



Paper Type: Research Paper



# Data Mining and Diagnosis of Heart Diseases: A Hybrid Approach to the B-Mine Algorithm and Association Rules

Shokoofa Mostofi<sup>1</sup>, Sohrab Kordrostami<sup>1,\*</sup> , Amirhossein Refahi Sheikhani<sup>1</sup>, Marzieh Faridi Masouleh<sup>2</sup>, Soheil Shokri<sup>1</sup>

<sup>1</sup> Department of Mathematics, Lahijan Branch, Islamic Azad University, Lahijan, Iran; shokoofamostofi@yahoo.com; sohrabkordrostami@gmail.com; ah-refahi@gmail.com; soheilshokri@gmail.com.

<sup>2</sup> Computer and Information Technology Department, Ahrar Institute of Technology and Higher Education, Rasht, Iran; m.faridi@ahrar.ac.ir.

Citation:



Mostofi, Sh., Kordrostami, S., Refahi Sheikhani, A., Faridi Masouleh, M., & Shokri, S. (2022). Data mining and diagnosis of heart diseases: a hybrid approach to the b-mine algorithm and association rules. *International journal of research in industrial engineering*, 11 (1), 77-91.

Received: 01/09/2021

Reviewed: 17/10/2021

Revised: 05/03/2022

Accepted: 17/03/2022

## Abstract

Existing systems for diagnosing heart disease are time consuming, expensive, and prone to error. In this regard, a diagnostic algorithm has been proposed for the causes of heart disease based on a frequent pattern with the B-mine algorithm optimized by association rules. Initially, a data set of disease is used to select a feature, so that it deals with a set of training features. Then, association rules are used to classify educational and experimental sets, and then the factors affecting heart disease are analyzed. The numerical results from the experiments of real and standard datasets of cardiac patients show that the average accuracy of the proposed method is approximately 98%, which has been tested on the Cleveland database that includes 76 features in the case of heart disease dataset, 14 features of which are related to heart disease. This paper also uses four common categories such as decision tree to build the model. The data set studied in this article contains 270 records as well as 14 features. The accuracy of predicting the results of the support vector machine classifications, k nearest neighbor, decision tree and simple Bayesian is 81.11%, 66.67%, 59.72% and 19.85%, respectively, which are relatively satisfactory results.

**Keywords:** Frequent pattern, Heart disease, Data mining, B-mine algorithm, Association rules.

## 1 | Introduction

Cardiovascular Disease (CVD) is the leading cause of death in the world. Every year, 17.3 million people die from it and this number is expected to increase to 23.3 million by 2030. Therefore, it is very important to minimize the high mortality rate of heart diseases [1]. Data mining uses statistical methods and Artificial Intelligence (AI) to extract behavioral patterns from users and gain a variety of insights about them. Data mining can be widely used to support marketing decisions. The most important tasks in data mining are the discovery of repetitive elements and association rules. The dependencies and connections between data in a database can be discovered using association rules [2]. Data mining came up in the late 1980s.

 Licensee  
International Journal of Research in Industrial Engineering. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license  
(<http://creativecommons.org/licenses/by/4.0>).



Corresponding Author: sohrabkordrostami@gmail.com



<http://dx.doi.org/10.22105/riej.2022.302672.1243>

Great strides were made in this branch of science in the 1990s, and it is expected to continue to grow and develop in this century, and predictions indicate that it will experience revolutionary developments in the coming decades. The Massachusetts Institute of Technology considers data mining as one of the top ten technologies that will play a significant role in the development of the world [3]. With the growth of information technology and methods of data creation and collection, databases of data in commerce, agriculture, the internet, details of phone calls, medical records, etc. are being captured and stored at a faster rate every day. Therefore, since the late 1980s, the idea of accessing the information hidden in these large databases has been in the forefront, as traditional systems have not been able to do so. Due to competition in the political, military, economic and scientific fields and the importance of accessing information in the shortest possible time without human intervention, the science and analysis of data or data mining came into play. Data mining is a technique proposed in the early 1990s to deal with the problem of extracting data from a database using a new approach. Data mining has seriously entered the field of statistics since 1995 and the first issue of Knowledge Discovery Journal was published from the database in 1996.

Although knowledge discovery entered the healthcare field with the aim of detecting misappropriation, it was gradually applied in the clinical field as well. This is due to the rapid change in awareness of information in the healthcare field.

The healthcare industry is constantly generating large amounts of data [4] and [5], and people who are exposed to this type of data have found that there is a large gap between collecting and interpreting this data. The relatively young and growing field of data mining in healthcare [6] is one of the methods that can benefit this industry through the in-depth analysis of these data and lead to the development of medical research and scientific decisions in the field of diagnosis and treatment [5] and [6].

Data mining in medicine and biology is an important part of biomedical informatics and is one of the most commonly applied computer sciences in this science, used in hospitals, clinics, laboratories and research centers [4].

The process of knowledge discovery in databases is a scientific process of identifying valid, new, potentially useful and understandable patterns in data. The most important part of this process is data mining, where certain algorithms are used to extract a set of patterns from the database. Data mining is the extraction of hidden and potentially useful information and knowledge in a database. Nowadays, this topic has become very important due to the large amount of growing data and its diversity [3]. Data mining is a set of techniques that allow to go beyond ordinary data processing and help to discover information hidden in the huge amounts of data. The discovery and extraction of valuable data in the form of a set of dependency relationships is an important area of research in data mining, which has recently drawn the attention of many researchers to the creation of useful techniques for extracting these relationships, given the increasing use of large databases and large transaction warehouses [7].

The innovation of this method is that, in the first case, a person is selected from the list of people who are ill due to common characteristics. We try to extract the patients related to the selected person in the considered groups. Then, based on the selected person, the system extracts the causal rules between the selected person and the other patients. In order to improve the system, this method uses two patient similarity criteria simultaneously, namely coverage and reliability.

In the rest of this paper, a review of the literature and the work done is given in Section 2. Details of the proposed hybrid method are given in Section 3, the evaluation criteria are presented in Section 4, and the test results are given in Section 5. Finally, a conclusion of the whole work is given in Section 9.

## 2 | Literature Review

Initially in 1994, a model of Intelligent Heart Disease Prediction System (IHDPS) integration of data algorithms, such as Naive-Bayes, Decision trees and Network Neural was proposed. The final output of this algorithm describes that each method has a different capacity to achieve data mining goals. IHDPS is simple, easy to use, scalable, upgradeable and reliable, based on the web-based prediction system. And high blood sugar is used to diagnose patients' symptoms [8]. In IHDPS, the author uses medical specifications such as age range, gender, high blood pressure, and high blood sugar to diagnose patients' symptoms [8]. This runs on the NET platform.

Ordonez [9] presented an improved study in 2004 using legal rules to predict heart disease. The study was conducted to diagnose heart disease. Evaluation data covered by the medical records of people with heart disease are related to features such as chest pain, blood pressure, cholesterol and blood sugar for risk factors. The heart flow cytometry was measured and the arterial lesions were observed by the author.

Yan et al. [10] used Multilayer Perceptron (MLP) with 40 input variables for input layer and output layer with 5 nodes. The Backpropagation algorithm has been improved for system training. 352 medical records were collected for system training and testing. Evaluation methods such as cross-validation, maintenance and bootstrapping were used to evaluate the system. MLP is known for its good architecture, which makes this method more important in NN models, and also its algorithms are very simple and easy to understand. MLP consists of 3 layers, namely the input layer, the hidden layer and the output layer. The hidden layer was obtained by the waterfall learning process. The experimental results were 90% accurate.

Domadiya and Rao [11] performed a clinical observation and a 6-month follow-up of 1,000 CHD cases. Based on the data, they used three popular data mining algorithms to develop the prediction model using 502 items. They also used 10-fold validation methods to measure the neutral estimate of the three prediction models for performance comparison purposes. The results showed that SVM is the best predictor with 92.1% accuracy and neural networks with 91.0% accuracy and decision tree with 89.6% accuracy.

Another study was conducted by Amin et al. [12] with the theme of data mining systems to evaluate heart event related risk factors using statistical analysis based on apriori algorithm. The events included MI, PCI and CABG. Our risk factors were gender (male), smoking, high-density lipoprotein, glucose, family history, and a history of hypertension.

De Cnudde [13] introduce techniques that use software version of the Statistical Analysis System (SAS) version 9.1.3 to investigate basic heart disease. The research was conducted on the heart disease dataset by a completely automatic method using 3 independent neural network models to develop group models. The authors received an accuracy of 89.01%, 95.91% and 80.95% for classification, specification and sensitivity, respectively, in the data extracted from the Cleveland Heart Disease Database.

Gupta et al. [14] worked on using a modular neural network to diagnose heart disease. In their work, mainly two types of diagnostic methods were used, one manually and the other an automatic diagnosis, which consists of diagnosing the disease with the help of an intelligent expert system.

There is another way to predict heart disease using classification techniques. Three classification algorithms, decision trees, naive Bayes, and neural networks have been used and their performances have been compared. The results show that the accuracy of neural networks is 100% and completely predicts heart disease, when the number of traits is 15. But accuracy is lower when the number of traits is 13 [15]. The authors also used the neural network to diagnose heart disease and then used the genetic algorithm to improve the accuracy of neural networks [16]. Jabbar [17] proposed an "intelligent data mining method for diagnosing heart disease". The authors tried to increase the accuracy for heart disease. They used segregation preprocessing techniques and genetic algorithms that were applied to the naive classification.

Ani et al. [18] used four data classification methods to predict CVD. The results of this paper show that the accuracy level of random forests, Naive Bayes, C4.5 and KNN decision trees is 89%, 84%, 81% and 77%, respectively. Also work on a specific dataset is provided using the classification method and feature selection. In the mentioned method, first the instruction set is divided into two groups of healthy and sick people, then in the second stage, 8192 subsets are extracted from the total of features with a clear cost for each subset (increase of cost from feature ranking). In the third step, the PSO algorithm for all subsets is the learning classification (FFBP) for all subsets to find the best subset with the highest accuracy and accessibility and the lowest cost and time [19]. A PSO-based algorithm is provided for extracting classification rules. The algorithm has been compared with the decision tree based on the C4.5 algorithm in the UCI machine learning database. Experimental results show that the PSO algorithm has been predicted to be accurate and the list of rules is much smaller than C4.5. Based on the average accuracy, the accuracy of the PSO method is 87% and the accuracy of C4.5 is 63%. Using PSO, effective classification rules can be extracted with acceptable accuracy [20]. Jabbar et al. [21] also points to the prediction of heart disease based on PSO and KNN. This article also examines the PSO-based feature selection measurement tool for selecting a small number of features and improving classification performance. The results show that the proposed approach can significantly improve learning accuracy. In a research paper, Jabbar et al. [21] presented an effective approach to predicting heart disease using a random forest. The use of random forests has shown high accuracy in predicting heart disease. Chauhan et al. [22] discussed a variety of classification techniques for predicting heart attacks and determining the best classification among them. The results can be useful for discovering patterns among health care professionals to determine which classification can be beneficial during a prediction process. Also, various data mining techniques for diagnosing heart disease [23] and based on the Markov chain [2] and so on were proposed. The increasing volume of information and the lack of new, useful and understandable knowledge have led the design of knowledge discovery algorithms to be considered by researchers. This algorithm has been designed to process discrete data, so we have to quantify the data to use it in continuous data. This causes some data to be lost and unrealistic data to be added to the data space, which in turn can affect the final results. We used the proposed B-mine algorithm in this study for knowledge discovery in database, which is discussed in the next part of the paper.

There are numerous papers dealing with disease prediction systems using various data mining techniques and machine learning algorithms in medical centers.

Giudici and Castelo [23] proposed the prediction of heart disease using the multiple regression model and proved that multiple linear regression is suitable for predicting the probability of heart disease. The work is done with a training dataset consisting of 3000 instances with 13 different attributes mentioned earlier. The dataset is divided into two parts, i.e. 70% of the data is used for training and 30% for testing. From the results, it is clear that the classification accuracy of the regression algorithm is better compared to other algorithms.

Polaraju and Prasad [24] developed a prediction of heart disease using KStar, j48, SMO and Bayes Net as well as a multi-layer perception using the software WEKA. Based on the performance of various factors, SMO and Bayes Net achieve better performance than KStar, Multilayer Perception and J48 techniques using kfold cross validation. However, the accuracy performances achieved by these algorithms are not satisfactory. Therefore, there is a need to improve the accuracy performance for better decision making in disease diagnosis.

Sultana and Haider [25] deals with chronic disease prediction techniques by analysing data from historical health records using Naïve Bayes, Decision Tree, Support Vector Machine (SVM) and Artificial Neural Network (ANN). A comparative study is conducted on the classifiers to measure better performance in terms of accuracy rate. From this experiment it is found that SVM gives the highest accuracy rate while Naïve Bayes gives the highest accuracy in diabetes.

Deepika and Seema [26] recommended the prediction and analysis of heart disease occurrence using data mining techniques. The main objective is to predict the occurrence of heart disease to enable early automatic diagnosis of the disease within a short period of time. The proposed methodology is also crucial for healthcare organisations where experts lack other knowledge and skills. It uses various medical attributes such as blood sugar and heart rate, age and gender to determine whether a person has a heart disease or not. The analyses of the dataset are calculated using the software WEKA.

Beyene and Kamat [27] suggested risk prediction system of CVD using machine learning techniques. It continues as the main cause of mortality in upcoming twenty years. Its major goal is to apply the findings of this study to existing methodologies. Machine learning techniques are utilized to develop a clinician's treatment plan and diagnosis using AI. This paper briefly addresses the key modules of systems as well as related theories. The suggested technique combines AI and data mining to get precise results with the fewest errors. This study establishes a foundation for the development of a new type of risk prediction system in the field of CVD.

Patil et al. [28] proposed to predict heart disease using data mining and machine learning algorithms. The aim of this study is to extract hidden patterns by using data mining techniques. The best algorithm J48 based on UCI data has the highest hit rate compared to LMT.

Gupta et al. [29] proposed an efficient system for heart disease prediction using data mining. This system helps physicians to make effective decisions based on certain parameters. By testing and training a particular parameter, it provides accuracy of 86.3% in testing phase and 87.3% in training phase.

Saxena and Sharma [30] proposed to predict several diseases using data mining techniques. By using data mining techniques the number of tests can be reduced. This paper is mainly about prediction of heart diseases, diabetes and breast cancer etc.

Gomathi and Priyaa [31] proposed prediction of heart diseases using ANN algorithm in data mining. Due to the increasing cost of diagnosis of heart diseases, there was a need to develop a new system that can predict heart diseases. The prediction model is used to predict the condition of the patient after evaluation based on various parameters like heart beat rate, blood pressure, cholesterol etc. The accuracy of the system is proved in Java.

Reddy et al. [32] have recommended to develop a prediction system to diagnose heart diseases from the patient's medical record. While developing the system, 13 risk factors were considered for the input attributes. After analysing the data from the dataset, data cleaning and data integration were performed.

Ramotra et al. [33] proposed data mining techniques and machine learning to predict heart diseases. There are two objectives to predict the heart system. This system does not require any prior knowledge about the patient's data. 2. The chosen system must be scalable to work with a large number of data sets. This system can be implemented using the software WEKA. The classification tools and explorer mode of WEKA are used for testing. Aldhyani et al. [34] developed various data mining techniques to evaluate the prediction and diagnosis of heart diseases. The main objective is to evaluate the different classification techniques like J48, decision tree, KNN, SMO and Naïve Bayes. Then the performance is evaluated and compared in terms of accuracy, precision, sensitivity and specificity. J48 and decision tree provide the best technique for heart disease prediction. Bahrami and Shirvani [35] used data mining for heart disease prediction.

## 3 | The Proposed Method

### 3.1 | Association Rules

Hidden associations between data values in a database are called association rules. The process of finding the association rules is called association rule discovery. The most important step in finding association rules is to find the set of frequent elements. Suppose  $I = \{a, b, c, d, e\}$  is a set of elements. Suppose  $D$  is a set of database records, each of  $T$  records is a subset of elements, i.e.  $T \subseteq I$ . Each record has a unique number called TID. Frequent patterns are patterns that occur frequently in a record. For example, in a database related to a store's purchase history, milk and bread that are often purchased together are examples of frequent patterns. We indicate a set of elements containing element  $k$  with  $k$ -element set. For example, the set  $\{\text{computer, monitor}\}$  is an itemset-2. The number of repetitions of a set of items is repetitive if the number of repetitions of that set of items is greater than or equal to the value (minimum support \* number of records in  $D$ ). If a set of elements has the required number of transactions for it, we call it a set of frequent elements. In this paper, we use the B-mine algorithm.

### 3.2 | B-Mine Algorithm

The discovery of a set of frequent items is used to explore information in transaction databases. In this proposed method, a proposed B-table indicates a relationship between transactions and contains values of zero and one, and zero indicates that there is no relationship between items in a transaction. It does not exist, and if the value of the table contains the value of one, then it indicates that there is a relationship between the items in a transaction. This means that in this study, the value of one indicates the connection of nodes (patients).

## 4 | Extraction of the Frequent Pattern

This section consists of two parts:

- I. First mode choosing a specific disease.
- II. Second mode selecting the group of nodes (patients).

### 4.1 | First Mode Choosing a Specific Disease

One person is selected from the list of people (patients, for example, heart patients) who have common characteristics, and we seek to extract the nodes related to the selected person in the favorite groups, then the system extracts the causal rules between the selected person and the other nodes according to the selected person.

The system consists of three phases.

**First phase.** Creating the required system table that is made from the total database of nodes.

- I. In this system, first a person is selected, the system creates the table related to the nodes (patients) in question, in fact, all those who have the above disease are selected.
- II. The system will examine people who have the disease in the member groups and other people will be pruned.
- III. The threshold is equal to half the number of people who have the disease and is used to select nodes (patients) with higher accuracy.

**Second phase.** At this stage, the set of the desired node is obtained.

- I. The cumulative frequency of all nodes (patients) that the individuals have is calculated. Nodes (patients) whose frequency is greater than or equal to the threshold are selected.
- II. Two-member sets are made up of nodes in the previous stage, and any set that is less than the threshold (number of simultaneous selections) is removed. Each set includes the disease selected by the individual and with another disease.
- III. The previous stage to make larger sets of nodes continues until the largest set whose frequency is not less than the threshold is created.

**Third phase.** At this stage, the obtained sets are examined according to the following two criteria.

- I. The confidence relationship is evaluated.
- II. The support relationship is evaluated.

#### 4.2 | Second Mode Selecting the Group of Nodes (Patients)

In this mode, no target disease is selected for the person and the group only determines the disease (such as the heart group) and the system recommends some diseases to the person according to previous information.

If the person chooses a disease that no one else has chosen before, the system will suggest a disease with its second mode.

The procedure is as follows:

- I. The person chooses the disease group.
- II. The system creates a table for the target group.
- III. All possible rules are produced according to the threshold.
- IV. Rules that are acceptable in terms of support and confidence are presented to the person as a suggestion.

#### 4.3 | Pre-Processing and Preparation of Data

Data preprocessing is the most important and time-consuming step in data mining projects. Approximately 60 to 90 percent of the time spent on a data mining project is spent at this stage, and 75 to 90 percent of the success of data mining projects depends on it. The processes performed in preprocessing are aggregation, sampling, dimensional reduction, and data conversion, and different techniques are used for each of these operations, depending on the type of application that the data mining operation is to perform.

#### 4.4 | Investigation and Evaluation of Effective Features of the Model

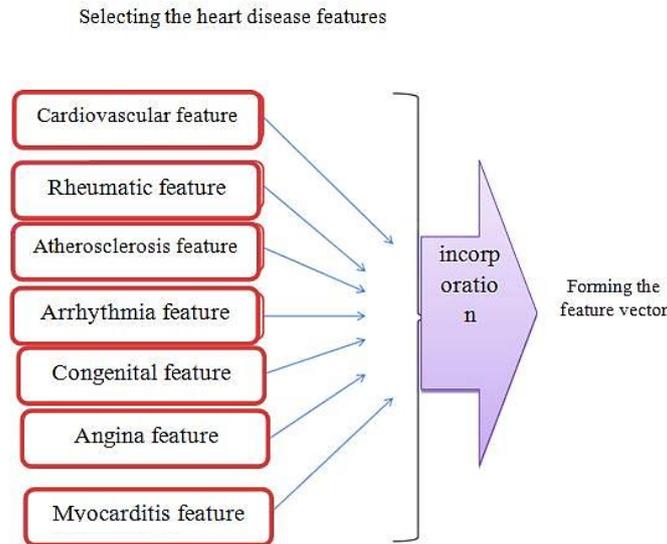
One of the most effective factors in improving the performance of data mining algorithms is the selection of appropriate features. In fact, choosing the right feature can be considered the main pillar of data mining. Many features can be used to build an intelligent model for diagnosing the presence or absence of heart disease. *Table 1* discusses the effective features in detecting heart disease, which are described below:

**Table 1. Introduction of all effective features in model construction.**

Domain	Adjective	Domain	Adjective
[71,202]	Maximum heart rate.	[29,77]	Age
[0,1]	Exercise-induced angina.	[0,1]	Gender
[0,62]	s.t. created when testing rest-dependent exercise.	[1,4]	Chest pain
[1,3]	The slope of the st piece at the time of maximum exercise.	[94,200]	Blood pressure at rest
[0,3]	The number of veins seen on fluoroscopy.	[126,564]	Cholesterol
[3,7]	Thallium scan.	[0,1]	Fasting blood sugar
{1,2}	Class (tag) 1 = no 2 = yes.	[0,2]	Resting ECG results

## 5 | Dataset

This study used the Cleveland Database, which is about heart disease, on the UCI website. The dataset includes 76 features, 14 of which are related to heart disease. The number of samples (records) is 303 and they were defined in 1988. *Fig. 1* shows some examples of these features.



**Fig. 1. Vector feature of the selection.**

The steps for building a decision tree include three steps:

**Determining the type of branch.** In each node, depending on the type of data, the branch can be binary with multiple. For attribute values that are continuous, it is most often a binary bifurcation type, and if the attribute values are discrete, the bifurcation may be multiple or binary.

**Select the best feature.** To select the best feature in the production of each node, you must pay attention to the purity and distribution of records based on classes. The bifurcation that increases the purity of the data with the greatest amount is the best bifurcation. To select the appropriate property for the node,  $Eg(1)$  is used, in which  $P_j$  is a fraction of the records labeled class  $j$  in node  $t$ , and  $m$  represents the number of classes.

**A. Ginny criterion.**

$$Gini(t) = 1 - \sum_{j=1}^m p_j^2 \tag{1}$$

**B. entropy.**

$$Entropy(t) = - \sum_{j=1}^m p(j|t) \log p(j|t) \tag{2}$$

**C. Categorization error.**

$$Error(t) = 1 - \max(p|t) \tag{3}$$

**Determining the stop time.** Determining the stop condition is one of the most important issues in recursive algorithms. The following modes are used to determine when the tree stops growing:

- I. All training records are in line with one class.
- II. Reach the maximum allowable depth
- III. The number of records in the current node is less than the threshold value.
- IV. The selection criterion is less than the threshold value.

A description of the decision tree for detecting heart disease using the features listed below is given.

Thal > 4.500

| ChestPainType > 1.500

| | MajorVessels > 0.500: 2.0 {2.0=59, 1.0=6}

| | MajorVessels ≤ 0.500

| | | RestBloodPressure > 109

| | | | Oldpeak > 24: 2.0 {2.0=6, 1.0=0}

| | | | Oldpeak ≤ 24

| | | | | age > 40.500

| | | | | | SerumCholesterol > 266.500

| | | | | | | ExerciseInduced > 0.500: 2.0 {2.0=9, 1.0=0}

| | | | | | | ExerciseInduced ≤ 0.500: 1.0 {2.0=1, 1.0=2}

| | | | | | | SerumCholesterol ≤ 266.500

| | | | | | | | age > 51: 1.0 {2.0=1, 1.0=10}

| | | | | | | | age ≤ 51

| | | | | | | | | age > 45: 2.0 {2.0=3, 1.0=0}

| | | | | | | | | age ≤ 45

| | | | | | | | | | age > 41.500: 1.0 {2.0=0, 1.0=3}

| | | | | | | | | | age ≤ 41.500: 2.0 {2.0=1, 1.0=1}

| | | | | | | | | | age ≤ 40.500: 2.0 {2.0=5, 1.0=0}

| | | | | | | | | | RestBloodPressure ≤ 109: 1.0 {2.0=0, 1.0=3}

| | | | | | | | | | ChestPainType ≤ 1.500: 1.0 {2.0=2, 1.0=6}

Thal ≤ 4.500

| Oldpeak > 35.500: 2.0 {2.0=3, 1.0=0}

| Oldpeak ≤ 35.500

| | MajorVessels > 2.500: 2.0 {2.0=5, 1.0=1}

| | MajorVessels ≤ 2.500

- | | | RestBloodPressure > 167: 2.0 {2.0=2, 1.0=1}
- | | | RestBloodPressure ≤ 167: 1.0 {2.0=23, 1.0=117}

## 6 | Evaluation Criteria

Different evaluation criteria are used in different issues, but some evaluation criteria are used in many studies as they are standard. The evaluation criteria are as follows:

- *True Positive (TP): The disease is classified correctly as the disease.*
- *False-positive (FP): The disease is classified incorrectly as similar cases.*
- *True Negative (TN): Similar cases are classified correctly as the similar cases.*
- *False Negative (FN): Similar cases are classified incorrectly as a disease.*
- *Precision criterion: This criterion is calculated by the following equation. In this equation, TP is the number of data that is correctly detected and FP is the number of data that is detected incorrectly as positive.*

$$\text{precision} = \frac{Tp}{TP + FP}. \quad (4)$$

Recall criterion the following equation is used to calculate it. In this equation, TP is the number of data that is detected correctly as positive and FN is the number of data that is detected incorrectly as negative.

$$\text{Recall} = \frac{Tp}{TP + FN}. \quad (5)$$

**F-measure criterion.** This criterion is calculated from the following equation, which is a criterion between recall and precision.

$$F1 = \frac{2 * \text{Re call} * \text{precision}}{\text{RECALL} + \text{PRECISION}}. \quad (6)$$

## 7 | Accuracy Criteria (ACC)

It is the ratio of the sensitivity criterion to the precision criterion. It is calculated as the accuracy criterion of the algorithm.

$$\frac{TP + TN}{TP + FP + FN + TN}. \quad (7)$$

## 8 | Experiments

In this section, the numerical results of the proposed method are presented in comparison with collinear algorithm algorithms. This comparison has been made for examples in different situations and challenges. MATLAB software has been used to implement the proposed method.

### 8.1 | Experiment 1: Algorithm Results

To demonstrate the accuracy, sensitivity, and feature of the classifier, the proposed algorithm is examined. Thus, the results of the algorithm for seven types of heart disease have been calculated separately:

- Atherosclerosis,
- Cardiovascular disease,
- Rheumatic heart disease,
- Congenital heart disease,
- Myocarditis,
- Angina,
- Arrhythmia.

The results of the classification are shown in *Table 2*. This table of the proposed method has high accuracy and sensitivity.

**Table 2. List of Arc lengths.**

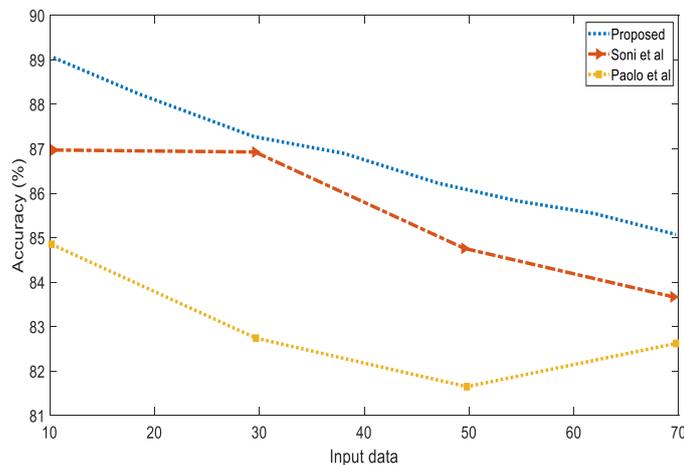
Disease	Precion	Criterion	
		Recall	Accuracy
Atherosclerosis	100	100	
Cardiovascular disease	100	100	
Rheumatic heart disease	97.22	100	
Congenital heart disease	100	92.59	98.96
Myocarditis	96.15	100	
Angina	100	100	
Arrhythmia	98.46	99.24	

## 8.2 | Experiment 2: Comparison of Results

Soni et al. [2] methods have been used to compare and evaluate the proposed method with other methods. The researchers used K-means clustering techniques and the Markov chain to classify heart disease, respectively.

### 8.2.1 | Comparison of average F-score

Accuracy is defined as the number of correct directions of the class to the total number of input data. This test has been repeated 10 times with different data numbers, and each time the accuracy has been calculated separately. This criterion shows that up to what percent the system has been able to correctly diagnose the disease in the data and non-disease, and the ratio is the sensitivity criterion to the accuracy criterion calculated as the accuracy criterion of the proposed algorithm. And the result of comparing the three methods can be seen in *Fig.2*.



**Fig. 2. The comparison of accuracy of proposed method with other works.**

In the proposed method, the accuracy of heart disease classification in 70 heart data is about 0.81%, and if we compare this value with the value obtained from other methods, we will see improvement in this method.

### 8.2.2 | Comparison of average F-score

Proper data and proper pre-processing and proper data mining methods provide good results in this regard. The more accurate the input data, the more accurate the output of our work. So from the wrong input, of course, we will have the wrong output as well. The better we know about the data, the more it will help us enter the right data into the model structure and therefore the more appropriate output. By this criterion, the proposed method performed better than the previous two methods in classifying heart disease in the data, and the average F-score of each of the methods can be seen in *Fig. 3*.

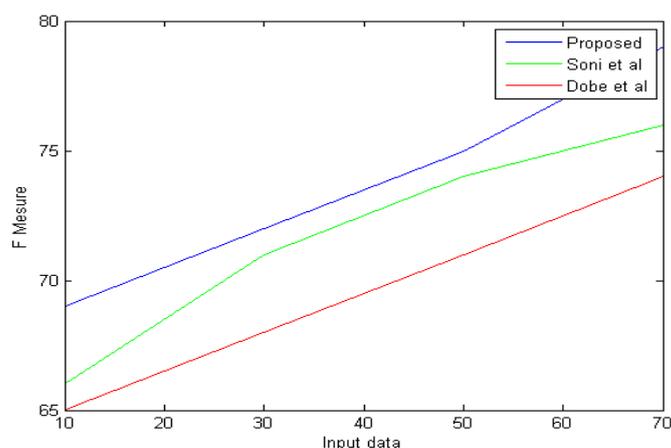


Fig. 3. Average F-score.

### 8.2.3 | Comparison of precision criterion

This criterion is one of the most important criteria for accuracy in heart disease classification algorithms in data. The precision criterion means a ratio of the negatives that the experiment marks correctly as negative. In mathematical terms, precision is the result of dividing TN by the sum of TN and false positives. The simulation result based on the above criterion is shown in *Fig. 4*.

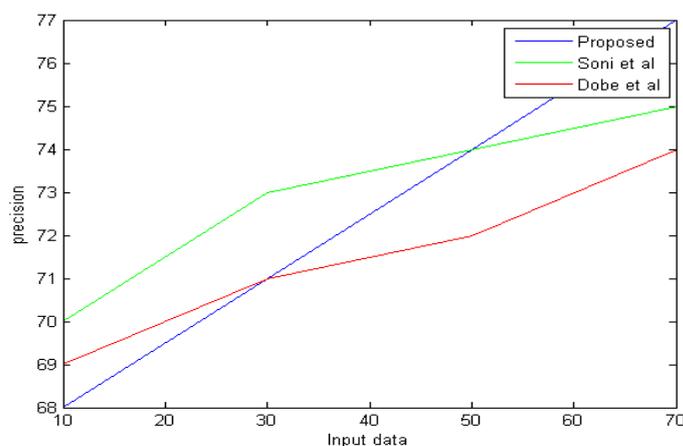


Fig. 4. Comparison of precision criterion.

In the first two steps, simulation of the proposed method is weaker than the previous two methods, but with the increase in the number of data, the proposed method shows good performance in classifying heart disease in the data. According to the proposed method diagram, about 77% have been efficient in

terms of precision criteria. The disadvantage of this method is the low number of data, i.e. 30, 10 and 50, which has less precision than the other two methods.

#### 8.2.4 | Comparison of precision criterion

The next criterion in classifying heart disease in data is the Recall criterion. In other words, Recall is the result of dividing TP by the sum of real positives and FN. Some cases are briefly presented. *Fig. 5* shows the simulation results for this criterion.

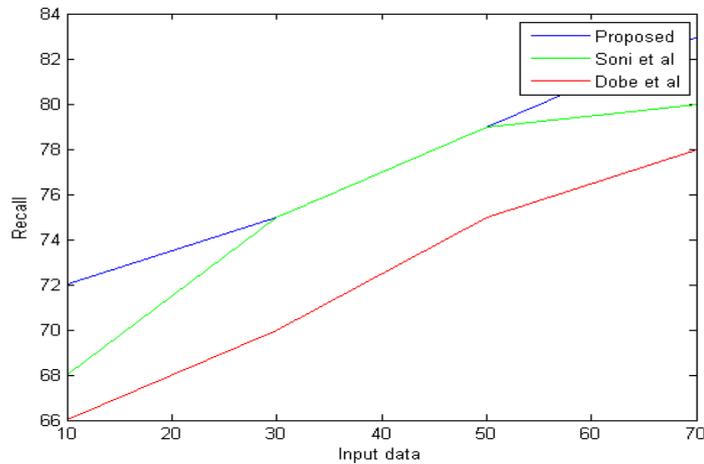


Fig. 5. Comparison of Recall criterion.

In this figure, 72, 75, 79 and 83% of the sensitivity of the proposed method have been obtained for 10 to 70 data, respectively. Based on the results obtained, the proposed method performed on average 3% better than the two similar methods. Also, with the increase in the number of data and inputs, the proposed method has a high sensitivity in classifying heart disease in the data.

## 9 | Conclusion

In recent years, there have been many important applications whose true datasets are causal data. When data mining techniques are considered for these causal databases, it can yield high quality results. Therefore, knowledge extraction from these data is important and one of the most important research topics conducted on this causal data is to extract frequent patterns. The extraction of frequent patterns is a very important field of research in data mining with a wide range of applications. Therefore, since its introduction, it has been the subject of numerous studies, and since databases often face over-elimination or change of transactions in many applications, frequent patterns extracted from them should be updated. In this study, an efficient approach to the diagnosis of heart disease based on abnormal features based on frequent pattern and the b-mine algorithm is presented. When making an association decision-making system, the type of structure and data mining parameters are the most important components. In this study, the b-mine algorithm along with decision tree in the structure of the association rules for data mining was used. The proposed method has a higher accuracy compared to other methods such as the Markov chain. The first advantage is that it has a running time of 3.12 seconds on a home computer. Extracting dependence association rules is one of the most important knowledge extractions in data that can save a lot of money and time. We intended to provide a method to do this. In addition to extracting association rules, we provided a model for predicting future association rules, i.e., the next data and their rules are predicted with the highest accuracy according to the recorded data and the extracted rules.

## Conflicts of Interest

All co-authors have seen and agree with the contents of the manuscript and there is no financial interest to report. We certify that the submission is an original work and is not under review at any other publication.



## References

- [1] Shouman, M., Turner, T., & Stocker, R. (2012, March). Using data mining techniques in heart disease diagnosis and treatment. *2012 Japan-Egypt conference on electronics, communications and computers* (pp. 173-177). IEEE. DOI: [10.1109/JEC-ECC.2012.6186978](https://doi.org/10.1109/JEC-ECC.2012.6186978)
- [2] Soni, J., Ansari, U., Sharma, D., & Soni, S. (2011). Predictive data mining for medical diagnosis: an overview of heart disease prediction. *International journal of computer applications*, 17(8), 43-48.
- [3] Latha, C. B. C., & Jeeva, S. C. (2019). Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. *Informatics in medicine unlocked*, 16, 100203. <https://doi.org/10.1016/j.imu.2019.100203>
- [4] Berka, P., Rauch, J., & Zighed, D. A. (Eds.). (2009). *Data mining and medical knowledge management: cases and applications: cases and applications*. IGI Global. DOI: [10.4018/978-1-60566-218-3](https://doi.org/10.4018/978-1-60566-218-3)
- [5] Maji, S., & Arora, S. (2019). Decision tree algorithms for prediction of heart disease. *Information and communication technology for competitive strategies* (pp. 447-454). Springer, Singapore. [https://doi.org/10.1007/978-981-13-0586-3\\_45](https://doi.org/10.1007/978-981-13-0586-3_45)
- [6] Cios, K. J. (2000). From the guest editor medical data mining and knowledge discovery. *IEEE engineering in medicine and biology magazine*, 19(4), 15-16. DOI: [10.1109/MEMB.2000.853477](https://doi.org/10.1109/MEMB.2000.853477)
- [7] Saraçoğlu, R. (2012). Hidden Markov model-based classification of heart valve disease with PCA for dimension reduction. *Engineering applications of artificial intelligence*, 25(7), 1523-1528. <https://doi.org/10.1016/j.engappai.2012.07.005>
- [8] Agrawal, R., & Srikant, R. (1994, September). Fast algorithms for mining association rules. *Proc. 20th int. conf. very large data bases, VLDB* (Vol. 1215, pp. 487-499).
- [9] Ordonez, C. (2004). Improving heart disease prediction using constrained association rules. In *Seminar presentation at university of Tokyo* (Vol. 4).
- [10] Yan, H., Jiang, Y., Zheng, J., Peng, C., & Li, Q. (2006). A multilayer perceptron-based medical decision support system for heart disease diagnosis. *Expert systems with applications*, 30(2), 272-281. <https://doi.org/10.1016/j.eswa.2005.07.022>
- [11] Domadiya, N., & Rao, U. P. (2019). Privacy preserving distributed association rule mining approach on vertically partitioned healthcare data. *Procedia computer science*, 148, 303-312. <https://doi.org/10.1016/j.procs.2019.01.023>
- [12] Amin, M. S., Chiam, Y. K., & Varathan, K. D. (2019). Identification of significant features and data mining techniques in predicting heart disease. *Telematics and informatics*, 36, 82-93. <https://doi.org/10.1016/j.tele.2018.11.007>
- [13] De Cnudde, S., Martens, D., Evgeniou, T., & Provost, F. (2020). A benchmarking study of classification techniques for behavioral data. *International journal of data science and analytics*, 9(2), 131-173. <https://doi.org/10.1007/s41060-019-00185-1>
- [14] Gupta, A., Kumar, R., Arora, H. S., & Raman, B. (2019). MIFH: a machine intelligence framework for heart disease diagnosis. *IEEE access*, 8, 14659-14674. DOI: [10.1109/ACCESS.2019.2962755](https://doi.org/10.1109/ACCESS.2019.2962755)
- [15] Kirankumar, V., Ramasubbareddy, S., Kannayaram, G., & Nikhil Kumar, K. (2019). Classification of heart disease using support vector machine. *Journal of computational and theoretical nanoscience*, 16(5-6), 2623-2627. DOI: <https://doi.org/10.1166/jctn.2019.7941>
- [16] Bajaj, P., & Gupta, P. (2014). Review on heart disease diagnosis based on data mining techniques. *International journal of science and research (IJSR)*, 3(5).
- [17] Jabbar, M. A. (2017). Prediction of heart disease using k-nearest neighbor and particle swarm optimization. *Biomed. Res*, 28(9), 4154-4158. <https://www.alliedacademies.org/articles/prediction-of-heart-disease-using-knearest-neighbor-and-particle-swarm-optimization.html>

- [18] Ani, R., Augustine, A., Akhil, N. C., & Deepa, O. S. (2016). Random forest ensemble classifier to predict the coronary heart disease using risk factors. *Proceedings of the international conference on soft computing systems* (pp. 701-710). Springer, New Delhi. [https://doi.org/10.1007/978-81-322-2671-0\\_66](https://doi.org/10.1007/978-81-322-2671-0_66)
- [19] Feshki, M. G., & Shijani, O. S. (2016, April). Improving the heart disease diagnosis by evolutionary algorithm of PSO and Feed Forward Neural Network. *2016 artificial intelligence and robotics (IRANOPEN)* (pp. 48-53). IEEE. DOI: [10.1109/RIOS.2016.7529489](https://doi.org/10.1109/RIOS.2016.7529489)
- [20] Alkeshuosh, A. H., Moghadam, M. Z., Al Mansoori, I., & Abdar, M. (2017, September). Using PSO algorithm for producing best rules in diagnosis of heart disease. *2017 international conference on computer and applications (ICCA)* (pp. 306-311). IEEE. DOI: [10.1109/COMAPP.2017.8079784](https://doi.org/10.1109/COMAPP.2017.8079784)
- [21] Jabbar, M. A., Deekshatulu, B. L., & Chandra, P. (2016). Prediction of heart disease using random forest and feature subset selection. In *Innovations in bio-inspired computing and applications* (pp. 187-196). Springer, Cham. [https://doi.org/10.1007/978-3-319-28031-8\\_16](https://doi.org/10.1007/978-3-319-28031-8_16)
- [22] Chauhan, R., Jangade, R., & Rekapally, R. (2018). Classification model for prediction of heart disease. In *Soft computing: theories and applications* (pp. 707-714). Springer, Singapore. [https://doi.org/10.1007/978-981-10-5699-4\\_67](https://doi.org/10.1007/978-981-10-5699-4_67)
- [23] Giudici, P., & Castelo, R. (2003). Improving Markov chain Monte Carlo model search for data mining. *Machine learning*, 50(1), 127-158. <https://doi.org/10.1023/A:1020202028934>
- [24] Polaraju, K., & Prasad, D. D. (2017). Prediction of heart disease using multiple linear regression model. *International journal of engineering development and research*, 5(4), 2321-9939.
- [25] Sultana, M., & Haider, A. (2017, March). Heart disease prediction using WEKA tool and 10-Fold cross-validation. In *The institute of electrical and electronics engineers* (pp. 17-33).
- [26] Deepika, K., & Seema, S. (2016, July). Predictive analytics to prevent and control chronic diseases. *2016 2nd international conference on applied and theoretical computing and communication technology (iCATccT)* (pp. 381-386). IEEE. DOI: [10.1109/ICATCCT.2016.7912028](https://doi.org/10.1109/ICATCCT.2016.7912028)
- [27] Beyene, C., & Kamat, P. (2018). Survey on prediction and analysis the occurrence of heart disease using data mining techniques. *International journal of pure and applied mathematics*, 118(8), 165-174.
- [28] Patil, P. B., Shastry, P. M., & Ashokumar, P. S. (2020). Machine learning based algorithm for risk prediction of cardio vascular disease (Cvd). *Journal of critical reviews*, 7(9), 836-844.
- [29] Gupta, A., Kumar, R., Arora, H. S., & Raman, B. (2019). MIFH: a machine intelligence framework for heart disease diagnosis. *IEEE access*, 8, 14659-14674. DOI: [10.1109/ACCESS.2019.2962755](https://doi.org/10.1109/ACCESS.2019.2962755)
- [30] Saxena, K., & Sharma, R. (2016). Efficient heart disease prediction system. *Procedia computer science*, 85, 962-969. <https://doi.org/10.1016/j.procs.2016.05.288>
- [31] Gomathi, K., & Priyaa, D. D. S. (2016). Multi disease prediction using data mining techniques. *International journal of system and software engineering*, 4(2), 12-14.
- [32] Reddy, M. P. S. C., Palagi, M. P., & Jaya, S. (2017). Heart disease prediction using ANN algorithm in data mining. *International journal of computer science and mobile computing*, 6(4), 168-172.
- [33] Ramotra, A. K., Mahajan, A., Kumar, R., & Mansotra, V. (2020). Comparative analysis of data mining classification techniques for prediction of heart disease using the weka and SPSS modeler tools. In *Smart trends in computing and communications* (pp. 89-97). Springer, Singapore. [https://doi.org/10.1007/978-981-15-0077-0\\_10](https://doi.org/10.1007/978-981-15-0077-0_10)
- [34] Aldhyani, T. H., Alshebami, A. S., & Alzahrani, M. Y. (2020). Soft clustering for enhancing the diagnosis of chronic diseases over machine learning algorithms. *Journal of healthcare engineering*. <https://doi.org/10.1155/2020/4984967>
- [35] Bahrami, B., & Shirvani, M. H. (2015). Prediction and diagnosis of heart disease by data mining techniques. *Journal of multidisciplinary engineering science and technology (JMEST)*, 2(2), 164-168.