher precision they have, the more accurate the output will be.

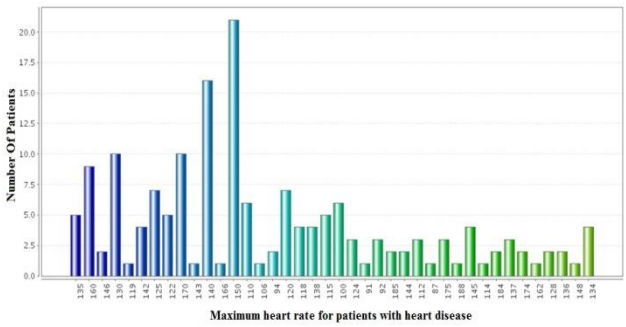


Fig. 2 Maximum heart rate for patients with heart disease

2.1 Prognostic factors relationships

In Figures 3-8, the relationship between prognostic parameters and positive diagnosis of heart disease in patients, has been studied. As it is evident in these figures, most people whom diagnosis were positive, have an age between 37-60 years, asymptomatic chest pain, resting blood pressure between 120 and 150, normal condition for the ECG at rest, maximum heart rate of 95, 115, 120 to 140, and exercise-induced angina.

3. Modeling

After studying the data and preparing them, now we can start the modeling phase. In the first step, we choose the appropriate technique which is a very crucial. Then, the necessary model parameters must be specified. After selecting the model and determine the parameters, small parts of the project will be defined, implemented at every step and carefully tested in order to the quality of the created model be insured. At this point, if the model lacks the acceptable accuracy or its quality is not satisfactory, first we will change the model parameters and re-examine our test model. If the necessary quality is not obtained yet, we should change the model and create a new one.
Classification and prediction are two types of operations for data analysis and modeling, which are used to classify data sets and to understand and predict their future behaviors. The classification models are applied in the analysis of discrete and categorical data, and prediction or regression models are often used in continuous data. The existing data are divided into two groups: training and testing. The training data will be used by system to learn the rules, and testing data are employed to examine the accuracy of the model. In the current work, K-nearest-neighbor, Bagging, PSO, and proposed hybrid algorithms are used.

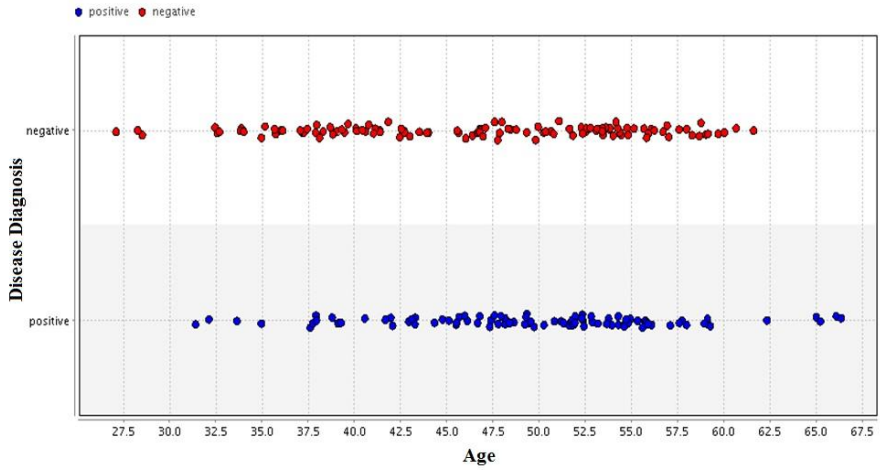


Fig. 3 The relationship between age and disease diagnosis

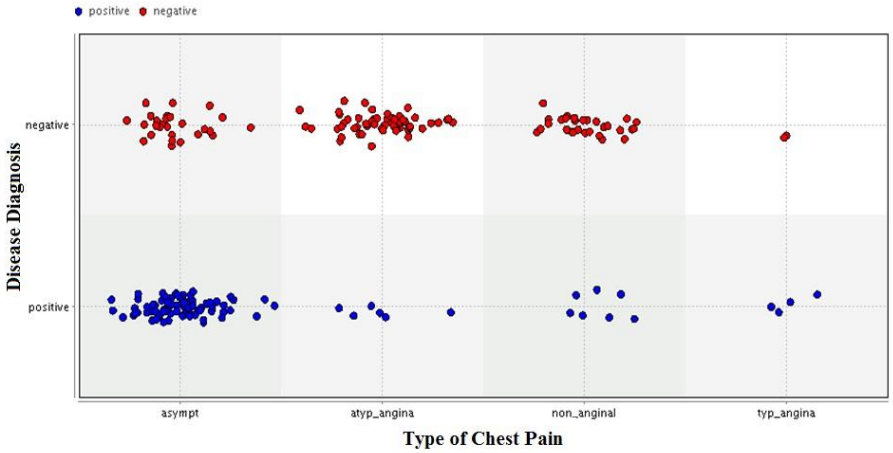


Fig. 4 The relationship between type of chest pain and diagnosis

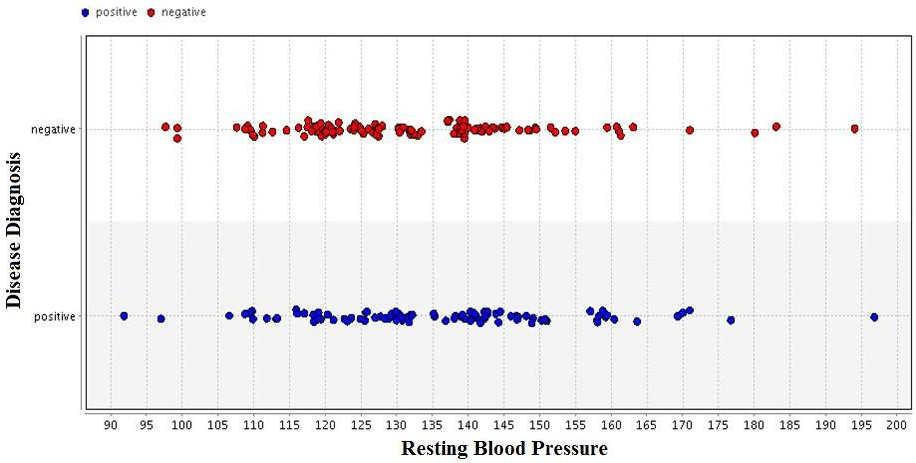


Fig. 5 The relationship between blood pressure at rest and diagnosis

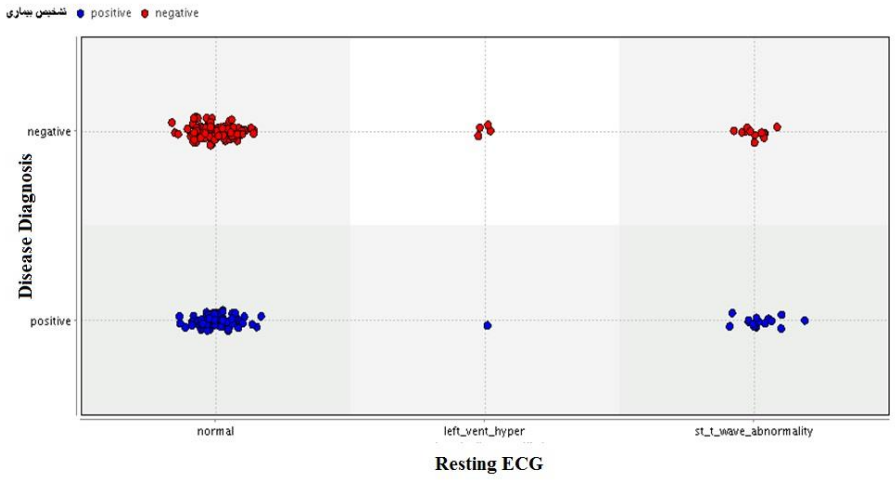


Fig. 6 The relationship between resting ECG and diagnosis.

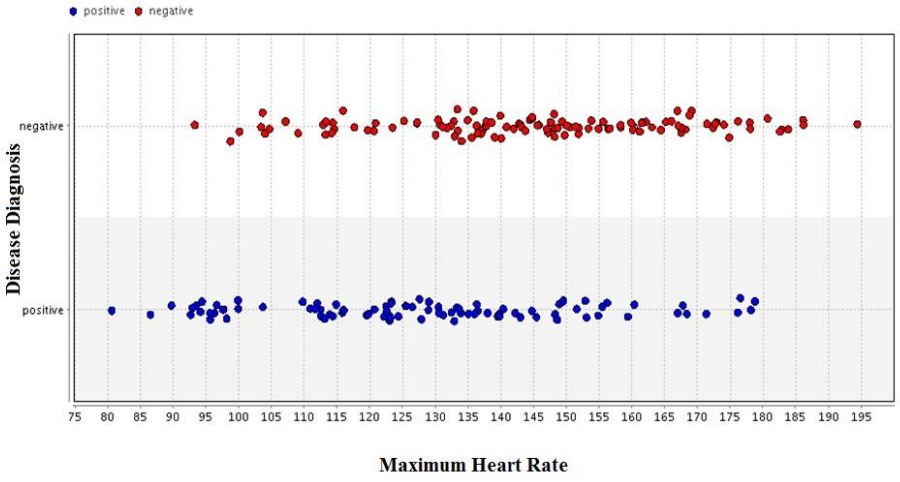


Fig. 7 The relationship between maximum heart rate and diagnosis.

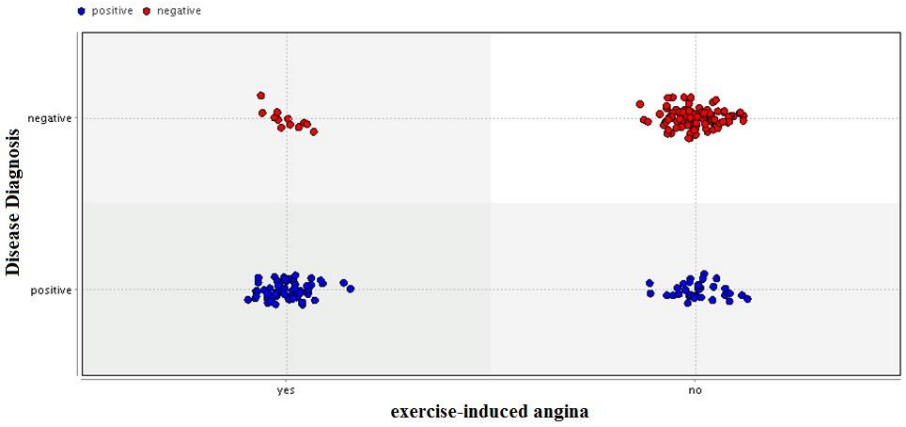


Fig. 8 The relationship between the exercise-induced angina and diagnosis

3.1 PSO Algorithm

The concept of PSO was first suggested by Kennedy and Eberhart [2]. Since its development is 1995, PSO has emerged as one of the most promising optimizing technique for solving global optimization problems. Its mechanism is inspired by the social and cooperative behavior displayed by various species like birds, fish etc including human beings [Particle Swarm Optimization: Performance Tuning and Empirical Analysis].

The working of the Basic Particle Swarm Optimization (BPSO) may be described as: For a D-dimensional search space the position of the ith particle is represented as $X_i = (x_{i1}, x_{i2}, \dots x_{iD})$. Each particle maintains a memory of its previous best position $P_{besti} = (p_{i1}, p_{i2} \dots p_{iD})$. The best one among all the particles in the population is represented as $P_{gbest} = (p_{g1}, p_{g2}, \dots, p_{gD})$. The velocity of each particle is represented as $V_i = (v_{i1}, v_{i2}, \dots v_{iD})$. In each iteration, the P vector of the particle with best fitness in the local neighborhood, designated g, and the P vector of the current particle are combined to adjust the velocity along each dimension and a new position of the particle is determined using that velocity. The two basic equations which govern the working of PSO are that of velocity vector and position vector given by:

$$v_{id} = wv_{id} + c_1 r_1 (p_{id} - x_{id}) + c_2 r_2 (p_{gd} - x_{id})$$

$$x_{id} = v_{id} + x_{id}$$

The first part of equation (1) represents the inertia of the previous velocity, the second part is the cognition part and it tells us about the personal experience of the particle, the third part represents the cooperation among particles and is therefore named as the social component. Acceleration constants c_1, c_2 and inertia weight w are the predefined by the user and r_1, r_2 are the uniformly generated random numbers in the range of [0, 1] [3].

As it could be seen in Table 1, the accuracy of this model in the current study is 51.66%. Also, precision, recall and level of classification error of this method, have been summarized in table 2.

Table 1: The model accuracy in PSO algorithm

Accuracy : 66.51 %			
	True Positive	True Negative	Class Prediction
Pred. positive	57	35	61.96 %
Pred. negative	35	82	70.09 %
Class recall	61.96 %	70.09 %	

Table 2: Summary of software output in the training algorithm of PSO

Category	***Precision***		***Recall***
Positive	61. 96%		61. 96%
Negative	70. 09%		70. 09%
Accuracy	66. 51%	Classification Error	33. 49%

Assessment of PSO classification algorithm in the testing phase

As you can see in Table 3, the model accuracy is 33.60%. Precision, recall and level of classification error of the method, have been summarized in table 4.

Table 3: The model accuracy in PSO algorithm, in the testing phase.

Accuracy : 60.33 % ± 8.08%			
	True Positive	True Negative	Class Prediction
Pred. positive	32	23	58.18 %
Pred. negative	60	94	61.04 %
Class recall	34.78 %	80.34 %	

Table 4: Summery of PSO algorithm outputs, in the testing phase.

Category	***Precision***		***Recall***
Positive	58. 18%		34. 78%
Negative	61. 04%		80. 34%
Accuracy	60. 33%	Classification Error	39. 67

3.2 Bagging Algorithm

Bagging is a method for improving results of machine learning classification algorithms. This method was formulated by Leo Breiman and its name was deduced from the phrase “bootstrap aggregating” [4].

In case of classification into two possible classes, a classification algorithm creates a classifier H: D → (-1, 1) on the base of a training set of example descriptions (in our case played by a document collection) D. The bagging method creates a sequence of classifiers H_m, m=1,… ,M in respect to modifications of the training set. These classifiers are combined into a compound classifier. The prediction of the compound classifier is given as a weighted combination of individual classifier predictions

$$H(d_i) = \text{sign}\left(\sum_{m=1}^{M} \alpha_m H_m(d_i)\right)$$

The meaning of the above given formula can be interpreted as a voting procedure. An example d_i is classified to the class for which the majority of particular classifiers vote. Parameters α_m, m=1,…,M are determined in such way that more precise classifiers have stronger influence on the final prediction than less precise classifiers. The precision of base classifiers H_m can be only a little bit higher than the precision of a random classification [5].

As Table 5 shows, the model accuracy is 21.84%. Precision, recall and level of classification error of the method, have been summarized in table 6.

Table 5: The model accuracy in bagging algorithm in the training phase.

Accuracy : 84.21 %			
	True Positive	True Negative	Class Prediction
Pred. positive	70	11	86.42 %
Pred. negative	22	106	82.81 %
Class recall	76.09 %	90.60 %	

Table 6: Summery of the software output in the hybrid algorithm training.

Category	***Precision***		***Recall***
Positive	76. 09 %		86. 42 %
Negative	90. 60 %		82. 81 %
Accuracy	84. 21 %	Classification Error	15. 79 %

Assessment of bagging classification algorithms in the testing phase

The model accuracy in Table 3 is 33.60%. Also, precision, recall and level of classification error of the method, have been presented in table 8:

Table 7: The model accuracy in bagging algorithm, in the testing phase.

Accuracy : 76.55 % ± 7.22%			
	True Positive	True Negative	Class Prediction
Pred. positive	70	27	72.16 %
Pred. negative	22	90	80.36 %
Class recall	76.09 %	76.92 %	

Table 8: Summery of bagging algorithm outputs, in the testing phase

Category	***Precision***		***Recall***
Positive	72. 16%		76. 61%
Negative	80. 36%		76. 92%
Accuracy	76. 55 %	Classification Error	23. 45

3.3 K-NNAlgorithm

K nearest neighbor (KNN) is a simple algorithm, which stores all cases and classify new cases based on similarity measure. KNN algorithms have been used since 1970 in many applications like statistical estimation and pattern recognition etc.KNN is a non parametric classification method which is broadly classified into two types 1) structure less NN techniques 2) structure based NN techniques. In structure less NN techniques whole data is classified into training and test sample data. From training point to sample point distance is evaluated, and the point with lowest distance is called nearest neighbor. Structure based NN techniques are based on structures of data like orthogonal structure tree (OST), ball tree, k-d tree, axis tree, nearest future line and central line [6]. Nearest neighbor classification is used mainly when all the attributes are continuous. Simple K nearest neighbor algorithm is as follows:

Steps 1) find the K training instances which are closest to unknown instance

Step2) pick the most commonly occurring classification for these K instances [7].

Table 9: The model accuracy in K-NN algorithm in the training phase

Accuracy : 81.82 %			
	True Positive	True Negative	Class Prediction
Pred. positive	73	19	79.35 %
Pred. negative	19	98	83.76 %
Class recall	79.35 %	83.76 %	

As you can see, the accuracy of the model is 81.82 %.

Table 10: Summery of the software output in the hybrid algorithm training.

Category	***Precision***		***Recall***
Positive	79. 35%		79. 35%
Negative	83. 76		83. 76%
Accuracy	81. 82	Classification Error	18. 18

Assessment of K-NN classification algorithms in the testing phase

As it can be seen in Table 11, the accuracy of the model is 72.74%.

Table 11: The model accuracy in K-NN algorithm, in the training phase.

Accuracy : 72.74 % ± 9.74%			
	True Positive	True Negative	Class Prediction
Pred. positive	62	27	69.66 %
Pred. negative	30	90	75.00 %
Class recall	67.39 %	76.92 %	

Table 12: Summery of K-NN algorithm outputs, in the testing phase

Category	***Precision***		***Recall***
Positive	69. 66%		67. 39%
Negative	75. 00%		76. 92
Accuracy	72. 74	Classification Error	27. 26

3.4 Combination of PSO and SVM

Support Vector Machines(SVMs) are one of the binary classifiers based on maximum margin strategy introduced by Vapnik [8]. Suppose we are given l training examples

(xi; yi); (1≤i ≤l), where xi is a feature vector in n dimensional feature space, yi is the class label {-1, +1}(positive or negative) of xi. SVMs find a hyperplane w.x + b = 0 which correctly separates training examples and has maximum margin which is the distance between two hyperplanes w. x + b≥1 and w. x + b≤1. The optimal hyperplane with maximum margin can be obtained by solving the following quadratic programming.

$$\min \quad \frac{1}{2}\|w\| + C\sum_{i}^{1}\xi_i$$
$$\text{s.t.} \quad y_i(w.x + b) \geq 1 - \xi_i$$
$$\xi_i \geq 0$$

where C is the constant and ξ_i is called a slack variable for a non-separable case [9].

In RapidMiner software, PSO is designed with an integrative approach, in order to be integrated with classification of support vector optimization. The goal is to improve the capabilities of each individual technique, and to compensate for disabilities of the other models such as Kernel. In the proposed hybrid algorithm, based on the nature of the data, radial type of kernel algorithm has been selected. Then, due to compatibility of this algorithm with support vector graph, it will be integrated with SVM algorithm.

The accuracy of this model in Table 13 is 52.82%. Also, precision, recall and level of classification error of this method, have been summarized in table 14.

Table 13: The model accuracy in hybrid algorithm in the training phase

Accuracy : 94.74 %			
	True Positive	True Negative	Class Prediction
Pred. positive	92	11	89.32 %
Pred. negative	0	106	100.00 %
Class recall	100.00 %	90.60 %	

Table 14: Summary of software output in hybrid algorithm in the training phase.

Category	***Precision***		***Recall***
Positive	99. 99%		99. 99%
Negative	90. 60%		90. 60 %
Accuracy	94. 74	Classification Error	5. 26

Assessment of hybrid classification algorithm

As you can see in Table 15, the model accuracy is 29.60%. Precision, recall and level of classification error of the method, have been summarized in table 16:

Table 15: The model accuracy of hybrid algorithm, in the testing phase.

Accuracy : 60.29 % ± 6.74%			
	True Positive	True Negative	Class Prediction
Pred. positive	46	37	55.42 %
Pred. negative	46	80	63.49 %
Class recall	50.00 %	68.38 %	

Table 16: Summery of hybrid algorithm outputs, in the testing phase.

Category	***Precision***		***Recall***
Positive	60. 23%		57. 61%
Negative	67. 77%		70. 09%
Accuracy	60. 49	Classification Error	39. 51

3.5 Summary

Using data mining in the analysis of massive data, provides a powerful tool to examine the relationship between the variables. Despite their Simplicity, K-NN and Bagging Algorithms, provide acceptable results regarding the feature reduction of collected data. Access to the appropriate data, proper preprocessing and applying suitable data mining methods, are the elements that could be useful in achieving good results. A comparison between the obtained accuracies of all three models, in training and testing phases are presented in table 17. As it could be seen, the highest accuracy belongs to the proposed hybrid algorithm (SVM + PSO), which is 94.74%.

Table 17: summary of algorithms accuracies.

Model	***Accuracy in Testing***	***Accuracy in Training***
K-NN	74.72%	82.81%
PSO	33.6%	51.66%
Bagging	55.76%	21.84%
SVM + PSO	49.6%	74.94%

4. Conclusions

The survey carried out by the researcher suggests that so far, in the majority of investigations, the standard data published by UC have been used, but given that the type of data impacts on the accuracy of the work, it would be more suitable to use the information and data collected in specific health centers and the results be referred to that specific center. Also, in most of conducted researches, the focus was on the old algorithms which in this case, the

newer and more efficient algorithm and even preferably, a hybrid algorithm could be used to achieve better results.
With reviewing the conducted researches and based on the findings of this study, the following results are achieved:

- Data mining, as a semi-automatic process of extracting knowledge from available information, includes the selection, processing, aggregation of information, knowledge extraction and representation and interpretation.
- Data mining process could be applied in commercial, scientific and security applications and the main techniques include classification, clustering and association rule extraction.
- New issues of in the field of data mining such as the diagnosis of heterogeneous, distributed data mining and maintaining the knowledge confidentiality and privacy, indicate the need for the development of scalable data mining techniques based on the available information, as a standard and distributed process.
- Different data mining algorithms have different accuracy and the type of effects on the accuracy of data.
- In order to have a more useful of data mining techniques in hospitals and medical centers in Iran, the best way is to create electronic medical and health records for patients, and to document patients information in hospitals.

References

[1] A. Khasseh, "data mining, text mining, and web mining: definitions and applications.," 2009.

[2] R. C. Eberhart and J. Kennedy, "A new optimizer using particle swarm theory," in *Proceedings of the sixth international symposium on micro machine and human science*, 1995, pp. 39-43.

[3] M. Pant, R. Thangaraj, and A. Abraham, "Particle swarm optimization: performance tuning and empirical analysis," in *Foundations of Computational Intelligence Volume 3*, ed: Springer, 2009, pp. 101-128.

[4] L. Breiman, "Bagging predictors," *Technical Report 421, Department of Statistics, University of California at Berkeley*, 1994.

[5] K. Machová, F. Barcak, and P. Bednár, "A bagging method using decision trees in the role of base classifiers," *Acta Polytechnica Hungarica*, vol. 3, pp. 121-132, 2006.

[6] N. Bhatia, "Survey of nearest neighbor techniques," *arXiv preprint arXiv:1007.0085*, 2010.

[7] B. Deekshatulu and P. Chandra, "Classification of heart disease using K-nearest neighbor and genetic algorithm," *Procedia Technology*, vol. 10, pp. 85-94, 2013.

[8] V. Vapnik, *The nature of statistical learning theory*: Springer Science & Business Media, 2013.

[9] H. Yamada and Y. Matsumoto, "Statistical dependency analysis with support vector machines," in *Proceedings of IWPT*, 2003, pp. 195-206.