Application of Data Mining Techniques in Predicting Coronary Heart Disease: A Systematic Review

Saeed Saeedbakhsh, Mohammad Sattari, Maryam Mohammadi[1], Jamshid Najafian[2]

Health Information Technology Research Center, [1]Department of Management and Health Information Technology, School of Management and Medical Information Sciences, [2] Isfahan Cardiovascular Research Center, Cardiovascular Research Institute, Isfahan University of Medical Sciences, Isfahan, Iran

Abstract

Aim: The early detection of cardiovascular diseases by noninvasive and low-cost methods such as data mining techniques has been considered by many researchers. This study intends to review the studies performed on the prognosis of coronary heart disease using data mining techniques. **Materials and Methods:** The published studies in English between 2001 and 2021 that the use classification methods to predict coronary heart disease were considered. Databases such as ScienceDirect, Web of Science, and ScoPURs were considered as searchable databases. After searching, 348 articles were retrieved. After removing duplicates and evaluating the articles, finally, 20 articles were used. **Results:** The three data mining techniques support vector machine (SVM), neural network, and naive Bayes which were the most used among the studies. In the most studies, risk factors age, blood pressure, gender, diabetes, and chest pain were used. The accuracy was the most-used measure. The Alizadeh Sani dataset was the most used among the studies. **Conclusion:** Techniques such as SVM and neural network have performed better than other techniques. The output of these techniques can be used as a decision support system so that clinicians can enter various risk factors such as age, blood pressure, gender, diabetes, and chest pain and then view system output.

Keywords: Coronary heart disease, data mining, diagnosis, predicting

Introduction

Cardiovascular disease is a group of diseases that indicates in the heart or arteries.[1,2] Heart disease includes a variety of conditions, including congenital diseases, coronary heart disease, and rheumatoid arthritis.[3] The World Health Organization has identified coronary heart disease as the most common type of cardiovascular disease.[4] Coronary heart disease is caused by the accumulation of platelets in the coronary arteries.[5] The definitive diagnosis of this disease is when the stenosis of at least, one of the coronary arteries is >50%.[6] According to the World Health Organization, coronary heart disease has remained in the top 10 causes of death in the world for the past 15 years.[7] According to statistics, >17 million of deaths worldwide are due to the disease.[8] Various methods such as coronary angiography and cardiac catheterization are known as standard methods for assessing the presence of coronary heart disease, but these methods are invasive and expensive.[5] Therefore, the use of noninvasive and low-cost methods such as data mining techniques for early detection of this disease has been considered by many researchers.[1,9] Medical data contain valuable information that can be a good source of knowledge. The volume of this data is increasing day by day, and physicians can obtain usable information about diseases from this volume of data. Data mining is one of the evolving sciences that have proven its place in all fields in recent years; in such a way that its growth is increasing compared to other superior sciences. The collaboration of computer and medical experts offers a new solution in analyzing medical data and obtaining useful and practical models, which is the same as medical data mining. Mazaheri developed a data-driven model for predicting heart disease. The best results were obtained from tree classification and regression algorithms. Values of 2.89% for the overall average accuracy of the model, 2.59% for accuracy, 6.81% for

Access this article online

Quick Response Code:

Website: www.ijehe.org

DOI: 10.4103/ijehe.ijehe_1_21

Address for correspondence: Dr. Mohammad Sattari, Health Information Technology Research Center, Isfahan University of Medical Sciences, Isfahan, Iran. E-mail: msattarimng.mui@gmail.com

How to cite this article: Saeedbakhsh S, Sattari M, Mohammadi M, Najafian J. Application of data mining techniques in predicting coronary heart disease: A systematic review. Int J Env Health Eng 2021;XX:XX-XX.

Received: 10-03-2021, **Accepted:** 15-05-2021, **Published:** 30-09-2021

specificity, and 2.77% for sensitivity by the tree classification and regression algorithm indicate that the generated tree can provide comprehensive rules for predicting the status of future patients.[10] Dutta *et al.* presented a torsional neural network model with the aim of predicting the indication of coronary heart disease. The model was compared with support vector machine (SVM) and random forest (RF). The accuracy of the model was 79.5% and had a higher accuracy than the other methods. Furthermore, the torsional neural network model had better sensitivity, specificity, and the area under the curve.[11] Many studies have used various data mining techniques to predict coronary heart disease under different conditions, and each of these techniques has used one or two data mining techniques independently or in comparison with each other. Taking into account the importance of predicting coronary heart disease, researchers in this study assess all studies have used data mining techniques to predict coronary heart disease, search, study, and analysis, and finally put a scientific framework to do future research in this field.

Materials and Methods

This study considers studies published in English between 2001 and 2021 that use classification methods to predict coronary heart disease. Databases such as ScienceDirect, Web of Science, and Scopurs are considered [Table 1]. Table 1 shows the number of studies published between 2001 and 2021 in the listed databases.

Search strategy

In terms of keywords and the list of synonyms, a combination of keywords and synonyms is searched based on Boolean logic (OR). The results are combined and searched using Boolean logic (AND). These searches utilized keywords including (Coronary artery disease) AND (Diagnosis OR Prediction) AND (Data mining).

Inclusion criteria

Studies in English between 2001 and 2021 using classification techniques to predict coronary heart disease are included.

Exclusion criteria

Studies that are not in English are not considered. In addition, studies that have used text mining techniques to predict coronary heart disease are excluded.

Selection of studies

Duplicate records are removed first, and then, the title and abstract of the remaining studies are considered based on input and output criteria. Unrelated studies were excluded in terms of title and abstract. Then, the full text of the articles was considered. Then, among the remaining articles, the articles whose full text was not related were deleted. Finally, the available articles were considered.

Table 1: The number of records in each database

Database	The number of records
PubMed	87
Web of Science	39
ScienceDirect	69
Scopus	153
Total	348

Data extraction and classification

Information on authors' names, year of publication, data set, risk factors and techniques used, and evaluation criteria for each technique were extracted from the studies. The factors extracted by the researchers were then analyzed. 348 studies were retrieved after the initial search. Finally, 20 cases had the necessary criteria to enter this study. Table 1 shows the number of initial studies retrieved from each database. Thirty-four duplicate studies were excluded. The available studies (314) were reviewed and evaluated based on the title and abstract, and 278 studies that did not meet the inclusion criteria (their title and purpose did not meet the criteria of this study) were left out. After assessing the full text of the remaining studies (36 cases), 16 studies did not qualify for this study and were deleted, and finally, 20 studies were selected and used [Figure 1].

Results

Kolukisa *et al.* have used a combination of various machine learning algorithms to predict coronary heart disease. In this study, techniques such as k-Nearest Neighbor (KNN), logistic regression, and SVM were tested in two datasets. The results showed that the SVM algorithm with 83.43% and 88.38% accuracy in both datasets performed better than other algorithms.[12] Dutta *et al*. presented a Convolutional Neural Network (CNN) model with the aim of predicting the occurrence of coronary heart disease. In this study, other machine learning methods, such as SVM and RF, were compared with the proposed CNN model. The accuracy of the proposed model was 79.5% and had a higher accuracy than the RF and SVM. Furthermore, the CNN model had better sensitivity and specificity and the area under the curve.[11] Ayatollahi *et al*. used Artificial Neural Network (ANN) and SVM to predict coronary heart disease. The results showed that the SVM algorithm with 92.32% sensitivity and 74.42% specificity performed better

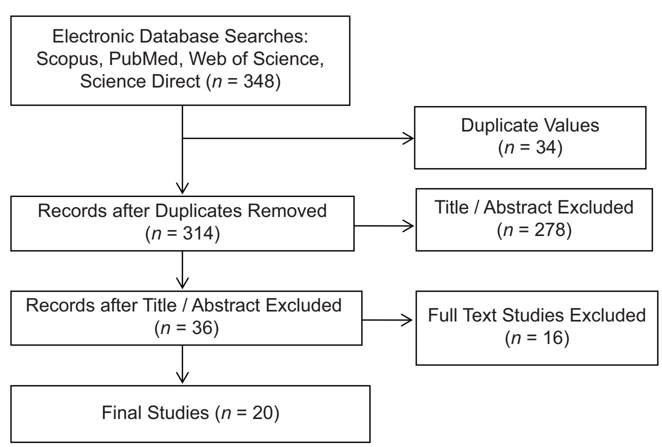

Figure 1: Strategy for extraction of the studies

than the ANN. Since the area under the ROC curve in the SVM algorithm was more than this area in the ANN model, it can be concluded that the SVM model is more accurate than the ANN model.[13] Velusamy and Ramasamy proposed a new heterogeneous ensemble method combining three base classifiers, KNN, RF, and SVM for effective diagnosis of coronary heart disease. The results of the base classifiers were combined using an ensemble voting technique based on average-voting (AVEn), majority-voting (MVEn), and weighted-average voting (WAVEn) to predict coronary heart disease. The proposed ensemble algorithm was evaluated using Z-Alizadeh Sani dataset. The result analysis showed that the WAVEn algorithm achieved 98.97%, 100%, 96.3%, and 98.3% accuracy, sensitivity, specificity, and accuracy, respectively, for the main data set. The WAVEn algorithm applied to the balanced dataset achieves 100% accuracy, sensitivity, specificity, and accuracy in diagnosing coronary heart disease. From the author's point of view, the accuracy obtained by WAVEn is the highest accuracy compared to the advanced algorithms in studies for the original and balanced data set.[14] Joloudari *et al*. conducted a study with the aim of increasing the accuracy in diagnosing coronary heart disease by selecting significant predictive features in order of their ranking. In this study, an integrated method using machine learning was proposed. The techniques of random trees (RTs), decision tree of C5.0, SVM, and decision tree of Chi-squared automatic interaction detection were used in this study. The results showed that the RT model with 40 important features and 91.47% accuracy had better performance than other classification models. Another achievement of this study was the important rules extracted for the diagnosis of coronary heart disease using a RT model.[3] Abdar *et al*. described an innovative machine learning method for accurately diagnosing coronary heart disease. In this study, ten traditional machine learning algorithms were tested on the Z-Alizadeh Sani dataset, and then, three types of SVC (C-SVC), NuSVM (nu-SVC), and LinSVM with the best performance were used in the rest of the study. The results showed that the proposed method increased the performance of all traditional machine learning algorithms used in this study. The accuracy obtained for SVC algorithm was 92%, for LinSVM algorithm was 93%, and for nuSVM algorithm was 92%. The study also introduced a new optimization method called N2Genetic optimizer (a new genetic training). Experiments showed that N2Genetic-nuSVM method achieved 93.08% accuracy and F1 score 91.51%.[15] Haruna *et al*. used an improved data mining algorithm C4.5 to diagnose coronary heart disease. Performance evaluation of the improved algorithm was performed against the traditional C4.5 algorithm. As a result, the improved data mining algorithm C4.5 showed better performance with 97.23% overall accuracy, 97.03% specificity, and 96.39% sensitivity.[16]

According to the results of the mentioned researches, data mining techniques are suitable tools for predicting heart diseases and can help health policymakers in developing preventive programs.

After searching, screening, and evaluation during the systematic review, eventually, the final analysis was performed on 20 articles. Findings were presented in five sections: risk factors, datasets, data mining techniques used, best performance of techniques in terms of accuracy, sensitivity, specificity, and area under the curve, as well as the criteria used for measuring the performance of the algorithms.

In total, after a critical review of the articles, 20 articles in the field of data mining on coronary heart disease were obtained, all of which were in English.

According to Table 2, among the 8 datasets used in these studies, Z-Alizadeh Sani dataset (10 studies) and Cleveland Heart dataset (5 studies) had the highest frequency. The other datasets were each used in only one study.

The results of the studies on data mining techniques used in these studies are also presented in Table 2. According to these studies, a total of 30 techniques were used in these studies. SVM (6 studies) and naïve Bayes (4 studies) data mining techniques had the highest frequency. Data mining techniques PSO, Neural Network, SMO, and MLP were each used in three studies. Most data mining techniques (18 cases) had the lowest frequency and each of them was used in one study.

Table 2 also shows the criteria used to measure the performance of algorithms in studies and their frequency percentage. Accuracy (16 studies) with 80% frequency had the highest use in studies. Sensitivity and specificity (13 studies each) with 65% frequency were used in studies. Furthermore, the area under the curve (4 studies) with 20% frequency had the lowest use in studies.

According to Table 3, a total of 63 risk factors related to coronary heart disease were considered in the studies. Among the risk factors associated with coronary heart disease, the features of age (17 studies), blood pressure (17 studies), gender (16 studies), diabetes (16 studies), chest pain (16 studies), and smoking (15 studies), respectively, had the highest frequency in these studies.

On the other hand, features such as length of hospital stay (1 study), recovery time (1 study), and treadmill score (1 study) had the lowest frequency.

The above factors can be classified into four categories including demographic features, symptom and examination features, electrocardiographic features, and experimental and echocardiographic features.

The results for the number of best-performing data mining techniques in terms of the accuracy, sensitivity, specificity, and area under the curve are given in Table 4. Among the studies, naive Bayes technique (4 cases) and SVM, and SMO techniques (3 cases each) had the highest number of best performances for the sensitivity and specificity. SVM, and naïve Bayes techniques (4 cases each) and PSO, Neural Network, and SMO techniques (3 cases each) had the highest number of best performances for the accuracy. KNN, SVM,

Table 2: Articles information

Authors	Year	Technique	Dataset	Criteria
Velusamy and Ramasamy[14]	2021	KNN, SVM, random forest, WAVEn	Z-Alizadeh Sani dataset	Accuracy, sensitivity, specificity, AUC
Shahid and Singh[17]	2020	PSO-NFIS, PSO-EmNN	Z-Alizadeh Sani dataset	Accuracy, sensitivity, specificity
Joloudari *et al*.[3]	2020	RTs, C5.0, SVM, CHAID	Z-Alizadeh Sani dataset	AUC
Shahid *et al*.[18]	2020	PSO-ELM, NN-GA	Z-Alizadeh Sani dataset	Accuracy
Ricciardi *et al*.[19]	2020	LDA, LDA-PCA	Department of Advanced Biomedical Sciences, University Hospital Federico II of Naples, between 2004 and 2017	Accuracy, sensitivity, specificity
Zomorodi-Moghadam *et al*.[20]	2019	Hybrid PSO	Z-Alizadeh Sani dataset	Accuracy, sensitivity, specificity
Abdar *et al*.[21]	2019	nuSVM, LinSVM, SVC	Z-Alizadeh Sani dataset	Accuracy
Subramaniam and Mylswamy[22]	2019	SMO, naïve Bayes, NN, bagging SMO	Z-Alizadeh Sani dataset	Accuracy, sensitivity, specificity
Ayatollahi *et al*.[13]	2019	ANN, SVM	AJA University of Medical Sciences	Sensitivity, specificity
Haruna *et al*.[16]	2019	C4.5	Murtala Muhammad General Hospital and Abdullahi Wase General Hospitals in Kano State	Accuracy, sensitivity, specificity
Kolukisa *et al*.[23]	2019	Naïve Bayes, random forest, KNN, MLP	Cleveland heart dataset	Accuracy, sensitivity, specificity, AUC
Abdar *et al*.[15]	2019	J48, Bloom Filter tree, Reduced Error Pruning Tree, and adaptive Naïve Bayes tree. A multi filtering approach	Z-Alizadeh Sani dataset	Accuracy
Normawati and Winarti[24]	2018	VPRS, RIPPER	Cleveland heart dataset	Accuracy, sensitivity, specificity
Dhanaseelan and Jeya Sutha[25]	2018	HCFI, predictive Apriori	UCI Cleveland heart dataset	-
Davari Dolatabadi *et al*.[26]	2017	SVM	Long-term ST database	Accuracy, sensitivity, specificity
Noreen *et al*.[27]	2016	MLP, MLR, FURIA, C4.5	IGMC data Cleveland heart data set	Accuracy
Verma *et al*.[28]	2016	SVM, K-means clustering	UCI Cleveland heart data	Accuracy
Alizadehsani *et al*.[29]	2013	SMO, naïve Bayes, Bagging, NN	Z-Alizadeh Sani data set	Accuracy, sensitivity, specificity
Alizadehsani *et al*.[30]	2012	Naïve Bayes, SMO, Ensemble	Z-Alizadeh Sani data set	Accuracy, sensitivity, specificity
Kurt *et al*.[31]	2008	LR, regression tree, MLP, RBF, SOFM	Cardiology Clinic of Trakya University Medical Faculty in Turkey between January 2002 and February 2003	Sensitivity, specificity, AUC

KNN: K-nearest neighbor, LR: Logistic regression, SVM: Support vector machine, AUC: Area under the curve, PSO-NFIS: Particle swarm optimization-neuro -fuzzy inference system, PSO-EmNN: PSO based emotional neural network, PSO-ELM: PSO-extreme learning machine, NN-GA: Neural network-genetic algorithm, LDA: Linear discriminant analysis, LDA-PCA: LDA-principal component analysis, SVC: Support vector classifier, ANN: Artificial NN, MLR: Multinomial LR, WAVEn: Weighted Average Voting Ensemble, CHAID: Chi-square Automatic Interaction Detection, SMO : Sequential Minimal Optimization, MLP: Multi-layer Perceptron, HCFI: Hash table based Closed Frequent Itemsets, FURIA: Fuzzy Unordered Rule Induction Algorithm, VPRS: Variable Precision Rough Set, RIPPER: Repeated Incremental Pruning to Produce Error Reduction, UCI: University of California, Irvine repository for the machine learning, SOFM: Self-Organizing Feature Map, RBF: Radial Basis Function

and RF techniques (2 each) had the highest number of best performances for the area under the curve.

Discussion

A review of data mining studies on coronary artery disease information found that 85% of these studies considered age as a risk factor. With age, the risk of cardiovascular disease increases throughout a person's life. Age is currently considered an independent risk factor for assessing the risk of cardiovascular disease.[32,33]

Studies have also shown that gender, with a frequency of 80%, is the second most important risk factor in studies related to coronary heart disease. In a study to investigate the presence or absence of gender differences in the management of risk factors for coronary heart disease, the results showed that risk factor management for secondary prevention of coronary heart disease was generally worse in women than in men.[34] In other studies, the results showed that gender is effective in the prevalence and mortality of coronary heart disease.[35-38]

Table 3: Frequency of risk factors

Risk factor	Frequency (%)
Age	17 (85)
Blood pressure	17 (85)
Chest pain	16 (80)
Diabetes mellitus	16 (80)
Sex	16 (80)
Smoking	15 (75)
BMI	13 (65)
Family history	13 (65)
High density lipoprotein	13 (65)
Hyper tension	13 (65)
T inversion	13 (65)
Fasting blood sugar	12 (60)
Low density lipoprotein	12 (60)
Triglyceride	12 (60)
Dyspnea	11 (55)
Ejection fraction	11 (55)
Erythrocyte sedimentation rate	11 (55)
Hemoglobin	11 (55)
Obesity	11 (55)
Potassium	11 (55)
Pulse rate	11 (55)
Region RWMA	11 (55)
ST elevation	11 (55)
Weight	11 (55)
Airway disease	10 (50)
Blood urea nitrogen	10 (50)
Cerebrovascular accident	10 (50)
Chronic renal failure	10 (50)
Congestive heart failure	10 (50)
Creatine	10 (50)
Diastolic murmur	10 (50)
Dyslipidemia	10 (50)
Edema	10 (50)
Function class	10 (50)
Left ventricular hypertrophy	10 (50)
Low threshold angina	10 (40)
Lymphocyte	10 (50)
Neutrophil	10 (50)
Platelet	10 (50)
Sodium	10 (50)
Systolic murmur	10 (50)
Thyroid disease	10 (50)
Valvular heart disease	10 (50)
Weak peripheral pulse lung rales	10 (50)
White blood cell	10 (50)
Poor R-wave progression	9 (45)
Q-wave	9 (45)
ST depression	9 (45)
Electrocardiography rhythm	7 (35)
Cholesterol	6 (30)
Atrial fibrillation	3 (15)
Rhythm sin	3 (15)
Thalach: Maximum heart rate	3 (15)

Contd...

Table 3: Contd...

Risk factor	Frequency (%)
Ca: Number of fluoroscopy colored major vessels	2 (10)
Exang: Highlights existence of exercise-induced angina	2 (10)
Glucose	2 (10)
Restecg: Electrocardiographic results	2 (10)
Resting	2 (10)
Slope: The slope characteristics of the peak exercise ST segment	2 (10)
Thal: Heart status	2 (10)
Duke tread mill score	1 (5)
Duration recovery	1 (5)
Length of hospitalization	1 (5)

BMI: Body mass index, RWMA: Regional wall motion abnormalities

Another important risk factor in the results of this study was the blood pressure factor with a frequency of 85%. High blood pressure is a major controllable risk factor for all clinical manifestations of coronary heart disease.[39] An overview of randomized controlled trials and prospective observational studies provides the most reliable data on the association between high blood pressure and coronary heart disease. The overall evidence suggests a strong association between blood pressure and coronary heart disease.[40]

In this study, the risk factor for diabetes with a frequency of 80% was another important factor. Diabetes mellitus is associated with an increased risk of cardiovascular death and a higher incidence of cardiovascular disease, including coronary artery disease.[41] Diabetes has been specifically described as a cardiovascular risk factor in developed countries. In the Framingham study, the incidence of cardiovascular disease in diabetic men was twice as high as in nondiabetic men, and similarly, in diabetic women, it was three times higher than in nondiabetic women.[42]

In this study, chest pain with a frequency of 80% was another important risk factor. Chest pain, which is a common complaint in all health-care settings, is one of the causes of coronary heart disease.[43] Chest pain raises concerns about the development of a serious illness such as coronary heart disease.[44] Cardiovascular risk factors and a history of chest pain are associated with coronary heart disease and have been extensively studied.[45]

Another risk factor with a 75% frequency is smoking. Smoking has been well established as a risk factor for coronary heart disease and peripheral vascular disease.[46]

In general, studies show that many environmental characteristics have a significant impact on the risk, progression, and severity of cardiovascular disease. Evidence supports the notion that ecological features such as daily cycles of light and day, sun exposure, seasons, and geographical features of the natural environment are important determinants of cardiovascular health. In highly developed societies, the effect of the natural

Table 4: Frequency of best performing data mining techniques in terms of sensitivity, accuracy, specificity, and area under the curve

Technique	Frequency of best sensitivity performance	Frequency of best accuracy performance	Frequency of best specificity performance	Frequency of best AUC performance
KNN	2	2	2	2
SVM	3	4	3	2
Random forest	2	2	2	2
Naïve Bayes	4	4	4	1
PSO	2	3	2	-
Random trees	-	-	-	1
C5.0	-	-	-	1
CHAID	-	-	-	1
NN	2	3	2	-
LDA	1	1	1	-
SMO	3	3	3	-
Bagging	1	1	1	-
ANN	1	1	-	-
C4.5	1	1	2	-
J48	-	-	1	-
BF tree	-	-	1	-
REP tree	-	-	1	-
Naive Bayes Tree	-	-	1	-
VPRS	1	1	1	-
RIPPER	1	1	1	-
HCFI	-	-	-	-
NAFCP	-	-	-	-
PredictiveApriori	-	-		-
MLR	-	-	1	-
FURIA	-	-	1	-
Ensemble	2	2	2	1
LR	1	1	1	1
CART	1	1	-	1
RBF	1	1	-	1
SOFM	1	1	-	1

LR: Logistic regression, KNN: K-nearest neighbor, SVM: Support vector machine, PSO: Particle swarm optimization, NN: Neural network, LDA: Linear discriminant analysis, ANN: Artificial NN, MLR: Multinomial LR, AUC: Area under the curve, CHAID: Chi-square Automatic Interaction Detection SMO : Sequential Minimal Optimization, VPRS: Variable Precision Rough Set, HCFI: Hash table based Closed Frequent Itemsets, NAFCP: N-listbased algorithm for mining FCPs (Frequent closed patterns), FURIA: Fuzzy Unordered Rule Induction Algorithm, RBF: Radial Basis Function, SOFM: Self-Organizing Feature Map

environment is balanced by the physical characteristics of social environments such as the built environment and pollution, as well as by socioeconomic status and social networks. These features of the social environment alter lifestyle choices that significantly alter the risk of cardiovascular disease. Understanding how different domains of the environment, individually and collectively, affect cardiovascular disease risk can lead to better assessment of the disease and help develop new prevention and treatment strategies to limit heart disease.[47]

The Z-Alizadeh Sani dataset was the most widely used dataset among the studies with a frequency of 50%. The Z-Alizadeh Sani dataset contains the records of 303 patients, each of whom has 54 features. These features are categorized into four groups: Demographics, symptoms and examination, ECG, and laboratory features and echoes. The Z-Alizadeh Sani dataset is made up of information provided by 303 random visitors to the Shahid Rajaei Cardiovascular, Medical and Research Center. 216 samples had coronary heart disease and the rest were healthy.[29] One of the important features of this data set is the absence of missing values.[48]

The second most widely used dataset among studies with a frequency of 25% was the Cleveland Heart dataset. The Cleveland dataset contains information on the diagnosis of heart disease. Data were collected from the Cleveland Clinic Foundation and are available in the UCI Machine Learning Repository. This dataset contains 303 data samples, of which only six samples have missing values. Each sample is classified into one of two groups of patients with coronary heart disease and healthy.[23,49]

SVM data mining technique has been most used in predicting coronary heart disease and has been able to achieve the best performance in 50% of cases. The SVM technique has shown its efficiency in many pattern recognition techniques.[50] This technique has a good ability to generalize hidden test data. Therefore, it can perform well in the field of survival detection, where hidden test data are important.[51] The neural network technique has been used 3 times in studies but has had the best performance in all three times. This is due to the flexibility of this technique, considering that this technique is multi-layered and is based on the neuron processing unit. Therefore, since neurons have high flexibility, they can apply this flexibility in different layers of neural network technique.[52] The naïve Bayes technique has performed best in all four studies. This technique is a probabilistic technique and has similarities with the linear regression technique, so it can somehow find the relationship between each variable and the target variable.[53]

Accuracy criterion with 80% frequency has been the most used among the studied criteria. This criterion is one of the most common quality evaluation criteria in data mining.[54] Sensitivity and specificity are the second criteria used with 65% frequency. These two criteria are always used together.[55] The sensitivity criterion corresponds to the recall criterion, which is one of the basic criteria in data mining.[56] Usually, these medical data have no formal structure and are in fact heterogeneous. Therefore, compliance with the above criteria with the characteristics of medical data is one of the reasons for their selection.[57] The criterion of the area under the curve has the lowest and has been used in only 20% of studies.

Conclusion

Techniques such as SVM and neural network performed better than other techniques. These techniques are mostly used for the field of health and are used in various fields. The use of these techniques can provide a good basis for clinicians in the field of cardiology to evaluate the characteristics of different patients at a lower cost so that they can increase the risk of prediction coronary heart disease in patients The output of these techniques can then be used as a decision support system so that clinicians can enter various risk factors such as age, blood pressure, gender, diabetes, and chest pain. They can view the output from the system and make the appropriate decision based on the output from the system. In fact, in these systems, with a retrospective approach, better decisions can be made in the future. Of course, there is a point that one technique may work well for one disease and not work well for another. Data mining specialists can suggest the best technique for these systems through numerous studies.

Ethics Code: IR.MUI.RESEARCH.REC.1399.785.

Financial support and sponsorship

This article is based on a research project : #399932 approved in the Isfahan University of Medical Sciences (IUMS). The authors wish to acknowledge the Vice Chancellery of Research of IUMS for financial support. Ethics Code IR.MUI.RESEARCH.REC.1399.785

Conflicts of interest

There are no conflicts of interest.

References

1. Ali F, El-Sappagh S, Islam SM, Kwak D, Ali A, Imran M, *et al*. A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion. Inf Fusion 2020;63:208-22.
2. Fuster V, Kelly BB. Promoting cardiovascular health in the developing world: A critical challenge to achieve global health. In: Promoting Cardiovascular Health in the Developing World: A Critical Challenge to Achieve Global Health. USA: National Academies Press; 2010. p. 1-482. Available from: https://www.ncbi.nlm.nih.gov/books/NBK45693/. [Last accessed on 2020 Aug 18].
3. Joloudari JH, Joloudari EH, Saadatfar H, GhasemiGol M, Razavi SM, Mosavi A, *et al.* Coronary artery disease diagnosis; Ranking the significant features using a random trees model. Int J Environ Res Public Health 2020;17:731.
4. Mastoi QU, Wah TY, Gopal Raj R, Iqbal U. Automated diagnosis of coronary artery disease: A review and workflow. Cardiol Res Pract 2018;2018:2016282.
5. Setiawan NA, Venkatachalam PA, Hani AF. Diagnosis of coronary artery disease using artificial intelligence based decision support system. arXiv 2020.
6. Bender JR. Yale University School of Medicine Heart Book. New York: William Morrow Company, Inc.,; 1992. p. 167-75.
7. Md Idris N, Chiam YK, Varathan KD, Wan Ahmad WA, Chee KH, Liew YM. Feature selection and risk prediction for patients with coronary artery disease using data mining. Med Biol Eng Comput 2020;58:3123-40.
8. Cardiovascular Diseases. Available form: https://www.who.int/health-topics/cardiovascular-diseases/#tab=tab_1. [Last accessed on 2020 Dec 21].
9. Amin MS, Chiam YK, Varathan KD. Identification of significant features and data mining techniques in predicting heart disease. Telemat Inform 2019;36:82-93.
10. Mazaheri S, Ashoori M, Bechari Z. Model to predict heart disease treatment using data mining. Payavard 2017;11:287-96.
11. Dutta A, Batabyal T, Basu M, Acton ST. An efficient convolutional neural network for coronary heart disease prediction. Expert Syst Appl 2020;159:18-34.
12. Kolukisa B, Yavuz L, Soran A, Bakir-Gungor B, Tuncer D, Onen A, *et al.* Coronary artery disease diagnosis using optimized adaptive ensemble machine learning algorithm. Int J Biosci Biochem Bioinform 2020;10:10(1):58-65.
13. Ayatollahi H, Gholamhosseini L, Salehi M. Predicting coronary artery disease: A comparison between two data mining algorithms. BMC Public Health 2019;19:448.
14. Velusamy D, Ramasamy K. Ensemble of heterogeneous classifiers for diagnosis and prediction of coronary artery disease with reduced feature subset. Comput Methods Programs Biomed 2021;198:57-70.
15. Abdar M, Nasarian E, Zhou X, Bargshady G, Wijayaningrum VN, Hussain S. Performance improvement of decision trees for diagnosis of coronary artery disease using multi filtering approach. In: 2019 IEEE 4th International Conference on Computer and Communication Systems, ICCCS 2019.Singapoor: IEEE; 2019. p. 26-30.
16. Haruna AA, Muhammad LJ, Yahaya BZ, Garba EJ, Oye ND, Jung LT. An improved C4.5 data mining driven algorithm for the diagnosis of coronary artery disease. In: Proceeding of 2019 International Conference on Digitization: Landscaping Artificial Intelligence, ICD 2019. UAE: IEEE; 2019. p. 48-52.
17. Shahid AH, Singh MP. A novel approach for coronary artery disease diagnosis using hybrid particle swarm optimization based emotional neural network. Biocybern Biomed Eng 2020;40:1568-85.
18. Shahid AH, Singh MP, Roy B, Aadarsh A. Coronary artery disease diagnosis using feature selection based hybrid extreme learning machine. In: Proceedings – 3rd International Conference on Information and Computer Technologies, ICICT 2020. USA: IEEE ; 2020. p. 341-6.
19. Ricciardi C, Valente AS, Edmund K, Cantoni V, Green R, Fiorillo A,

et al. Linear discriminant analysis and principal component analysis to predict coronary artery disease. Health Informatics J 2020;26:2181-92.
20. Zomorodi-Moghadam M, Abdar M, Davarzani Z, Zhou X, Pławiak P, Acharya UR. Hybrid particle swarm optimization for rule discovery in the diagnosis of coronary artery disease. Expert Syst 2021;38:e12485.
21. Abdar M, Książek W, Acharya UR, Tan RS, Makarenkov V, Pławiak P. New machine learning technique for an accurate diagnosis of coronary artery disease. Comput Methods Programs Biomed 2019;179:104992.
22. Subramaniam O, Mylswamy R. Ant colony optimization based support vector machine towards predicting coronary artery disease. Int J Recent Technol Eng 2019;7:210-5.
23. Kolukisa B, Hacılar H, Goy G, Kus M, Bakir-Gungor B, Aral A, *et al.* Diagnosis of coronary heart disease via classification algorithms and a new feature selection methodology. Int J Data Min Sci 2019;1:8-15.
24. Normawati D, Winarti S. Feature selection with combination classifier use rules-based data mining for diagnosis of coronary heart disease. In: Proceeding of 2018 12th International Conference on Telecommunication Systems, Services, and Applications, TSSA 2018. Indonesia: IEEE; 2018.
25. Dhanaseelan R, Jeya Sutha M. Diagnosis of coronary artery disease using an efficient hash table based closed frequent itemsets mining. Med Biol Eng Comput 2018;56:749-59.
26. Davari Dolatabadi A, Khadem SE, Asl BM. Automated diagnosis of coronary artery disease (CAD) patients using optimized SVM. Comput Methods Programs Biomed 2017;138:117-26.
27. Noreen K, Azween A, Belhaouari SB, Sellapan P, Saeed AB, Nilanjan D. Ensemble clustering algorithm with supervised classification of clinical data for early diagnosis of coronary artery disease. J Med Imaging Heal Inform 2016;6:78-87.
28. 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;40:178.
29. 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;111:52-61.
30. Alizadehsani R, Habibi J, Hosseini MJ, Boghrati R, Ghandeharioun A, Bahadorian B, *et al.* Diagnosis of coronary artery disease using data mining techniques based on symptoms and ECG features. Eur J Sci Res 2012;82:542-53.
31. Kurt I, Ture M, Kurum AT. Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. Expert Syst Appl 2008;34:366-74.
32. Dhingra R, Vasan RS. Age as a risk factor. Medical Clinics of North America. Vol. 96. USA: NIH Public Access; 2012. p. 87-91. Available from: https://pmc/articles/PMC3297980/. [Last accessed on 2021 Feb 09].
33. Huxley R. The impact of cardiovascular risk factors on the age-related excess risk of coronary heart disease. Int J Epidemiol 2006;35:1025-33.
34. Zhao M, Vaartjes I, Graham I, Grobbee D, Spiering W, Klipstein-Grobusch K, *et al*. Sex differences in risk factor management of coronary heart disease across three regions. Heart 2017;103:1587-94.
35. Brochier ML, Arwidson P. Coronary heart disease risk factors in women. Eur Heart J 1998;19 Suppl A: A45-52.
36. Judelson DR. Coronary heart disease in women: Risk factors and prevention. J Am Med Womens Assoc 1994;49:186-91.
37. Wingard DL. Sex differences and coronary heart disease. A case of comparing apples and pears? Circulation 1990;81:1710-2.
38. Castanho VS, Oliveira LS, Pinheiro HP, Oliveira HC, de Faria EC. Sex differences in risk factors for coronary heart disease: A study in a Brazilian population. BMC Public Health 2001;1:3.
39. Weber T, Lang I, Zweiker R, Horn S, Wenzel RR, Watschinger B, *et al*. Hypertension and coronary artery disease: Epidemiology, physiology, effects of treatment, and recommendations: A joint scientific statement from the Austrian Society of Cardiology and the Austrian Society of Hypertension. Wien Klin Wochenschr 2016;128:467-79.
40. Lawes CM, Bennett DA, Lewington S, Rodgers A. Blood pressure and coronary heart disease: A review of the evidence. Semin Vasc Med 2002;2:355-68.
41. Chiha M, Njeim M, Chedrawy EG. Diabetes and coronary heart disease: A risk factor for the global epidemic. USA:ACM, Int J Hypertens 2012;2012:1-7.
42. Kannel WB, McGee DL. Diabetes and glucose tolerance as risk factors for cardiovascular disease: The Framingham study. Diabetes Care 1979;2:120-6.
43. Haasenritter J, Stanze D, Widera G, Wilimzig C, Abu Hani M, Sonnichsen AC, *et al*. Does the patient with chest pain have a coronary heart disease? Diagnostic value of single symptoms and signs – A meta-analysis. Croat Med J 2012;53:432-41.
44. Luepker RV, Apple FS, Christenson RH, Crow RS, Fortmann SP, Goff D, *et al.* Case definitions for acute coronary heart disease in epidemiology and clinical research studies: A statement from the AHA Council on Epidemiology and Prevention; AHA Statistics Committee; World Heart Federation Council on Epidemiology and Prevention; the European Society of Cardiology Working Group on Epidemiology and Prevention; Centers for Disease Control and Prevention; and the National Heart, Lung, and Blood Institute. Circulation 2003;108:2543-9.
45. Gencer B, Vaucher P, Herzig L, Verdon F, Ruffieux C, Bösner S, *et al*. Ruling out coronary heart disease in primary care patients with chest pain: A clinical prediction score. BMC Med 2010;8:9.
46. McGill HC. The cardiovascular pathology of smoking. Am Heart J 1988;115:250-7.
47. Bhatnagar A. Environmental Determinants of Cardiovascular Disease. Circulation Research. Vol. 121. USA: Lippincott Williams and Wilkins; 2017. p. 162-80.
48. UCI Machine Learning Repository: Z-Alizadeh Sani Data Set. Available from: https://archive.ics.uci.edu/ml/datasets/Z-Alizadeh+Sani. [Last accessed on 2021 Feb 11].
49. Heartc: The Heart Cleveland Dataset in Dprep: Data Pre-Processing and Visualization Functions for Classification. Available from: https://rdrr.io/cran/dprep/man/heartc.html. [Last accessed on 2021 Feb 12].
50. Byun H, Lee SW. A survey on pattern recognition applications of support vector machines. Int J Pattern Recognit Artif Intell 2003;17:459-86.
51. Huang MW, Chen CW, Lin WC, Ke SW, Tsai CF. SVM and SVM Ensembles in Breast Cancer Prediction. PLoS One 2017;12:e0161501.
52. Nomi JS, Vij SG, Dajani DR, Steimke R, Damaraju E, Rachakonda S, *et al.* Chronnectomic patterns and neural flexibility underlie executive function. Neuroimage 2017;147:861-71.
53. Lowd D, Domingos P. Naive Bayes models for probability estimation. In: Proceedings of the 22nd International Conference on Machine Learning. USA:ACM; 2005. p. 529-36.
54. Guo T, Milanović JV. Probabilistic framework for assessing the accuracy of data mining tool for online prediction of transient stability. IEEE Trans Power Syst 2013;29:377-85.
55. Trevethan R. Sensitivity, specificity, and predictive values: Foundations, pliabilities, and pitfalls in research and practice. Front Public Heal 2017;5:307.
56. Cios KJ, Moore GW. Uniqueness of medical data mining. Artif Intell Med 2002;26:1-24.
57. Maciejewski T, Stefanowski J. Local neighbourhood extension of SMOTE for mining imbalanced data. In: 2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM). France: IEEE; 2011. p. 104-11.