

## Myocardial Infarction Prediction and Estimating the Importance of its Risk Factors Using Prediction Models

### Abstract

**Background:** According to World Health Organization (WHO), cardiovascular diseases (CVDs) are the leading cause of death globally. Although significant progress has been made in the diagnosis of CVDs, more investigation can be helpful. Therefore, this study aimed to predict the risk of myocardial infarction (MI) using data mining algorithms. **Methods:** The applied data were related to the admitted patients in Rajaei specialized cardiovascular hospital located in Tehran. At first, a literature review and interview with a cardiologist were conducted to understand MI. Then, data preparation (cleaning and normalizing the data) was performed. After all, different classification algorithms were applied in IBM SPSS Modeler (14.2) software on the prepared data; and, power of the applied algorithms and the importance of the risk factors in predicting the probability of getting involved with MI was calculated in the mentioned software. **Results:** This study was able to predict MI % 75.28 and 77.77% in terms of accuracy and sensitivity, respectively. The results also revealed that cigarette consumption, addiction, blood pressure, and cholesterol were the most important risk factors in predicting the probability of getting involved with MI, respectively. **Conclusions:** Predicting studies aim to support rather than replace clinical judgment. Our prediction models are not sufficiently accurate to supplant decision-making by physicians but have considerable tips about MI risk factors.

**Keywords:** Cardiovascular disease, data mining, myocardial infarction, risk factor

### Introduction

As the major cause of morbidity and mortality,<sup>[1]</sup> cardiovascular diseases (CVDs) are considered an issue in many countries around the world.<sup>[2]</sup> “CVDs include a wide range of conditions that affect the heart, blood vessels, and the way the heart pumps the blood and flows around the body”,<sup>[3]</sup> including coronary heart diseases, cerebrovascular diseases, peripheral arterial diseases, rheumatic heart diseases, congenital heart diseases, and deep vein thrombosis and pulmonary embolism.<sup>[1]</sup> In general, risk factors associated with CVDs are classified into two categories: modifiable and non-modifiable factors. Non-modifiable factors refer to factors that cannot be changed or controlled, such as age, gender, and family history of CVDs. Modifiable factors refer to those factors that can be changed or controlled, such as abnormal levels of blood lipids, hypertension, diabetes, cigarette consumption, obesity, and excessive weight, little physical

activity, excessive consumption of alcohol, stress, and unhealthy nutrition.<sup>[4-6]</sup> By being aware of the significance of these risk factors and identifying individuals susceptible to CVDs, the incidence of such diseases can be to a certain extent prevented. Considering the importance of this issue, significant achievements have been made so far in the area of diagnosis and treatment of CVDs; however, because of the heavy burden that these diseases place on society in economic and healthcare terms, continuous research is needed to be conducted in this respect. Multiple studies have demonstrated that computational predictive approaches play pivotal roles in the identification and prediction of the probability of suffering from CVDs. Moses *et al.* (2018)<sup>[7]</sup> managed to predict heart attack using a Linear Regression algorithm with an accuracy of 81%. Using the support vector machines (SVM) algorithm, Mukherjee *et al.* (2017)<sup>[8]</sup> managed to predict heart diseases with an accuracy of approximately 85%. Seenivasagam and Chitra (2016)<sup>[9]</sup> managed

Fatemeh Rahimi<sup>1,2</sup>,  
Mahdi Nasiri<sup>1</sup>,  
Reza Safdari<sup>3</sup>,  
Goli Arji<sup>3</sup>,  
Zahra Hashemi<sup>4</sup>,  
Roxana Sharifian<sup>5</sup>

<sup>1</sup>Department of Health, Information Management, School of Management and Medical Information Sciences, Shiraz University of Medical Sciences, Shiraz, <sup>2</sup>Department of Health Information Technology and Management, School of Allied Medical Sciences, Shahid Beheshti University of Medical Sciences, <sup>3</sup>Department of Health Information Management, Tehran University of Medical Sciences, Tehran, <sup>4</sup>Department of Pediatric Dentistry, School of Dentistry, Yasuj University of Medical Sciences, Yasuj, <sup>5</sup>Department of Health Information Management, School of Management and Medical Information Sciences, Health Human Resources Center, Shiraz University of Medical Sciences, Shiraz, Iran

**Address for correspondence:**  
Dr. Roxana Sharifian,  
School of Management  
and Medical Information  
Sciences, Qasroldasht  
29<sup>th</sup> Alley, Qasroldasht Ave,  
Shiraz - 7133654361, Iran.  
E-mail: Sharifianr@sums.ac.ir

### Access this article online

**Website:**  
[www.ijpvmjournal.net/www.ijpvm.net](http://www.ijpvmjournal.net/www.ijpvm.net)

**DOI:**  
10.4103/ijpvm.IJPVM\_504\_20

### Quick Response Code:



**How to cite this article:** Rahimi F, Nasiri M, Safdari R, Arji G, Hashemi Z, Sharifian R. Myocardial Infarction prediction and estimating the importance of its risk factors using prediction models. *Int J Prev Med* 2022;13:158.

This is an open access journal, and articles are distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 License, which allows others to remix, tweak, and build upon the work non-commercially, as long as appropriate credit is given and the new creations are licensed under the identical terms.

For reprints contact: WKHLRPMedknow\_reprints@wolterskluwer.com

to predict MI using the particle swarm optimized neural network algorithm with an accuracy of 89.61. With an accuracy of 100%, Dangare and Apte (2012)<sup>[10]</sup> managed to predict CVDs using the neural network model. Using the Naïve Bayes algorithm, Rajkumar and Reena (2010)<sup>[11]</sup> managed to diagnose heart diseases with an accuracy of 52.33%. Using K-means and MAFIA algorithms, Patil and Kumaraswamy (2009)<sup>[12]</sup> managed to develop patterns to predict the occurrence of heart attacks. Using a combination of neural networks and genetic algorithms, Amin (2013) managed to predict heart diseases based on the risk factors of this disease.<sup>[13]</sup> In general, a large volume of data is gathered in the area of healthcare. However, this size of data has not been effectively mined to detect hidden data and take more optimal decisions.<sup>[14,15]</sup> In line with this, data mining techniques can be tapped into as an efficient method to extract hidden patterns in a large volume of medical information.<sup>[14]</sup> Considering the significance of CVDs and the ability of data mining to extract hidden patterns from among a large volume of data in the area of healthcare, the present study aimed at predicting the development of myocardial infarction (MI) using numerous supervised data mining techniques and based on the risk factors of this disease.

## Methods

### Ethics

The patients' information was de-identified and confidentiality and privacy of the patients' information were maintained in all steps of the study. Also, this study was ethically approved by Shiraz University of Medical Sciences (Approval No. IR.SUMS.TEC.1396.S218).

### Study design and dataset

The present study is an exploratory data mining exerted on retrospective medical data. The data had been collected via a checklist from the medical files of patients visiting Rajaei Heart Hospital in 2013. The dataset included 350 records (200 patients and 150 People without MI) and 40 different variables. By people without MI in the study, it is meant individuals with whom the existence of MI had been rejected considering the information mentioned in their medical records. By patients, it is meant individuals with whom the presence of MI had been confirmed considering the information mentioned in their medical records.

### Data analysis

After data preparation (cleaning and normalizing the data) using the Microsoft Excel software, there were 14 variables in our dataset. Different classification algorithms were applied in IBM SPSS Modeler (Clementine 14.2) software on the prepared data; and, power of the applied algorithms and the importance of the risk factors in predicting the probability of getting involved with MI was calculated in the mentioned software. To evaluate the performance of the

models, we focused on the "Accuracy" and "Sensitivity" measures. The stages of the procedure are shown in Figure 1.

### Data preparation and modelling

The type and quality of the dataset used in the process of data mining affect the performance of data mining techniques.<sup>[16]</sup> Therefore, variables with high missing values (more than 50 percent of the data of each variable) were completely eliminated at the data preparation stage. Then, the data records of each patient were separately examined. The data belonging to some patients had plenty of missing values. For the reliability of the model, there were no other options but eliminating these records. Ultimately, by minimizing the missing values in each dataset, the mean values existing for each variable were calculated to use instead of the empty fields belonging to that variable. In the end, there were 274 records of normalized data, including 129 patients and 145 people without MI, and 14 variables in our dataset. Of these, 13 variables are the predictor variables [Table 1].

In this study, the effort was made to conduct the data cleaning, preparation, and modeling stages on the basis of the existing documents in the field of medicine. Therefore, variables with consecutive values were discretized based on the risk factors associated with CVDs classified by the National Institutes of Health (NIH) publications and the WHO.<sup>[17-19]</sup> Table 1 shows the method of discretization of these variables.

To develop the models, the four algorithms of the neural network, k-NN, Bayesian network, and decision tree were used. Then, the models resulting from the algorithms were combined together and the integrative model was developed. The four mentioned algorithms are supervised algorithms that are suitable for predictions in the field of health.<sup>[20,21]</sup>

- Neural network: It is a non-linear technique that works like a black-Box (It is difficult to see what happens in the prediction process)
- K-NN: It is a non-parametric classification method that predicts the class of each sample based on the class of its K nearest neighbor instances.
- Bayesian network: It is a probabilistic model that uses Bayesian inference computations to classify each sample.
- Decision tree: The decision tree algorithm works in a tree-like structure and classifies the instances by arranging them from the root to the leaves.

Before developing the models, a feature selection has been conducted. Then the data were clustered using the K-means algorithm. The results as to which data were placed in which cluster were recorded as a new field in the applied dataset and used in the modeling. Then, using the SPSS Clementine software, the data were classified into two partitions: training data (70%) and testing data (30%).

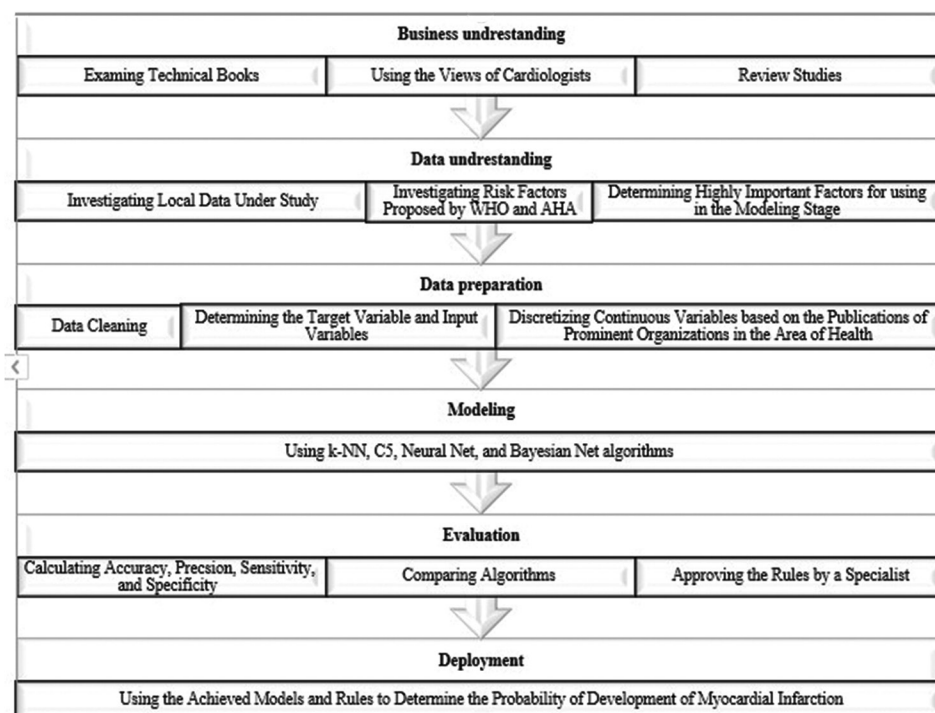


Figure 1: The stages of the procedure

After the models in mind were developed, the accuracy, precision, sensitivity, and specificity of the models were achieved using 10-fold cross validation. Then, the results obtained from each method were interpreted and the most important predictors were determined from the point of view of the applied models. To evaluate the performance of the models, we focused on the “Accuracy” and “Sensitivity” measures.

### Results

The data employed in the modeling consist of 129 (47%) records for the patients and 145 records for the People without MI (53%). The majority of data were for males (58%). The mean age (standard deviation) was 57 (10.66) and 60<sup>[11,12]</sup> years for the patients and the people without MI, respectively. The basic characteristics of the study participants have been provided in Table 2.

From among the algorithms used for modeling, k-NN, Bayesian network, decision tree (C5), and neural network algorithms managed to predict MI with an accuracy of 75.28%, 69.66%, 64.04%, and 51.69%, respectively [Table 3]. The rules obtained from the C5 algorithm have been provided in Table 4.

The accuracy of the model resulting from mixing the above said algorithms was equal to 73.03%, which was less than the value achieved for the k-NN algorithm.

Based on the Bayesian network algorithm, the following variables were the most important predictors of MI, respectively: cholesterol (0.14), triglyceride (0.12),

hypertension (0.10), cigarette consumption (0.09), addiction (0.08), DLP (0.08), diabetes (0.07), age (0.07), family history of CVDs (0.07), and HDL (0.06) [Figure 2a]. The C5 algorithm too showed that the following variables were the most important predictors of cardiac infarction, respectively: history of cigarette consumption (0.23), hypertension (0.22), addiction (0.19), age (0.18), and triglyceride (0.18) [Figure 2b]. The results obtained from the k-NN algorithm showed that the following variables were the most important predictors associated with cardiac infarction, respectively: history of cigarette consumption (0.09), hypertension (0.09), BMI (0.09), LDL (0.08), cholesterol (0.08), DLP (0.08), HDL (0.08), age (0.08), and diabetes (0.08) [Figure 2c]. The neural network algorithm revealed that the following variables were the most important predictors of cardiac infarction, respectively: history of cigarette consumption (0.33), LDL (0.17), addiction (0.14), cholesterol (0.09), HDL (0.08), triglyceride (0.06), DLP (0.03), age (0.03), BMI (0.02), and hypertension (0.02) [Figure 2d].

Table 5 shows that by applying the mean on the results of the four algorithms used, the variable “cigarette consumption” was the most important predictor of MI.

### Discussion

Considering the significance of CVDs and the capability of data mining to extract hidden patterns among a large volume of data in the field of healthcare, the present study aimed at predicting MI by using various supervised data mining techniques based on the risk factors of this disease.

**Table 1: Discretization of the applied variables based on medical reference**

ID	Risk factor	Discretization	Unit
1	Age	<45 Age 45<=Age <55 Age>=55	
2	Gender	Female Male	
3	Family history	Yes No	
4	Body Mass Index	Underweight: BMI <18.5 Normal: 18.5<=BMI <25 Overweight: 25 <BMI <30 Obese: BMI >=30	
5	Cigarette consumption	Yes No	
6	Addiction	Yes No	
7	Triglyceride value	Normal: TG <150 Borderline-High: 150 <=TG <200 High: 200 <=TG <500 Very High: TG >=500	(mg/dL)
8	Low-density lipoprotein	Optimal: LDL <100 Near optimal: 100 <=LDL <130 Borderline-High: 130 <=LDL <160 High: 160 <=LDL <190 Very High: LDL >190	(mg/dL)
9	High-density lipoprotein	Low: HDL <40 Normal: 40 <=HDL <60 Optimal: HDL >=60	(mg/dL)
10	Dyslipidaemia	Yes No	
11	Diabetes	Yes No	
12	Total cholesterol	Desirable: TC <200 Borderline-High: 200 <=TC <240 High: TC >=240	(mg/dL)
13	Blood pressure	Normal or prehypertension: SBP <140 hypertension: SBP >=140 Normal or prehypertension: DBP <90 Hypertension: DBP >=90	mmHg mmHg

**Table 2: Basic characteristics of the study participants**

	People with MI	People without MI
Age		
Mean	57	60
Std. Deviation	10.66	11.12
Min	33	30
Max	91	85
Gender		
Male	58 (%21)	101 (%37)
Female	71 (%26)	44 (%16)
Frequency	129 (%47)	145 (%53)

The results of the present study [Table 5] showed that from among the risk factors associated with CVDs, the following

variables were the most important predictors of MI, respectively, such as cigarette consumption, addiction, hypertension, cholesterol level, DLP, LDL, triglyceride level, HDL, diabetes, BMI, and family history of CVDs. The results of the study have also shown that the variable “gender” was not effective in predicting MI in terms of each of the algorithms used. The present study also showed that from among the algorithms used in this study, the algorithm k-NN with an accuracy of 75.28% was the most powerful algorithm for predicting MI. In other words, this algorithm is in general able to have a proper prediction of the occurrence or non-occurrence of the disease in 75.28% of the cases. The algorithm also had a better performance compared to other algorithms in terms of each of the indexes of sensitivity, specificity, and precision. The power of this algorithm in the proper identification of individuals that will suffer from MI (sensitivity or recall) is higher compared to the proper identification of individuals that will not suffer from MI in the future (specificity) (77.77% and 72.72%, respectively). Proper identification of individuals that will probably develop MI compared to proper identification of individuals that will probably not develop MI is of much more importance, as such vulnerable people can be helped to reduce the probability of developing the disease by timely identification of them and by training them. However, if such people are not properly identified, they will probably not use preventive measures or modify their high-risk behavior. However, if a person is likely to be healthy in the future and we wrongly consider him/her as a person vulnerable to MI, there will not be a lot of risks compared to the previous condition and he/she can use training programs and modifiable behavior like susceptible people. Therefore, as the most successful algorithm in this study, the k-NN algorithm has been able to properly identify people that will develop MI in the future in 77.77% of cases. Considering the importance of predicting CVDs, numerous other studies have so far been conducted in this area. One such study compared the performance of the decision tree and neural network in predicting the development of MI. The initial dataset used in the above study was identical to the initial data used in the present study.<sup>[22]</sup> The difference was, however, the method of data cleaning and preparation in the present study. In other words, in the present study, after the data were cleaned, only 274 records from the initial 350 records were used for the modeling. In addition, in the present study, the variables were discretized based on the classification of risk factors of CVDs as approved by prestigious organizations in the area of health and treatment [Table 1]. However, in the above said study, variables were discretized automatically by the SPSS Clementine software. In terms of the factors effective in the development of the disease, all the three variables introduced in the above said study, including hypertension, high levels of blood lipid, and cigarette consumption<sup>[14]</sup> were compatible with the variables introduced in the

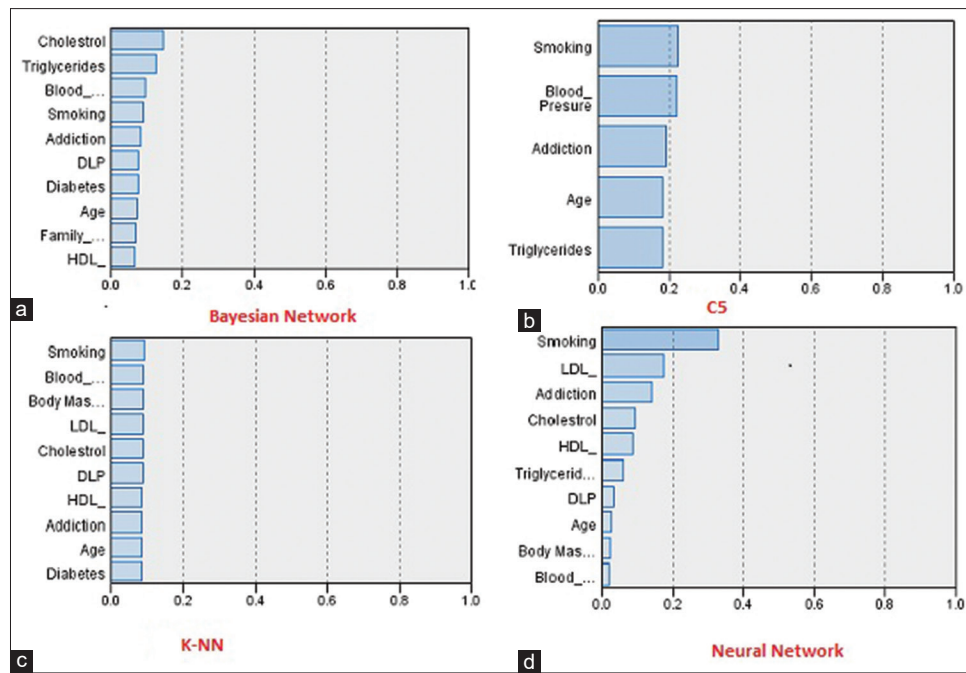


Figure 2: Importance of the predictor variables

**Table 3: Comparing the algorithms used in terms of accuracy, precision, sensitivity and specificity**

	Accuracy=(TP+TN)/ (TP+TN+FP+FN)	Precision=TP/(TP+FP)	Sensitivity (Recall)=TP/(TP+FN)	Specificity=TN/ (TN+FP)
K-NN	%75.28	%74.46	%77.77	72.72%
Bayesian Network	%69.66	%68	%75.55	%63.63
C5	%64.04	%64.44	%64.44	%63.63
Neural Network	%51.69	%53.84	%31.11	%72.72

(TP: True positive<sup>a</sup>, TN: True negative<sup>b</sup>, FP: False positive<sup>c</sup>, FN: False negative<sup>d</sup>). <sup>a</sup>Patients who were properly placed in the class of patients by the algorithm. <sup>b</sup>Healthy individuals who were properly placed in the class of healthy individuals by the algorithm. <sup>c</sup>Healthy individuals who were wrongly placed in the class of patients by the algorithm. <sup>d</sup>Patients who were wrongly placed in the class of healthy individuals by the algorithm

**Table 4: Rules achieved based on C5 algorithm and the percentage of accuracy of each rule**

Rule No	Rule Description	Rule accuracy
1	If an individual is not a smoker, he/she will not develop cardiac infarction.	68.6%
2	If an individual has a history of hypertension, he/she will not develop cardiac infarction.	68.6%
3	If an individual does not have a history of hypertension but is a smoker, he/she will develop cardiac infarction.	68.6%
4	If an individual is not addicted, he/she will not develop cardiac infarction.	64.3%
5	If an individual has a high level of triglyceride, he/she will develop cardiac infarction.	64.3%
6	If an individual is younger than 55 and has a history of addiction and hypertension, he/she will develop cardiac infarction.	64.3%
7	If an individual is a smoker, he/she will develop cardiac infarction.	60%
8	If an individual is addicted but has no hypertension, he/she will develop cardiac infarction.	63.8%

present study. Meanwhile, in the above said study, such variables as FBS, urea, creatinine, and blood type were used in addition to the risk factors associated with CVDs. [22] However, the present study specifically focused on determining the importance of risk factors associated with CVDs in predicting the development of MI. The results of the above said study indicated that the decision tree and neural network had the highest ability, respectively, to predict the development of MI. [22] Considering the difference in the way the present study was conducted compared to the above said study (in terms of the type of variables used and the method of cleaning and processing the initial data), the present study reinvestigated these two algorithms in addition to new algorithms. The present study not only supported the results of the above said study but also introduced the k-NN and Bayesian Network algorithms as the algorithms with higher ability than the two algorithms introduced in the above said study, namely, the decision tree and the neural network algorithms. Other advantages of the present study include the application of algorithms on local data. The reason for the importance of using local data is that the trend of progress and effect of

**Table 5: The mean importance of various risk factors in predicting the probability of development of MI**

ID	Risk factor	Predictor importance <sup>a</sup> using different algorithms				Average importance of predictors	Priority
		Bayesian Network	C5	K-NN	Neural Network		
1	Age	0.07	0.18	0.08	0.03	0.09	4
2	Gender	0	0	0	0	0.00	11
3	Family history	0.07	0	0	0	0.02	10
4	Body Mass Index	0	0	0.09	0.02	0.03	9
5	Cigarette consumption	0.09	0.23	0.09	0.33	0.19	1
6	Addiction	0.08	0.19	0.08	0.14	0.12	2
7	DLP	0.08	0	0.08	0.03	0.07	5
8	Triglyceride value	0.12	0.018	0	0.06	0.05	7
9	Low-density lipoprotein	0	0	0.08	0.17	0.06	6
10	High-density lipoprotein	0	0	0.08	0.08	0.04	8
11	Diabetes	0.07	0	0.08	0	0.04	8
12	Total cholesterol	0.14	0	0.08	0.14	0.09	4
13	Blood pressure	0.10	0.22	0.09	0.02	0.11	3

<sup>a</sup>Risk factor importance

the disease can be different among individuals living in a certain area compared to those living in other areas. Many studies conducted so far on CVDs have used general databases available on the World Wide Web. For instance, a study conducted on data of the UCI database<sup>[23]</sup> to predict CVDs suggested that Naïve Bayes had a better performance (with an accuracy of 85.03%) compared to the decision tree algorithm (with an accuracy of 84.01%).<sup>[24]</sup> There was also another study conducted on the data of the UCI database to predict CVDs. The results of the study showed that Naïve Bayes with an accuracy of 86.53% and the neural network with an accuracy of less than 1% compared to Naïve Bayes were appropriate algorithms for predicting CVDs. However, the decision tree technique was the most appropriate method for predicting individuals who have no CVDs (89%).<sup>[25]</sup> The results of this study emphasized that an algorithm might be appropriate for a particular purpose while it is less efficient than other algorithms for other purposes. Another study that was conducted to assess risk factors associated with CVDs was carried out using the decision tree (C4.5 algorithm). It revealed that important risk factors for MI include age, cigarette consumption, and hypertension, a finding that was compatible with the results of the present study. In this study, the highest accuracy for MI was equal to 66%.<sup>[26]</sup> In addition, the results of a study showed that a neural network with an accuracy of 100% is more efficient in predicting CVDs than the decision tree with an accuracy of 99.62%.<sup>[15]</sup> A review study conducted to investigate studies in the area of diagnosis of CVDs argued that the decision tree and SVM had, in general, more optimal results in predicting CVDs compared to other methods of classification. This study has suggested not only the previous history of the disease, cholesterol, and age, which are identical to the risk factors of the present study but also other factors including gender, chest pain, FBS, electrocardiograph results, and maximum heart rate as the effective and reliable factors in predicting CVDs.<sup>[27]</sup> Based

on the results of two recent studies, clinical data is not lonely enough for estimating the risk of Heart diseases (MI<sup>[28]</sup> and Coronary heart diseases<sup>[29]</sup>). However, Screening is of paramount importance when it comes to the identification of cardiac patients at the early stages of detection by investigating the existence or non-existence of risk factors associated with the development of these diseases in individuals under study. If patients vulnerable to the diseases are identified at early stages, the economic burden and the rate of mortality induced by CVDs will decrease in society.<sup>[30]</sup> Thus, it is necessary to identify risk factors that have a higher share, compared to other factors, in the development of such diseases and to identify individuals vulnerable to CVDs. This can have a significant effect on improving screening programs, increasing cost-effectiveness, and improving the health condition of society. Therefore, identification of these factors is of paramount importance to policymakers in the area of health, as it can concentrate on and invest in controlling modifiable risk factors and make society more aware of modification of high-risk behavior relating to CVDs.

In general, knowledge of risk factors associated with non-communicable diseases should be considered as the top priority in programs relating to the prevention and control of non-communicable diseases.<sup>[31]</sup>

## Conclusions

Predicting studies aim to support rather than replace clinical judgment. Our prediction models are not sufficiently accurate to supplant decision making by physicians but have considerable tips about MI risk factors. The results of this study revealed that cigarette consumption, addiction, hypertension, cholesterol level, DLP, LDL, triglyceride level, HDL, diabetes, BMI, and family history of CVDs were the most important predictors of possible development of MI. The results of the study also showed that gender was not effective in predicting MI in terms

of any of the algorithms applied. The history of cigarette consumption was the most important risk factor associated with the development of MI. Since cigarette consumption is a modifiable behavior, policymakers in the area of health should focus on this issue, implement educational and modifying programs on high-risk behavior and thus help to reduce the risk of development of CVDs.

**Limitations of the Study:** Data quality control is important for increasing the quality of data including data completeness.<sup>[32]</sup> The limitation of the study includes a large number of fields with missing values in the initial information relating to the patients. To compensate for this limitation and prevent the effect of such information on the process of education, information relating to a large number of patients was eliminated from the study. In order to improve the accuracy of the model, the recommended model can be taught using information relating to more patients so that it can have better accuracy in identifying unseen data.

Also, despite the proven advantages of using data mining techniques in predicting diseases and their relevant predictors, some physicians do not prefer to use some of these techniques (such as neural networks), due to their “Blackbox” entity. To facilitate this limitation, we used the Decision Tree algorithm (C5) which is more visible. But, the accuracy of this algorithm was not acceptable.

### Ethics approval and consent to participate

This project was ethically approved by Shiraz University of Medical Sciences (Approval No. IR.SUMS.TEC.1396.S218). Patients’ information was de-identified and confidentiality and privacy of the patients’ information were maintained in all steps of the study.

### Acknowledgments

Hereby, the authors would like to thank the Shahid Rajaei specialized cardiovascular hospital for their cooperation in data gathering.

### Financial support and sponsorship

Nil.

### Conflicts of interest

There are no conflicts of interest.

**Received:** 29 Aug 20 **Accepted:** 17 Oct 21

**Published:** 26 Dec 22

### References

- World Health Organization. Fact sheet 317: cardiovascular disease. [Last updated on 2021 June]. Available from: <https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-cvds>. [Last accessed on 2022 Nov 04].
- Mensah GA, Roth GA, Fuster V. The Global burden of cardiovascular diseases and risk factors: 2020 and beyond. *J Am Coll Cardiol* 2019;74:2529–32.
- Soni J, Ansari U, Sharma D, Soni S. Predictive data mining for medical diagnosis: An overview of heart disease prediction. *Int J Comput Appl* 2011;17:43–8.
- Fact sheet: Risk factors for cardiovascular disease (CVD): HEART UK. The Cholesterol charity. Available from: <https://heartuk.org.uk>. [Last accessed on 2020 Aug 09].
- World Health Organization. Global atlas on cardiovascular disease prevention and control. Edited by Mendis *et al.* 2011.
- Cardiovascular disease risk factors. World Heart Federation; 2014. Available from: <http://www.world-heart-federation.org>. [Last accessed on 2020 Aug 09].
- Moses A, Sathishkumar R, Meghana M, Meghana Raju M, Madhumitha M. Forecasting myocardial infarction using machine learning algorithms. *Int J Pure Appl Math* 2018;118:859–63.
- Mukherjee S, Kapoor S, Banerjee P. Diagnosis and identification of risk factors for heart disease patients using generalized additive model and data mining techniques. *J Cardiovasc Dis Res* 2017;8:137–44.
- Seenivasagam V, Chitra R. Myocardial infarction detection using intelligent algorithms. *Neural Netw World* 2016;26:91.
- Dangare C, Apte S. A data mining approach for prediction of heart disease using neural networks. *Int J Comput Eng Technol*. 2012;3:30–40.
- Rajkumar A, Reena GS. Diagnosis of heart disease using datamining algorithm. *Glob J Comput Sci Technol* 2010;10:38–43.
- Patil SB, Kumaraswamy YS. Extraction of significant patterns from heart disease warehouses for heart attack prediction. *IJCSNS*. 2009;9:228–35.
- Amin SU, Agarwal K, Beg R. Genetic neural network based data mining in prediction of heart disease using risk factors. In: 2013 IEEE Conference on Information & Communication Technologies. IEEE; 2013. p. 1227–31.
- Subbalakshmi G, Ramesh K, Rao MC. Decision support in heart disease prediction system using naive bayes. *Indian J Comput Sci Eng* 2011;2:170–6.
- Dangare CS, Apte SS. Improved study of heart disease prediction system using data mining classification techniques. *Int J Comput Appl* 2012;47:44–8.
- Tomar D, Agarwal S. A survey on data mining approaches for healthcare. *Int J Bio-Sci Bio-Technol* 2013;5:241–66.
- Sandmaier M. Your guide to healthy heart. U.S. Department of Health and Human Services, National Institutes of Health, & National Heart, Lung, and Blood Institute 2005. Available from: <https://www.nhlbi.nih.gov>. [Last accessed on 2022 Dec 14].
- The healthy heart handbook for woman. U.S. Department of Health and Human Services, National Institutes of Health, National Heart, Lung, and Blood Institute 2007. Available from: <https://www.nhlbi.nih.gov>. [Last accessed on 2022 Dec 14].
- Prevention of Cardiovascular Disease, Guidelines for assessment and management of cardiovascular risk. World Health Organization; 2007.
- Berner ES, editor. *Clinical Decision Support Systems: Theory and Practice*. Switzerland: Springer International Publishing; 2016.
- Uddin S, Khan A, Hossain M, Moni MA. Comparing different supervised machine learning algorithms for disease prediction. *BMC Med Inform Decis Mak* 2019;19:281.
- Safdari R, Gharooni M, Nasiri M, Argi G. Comparing performance of decision tree and neural network in predicting myocardial infarction. *J Paramed Sci Rehabil* 2014;3:26–35.
- The UCI machine learning repository [Data base]. Available from: <http://www.ics.uci.edu>. [Last accessed on 2022 Dec 14].
- Venkatalakshmi B, Shivsankar MV. Heart disease diagnosis using predictive data mining. *Int J Innov Res Sci Eng Technol*

- 2014;3:1873–7.
25. Palaniappan S, Awang R. Intelligent heart disease prediction system using data mining techniques. In: 2008 IEEE/ACS international conference on computer systems and applications. IEEE; 2008. p. 108–15.
  26. Karaolis MA, Moutiris JA, Hadjipanayi D, Pattichis CS. Assessment of the risk factors of coronary heart events based on data mining with decision trees. *IEEE Trans Inf Technol Biomed* 2010;14:559–66.
  27. Ahmed A, Hannan SA. Data mining techniques to find out heart diseases: An overview. *Int J Innov Technol Explor Eng* 2012;1:18–23.
  28. Zarrabi M, Parsaei H, Boostani R, Zare A, Dorfeshan Z, Zarrabi K, *et al.* A system for accurately predicting the risk of myocardial infarction using PCG, ECG and clinical features. *Biomed Eng Appl Basis Commun*. 2017;29:1750023.
  29. Korley FK, Gatsonis C, Snyder BS, George RT, Abd T, Zimmerman SL, *et al.* Clinical risk factors alone are inadequate for predicting significant coronary artery disease. *J Cardiovasc Comput Tomogr* 2017;11:309–16.
  30. Eyre H, Kahn R, Robertson RM, Committee AC, Members AC, Clark NG, *et al.* Preventing cancer, cardiovascular disease, and diabetes: A common agenda for the American Cancer Society, the American Diabetes Association, and the American Heart Association. *Circulation* 2004;109:3244–55.
  31. Karimi S, Javadi M, Jafarzadeh F. Economic burden and costs of chronic diseases in Iran and the world. *Health Inf Manag* 2012;8:984–96.
  32. Sharifian R, SedaghatNia MH, Nematollahi M, Zare N, Barzegari S. Estimation of completeness of cancer registration for patients referred to shiraz selected centers through a two source capture re-capture method, 2009 data. *Asian Pacific J Cancer Prev* 2015;16:5549–56.