Journal of Diabetes & Metabolic Disorders (2021) 20:2055–2071
https://doi.org/10.1007/s40200-021-00884-2

COMMENTARY

A survey on data mining techniques used in medicine

Saba Maleki Birjandi[1] · **Seyed Hossein Khasteh**[1,2]

Received: 10 May 2021 / Accepted: 22 August 2021 / Published online: 31 August 2021

Abstract
Data mining is the process of analyzing a massive amount of data to identify meaningful patterns and detect relations, which can lead to future trend prediction and appropriate decision making. Data mining applications are significant in marketing, banking, medicine, etc. In this paper, we present an overview of data mining applications in medicine to provide a clear view of the challenges and previous works in this area for researchers. Data mining techniques such as Decision Tree, Random Forest, K-means Clustering, Support Vector Machine, Logistic Regression, Neural Network, Naive Bayes, and association rule mining are used for diagnosing, prognosis, classifying, constructing predictive models, and analyzing risk factors of various diseases. The main objective of the paper is to analyze and compare different data mining techniques used in the medical applications. We present a summary of the results and provide comparison analysis of the data mining methods employed by the reviewed articles.

Keywords Data mining · Statistical methods · Decision tree · Linear regression · Association rules · Medical data mining

Introduction

With the increasing rate of data generation in various areas of human life, knowledge extraction from the generated data is strongly needed. Data mining is the process of discovering patterns within the data collected in a specific domain and finding associations between attributes of data samples based on the computer science and statistical techniques. An outstanding data mining capability is to handle massive amount of data. Data mining is a step in Knowledge Discovery in Database (KDD) which consists of data selection, data preprocessing, data transformation, data mining, interpretation or evaluation of the model and using the discovered knowledge [1]. Data mining applications include classification, clustering, prediction, and finding associations. Banking, medicine, marketing, and transportation are some areas in which data mining has been widely used.

✉ Seyed Hossein Khasteh
khasteh@kntu.ac.ir

Saba Maleki Birjandi
saba_maleki@email.kntu.ac.ir

[1] School of Computer Engineering, K. N. Toosi University of Technology, 16317-14191 Tehran, Iran

[2] Faculty of Computer Engineering, Seyed Khandan, Shariati Ave, Tehran, Iran

Recently, the use of data mining methods has increased in medicine. Access to medical data is difficult due to confidentiality of patients' records and the importance of critical individual information. However, the huge amount of hospital and clinical data has a great potential for discovering relations and hidden patterns. Using data mining methods reduces time and cost in prognosis and diagnosis of diseases [2]. It also has a special role in the treatment plan and medical decisions and helps construct accurate and reliable models.

In enormous sets of clinical data, the medical big-data mining is an appropriate method for knowledge discovery. Bag of Words technique for the automatic classification of the medical document data is helpful in disease diagnoses and making clinical decisions [3]. In addition to prognosis and diagnosis of diseases, data mining is used in other medical aspects like gene prioritization, gene function prediction, drug repositioning, pharmacogenomics, and toxicology [4]. In general, data mining methods are suitable for identifying risk factors, complications, therapies, health policies and decisions, genetic effects, and environmental effects of diseases [5].

We intend to clarify the path for the researchers to deal with challenges in medical data mining. In this article, we reviewed the various methods adopted in medical data mining and present a report on the results of previous works.

This study helps researchers to find the unresolved issues in the field of medical data mining and endeavor to find the solution.

The rest of this article is organized as follows. Section 2 presents the comparison between data mining and statistical methods employed in medical applications. Section 3 discusses various applications and models of data mining. In Sect. 4, the data mining methods used in medical area are surveyed and their results are compared in different applications. Finally, we present the conclusion in Sect. 5.

Statistics vs. data mining

There is a close relation between statistical and data mining techniques. In some studies, only statistical methods are used to tackle the problem. First, we survey these articles.

YoussefAgha et al. studied 657,068 students to analyze their Body Mass Index (BMI) to discover hidden patterns. They ran the chi-square test to determine three states in patients, namely normal, overweight, and obese. The study showed that the BMI is greater in elementary students than among middle and high school students [6].

Hosseini et al. applied ANOVA on more than 14,000 children and adolescents aged 7- 18 years old. They attempted to investigate the effect of BMI on systolic and diastolic blood pressure. They showed that the average of systolic and diastolic blood pressure of a patient is related to their BMI and age [7].

In general, data mining methods are suitable for large data sets while statistical methods are appropriate for smaller data sets. The statistical methods are adopted to formalize the relationship in data but data mining algorithms learn from data without using any programming during the learning. Unlike statistics, Data mining is not concerned with discovering the best way to collect data, rather it focuses on extracting knowledge. The statistics operate according to mathematical formulas and concepts, making it easier to understand and interpret the results [8]. In the problems which are suitable for the statistical methods the computational time of machine learning and data mining methods is much greater compared to statistical methods [9].

Statistical and data mining methods have their own disadvantages and advantages for different purposes. Since the focus of our article is on the data mining methods, we do not elaborate on the statistical methods.

Data mining, an overview

Statistics and machine learning are the two most important bases for data mining. Statistics emphasize on mathematics and formalization of associations in data. In contrast, machine learning evaluates various models without trying to prove the effectiveness of solutions [10].

Data mining models are classified into two categories: 1) Predictive models, and 2) Descriptive models. The predictive models attempt to identify hidden patterns among the data and predict the future trends [10]. Supervised learning is mostly employed in predictive models in the tasks of classification, regression, time-series analysis, and prediction [11].

The descriptive models attempt to interpret the patterns within the data for domain experts and often use the unsupervised learning techniques [10].

In unsupervised learning, the class label of each sample is unknown. Moreover, the number of classes to be learned may not be known in advance. Unsupervised learning includes clustering, summarization, association rules, and frequent rules tasks [11].

Some of the data mining methods we investigate here include Decision Tree, Random Forest, K-means Clustering, Support Vector Machine, Logistic Regression, Neural Networks, Naive Bayes and Association Rule Mining.

To classify and construct predictive models that are the main subject of most medical data mining articles, data is divided into two subsets. The large subset is called the training set used for constructing the classifier. The training set is made up of data set samples and their related class labels. The small subset is called the test set utilized for evaluating the classifier. Similar to the training set, the test set consists of data set records but the test set is independent of the training set, so it is not used for building the classifier. Generally, 70% of the data set is the training set and 30% is the test set [11].

To avoid overfitting, k-fold cross validation is applied. In this method, the original data set is divided randomly into k partitions. For the training set, k-1 of partitions is used and one remaining is employed for the test set. Cross-validation is performed k times with a different partitioning of the data set and the results are averaged [12].

Materials and Methods

In this section, the effective data mining methods are introduced in various fields especially in medicine. These methods are widely adopted in clinical issues. According to previous works, it can be mentioned that the constructing disease predictive models, extraction and analyzing disease risk factors, and classification and clustering the related features of diseases are the significant applications of data mining methods in medicine. As common data mining applications, disease predictive models are applied for many diseases like cancers, diabetes, stroke, heart diseases, etc. [12].

There are several predictive models constructed by using different data mining methods. To compare these models, we must evaluate their performance in different tasks.

To evaluate the performance of these models, the statistical metrics are used which are comprised of accuracy, sensitivity, specificity, the area under the ROC curve (AUC), positive predictive value (PPV) and negative predictive value (NPV). These metrics are calculated using the following equations:

$$Accuracy = \frac{TP + TN}{(TP + TN + FP + FN)} \quad (1)$$

$$Sensitivity = \frac{TP}{(TP + FN)} \quad (2)$$

$$Specificity = \frac{TN}{(TN + FP)} \quad (3)$$

$$PPV = \frac{TP}{(TP + FP)} \quad (4)$$

$$NPV = \frac{TN}{(TN + FN)} \quad (5)$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives and FN is the number of false negatives [13].

The most common data mining methods along with their usage in medical applications are detailed in this section.

Decision tree (DT)

The decision tree is one of the popular supervised machine learning methods used for classification, prediction and regression [14]. A decision tree has a tree structure and represents simple if–then rules which improve the interpretability of the domain expert knowledge. It consists of some nodes and branches. The internal nods denote a test on an attribute, with each branch corresponding to one of the possible values (class) for this attribute, and each leaf node holding a class label [11]. This method can also handle both categorical and numerical data. The Decision tree determines the best feature split by using the attribute selection measures, separates the data into two subsets and continues until all samples are classified [11].

Decision tree learning is applied to problems such as learning to classify patients by their disease and cope with some challenges like errors or missing values in the training sets [12]. It is also faster when dealing with large data with low dimensionality.

ID3, C4.5, and CART are included in the family of decision tree learning algorithms widely mentioned in various studies [11].

Shaikhina et al. attempted to predict antibody incompatible kidney transplantation within 9 years of incompatible renal allografts from 14 clinical indications based on 80 samples [15]. The authors used the DT model based on the standard CART algorithm. The split optimization criterion employed in this research was Gini's Diversity Index (GDI), which is a measure of node impurity. The experiment was repeated 600 times applying the pruning technique which can reduce the size of a DT and prevents overtraining, in order to penalize complexity of the DT and ensuring only the most significant splits to be discovered by the model. As a result, the DT model had 81.8% sensitivity and 88.90% specificity, AUC = 0.849 on training samples and AUC = 0.854 for the DT predictions on the test samples, the results are compared with random forest results presented in Table 1. Furthermore, the 6 key predictors identified by the DT were confirmed by previous works.

Tayefi et al. used a decision tree based on the CART algorithm to identify the associated risk factors for coronary artery disease from 14 clinical indications. The study was based on a dataset of 2346 individuals including 1159 healthy subjects and 1187 patients who had undergone angiography [405 participants with angiography (-), 782 participants with angiography (+)].

The target variable consisted of 3 classes as healthy, negative angiography and positive angiography.

Data were divided into training and test groups, 70% of total participants (1640 cases) were randomly selected to make the training group for constructing the decision tree. The remaining 30% (706 cases) were considered as a test group to evaluate the performance of the decision tree. In this study, the root node is divided the whole population with the highest information gain.

$$IG = -\sum_{c \in classes} P(c) \log (P(c)) \quad (6)$$

where P(c) represents the probability of a sample in class c.

The CART algorithm uses the Gini impurity index for selecting the best variable.

$$Gini(D) = 1 - \sum_{i=1}^{m} p_i^2 \quad (7)$$

where p_i is the probability that a record in D belongs to class C_i [11, 21].

Amongst 11 traditional related factors of coronary heart disease, 10 variables finally entered the decision tree model. This study indicated that hs-CRP as a new biomarker is strongly associated with coronary heart disease and used it in the root node. Also, the decision tree had an accuracy of 94%. The specificity and sensitivity of this tree were 87% and 96%, respectively with AUC = 95.4%.

In their study carried out in two phases, Chen et al. used data mining to explore risk factors of preterm birth based on 455 samples [22]. First, mining 15 most important factors from original factors. Second, a C5.0 decision tree which is

Table 1 Data mining techniques in medicine

Data mining techniques in medicine

Author	Subject	Methods								Description
			DT			RF				The 6 key predictors identified by the DT are confirmed by previous models Our DT and RF are equally well-equipped to handle partially missing data and managed to classify Classes
Shaikhina et al. [15] (2017)	Prediction of antibody incompatible kidney transplantation		train	test		train			test	
		C (%)	85.0	85.0		91.7			85.0	
		Sen (%)	85.7	81.8		93.9			92.3	
		Spe (%)	84.0	88.9		88.9			71.4	
		PPV (%)	88.2	90.0		91.2			85.7	
		NPV (%)	80.8	80.0		92.3			83.3	
		AUC	0.849	0.854		0.914			0.819	
		Acc (%)	N/A	N/A		N/A			N/A	
Meng et al. [16] (2013)	Prediction of diabetes or prediabetes		Acc (%)	Sen (%)		Spe (%)			AUC	The C5.0 decision tree model performed best on classification accuracy
		DT	77.87	80.68		75.13			N/A	
		Logistic regression	76.13	79.59		72.74			N/A	
		NN	73.23	82.18		64.49			N/A	
Tesfaye et al. [17] (2017)	Prediction of child mortality		Sen (%)	Spe (%)	AUC (%)	PPV (%)		NPV (%)	Acc (%)	Variables withthe p-value of 0.05 level and 95% CI Sub-model C that applies decision tree algorithm with pruning has a better result
		DT (J48)	94.3	92.4	94.8	92.2		94.5	N/A	
Acharya et al. [18] (2016)	Classification of breast lesion		Acc (%)	Sen (%)		Spe (%)			AUC (%)	A neural network with 4 layers. (Probabilistic neural network) tenfold cross-validation was used for model performance evaluation
		DT	88.46	84.93		91.57			N/A	
		SVM poly 1	92.95	87.67		97.59			N/A	
		SVM poly 2	91.03	84.93		96.39			N/A	
		SVM ploy 3	92.31	90.41		93.98			N/A	
		SVM RBF	92.95	90.41		95.18			N/A	
		PNN	91.67	83.56		98.80			N/A	

Table 1 (continued)

Data mining techniques in medicine							
Amin et al. [19] (2019)	Prediction heart disease		Acc (%)	Sen (%)	Spe (%)	AUC (%)	Vote (a hybrid technique with Naive Bayes and Logistic Regression) achieved accuracy = 87.41% obtained with identified 9 significant features
		Vote	78.20	N/A	N/A	N/A	
		Naive Bayes	78.20	N/A	N/A	N/A	
		SVM	78.15	N/A	N/A	N/A	
		LR	78.03	N/A	N/A	N/A	
		NN	75.18	N/A	N/A	N/A	
		DT	63.50	N/A	N/A	N/A	
Maji et al. [20] (2019)	Prediction of heart disease		Acc (%)	Sen(%)	Spe(%)	AUC (%)	The hybrid DT promotes better results for prediction of heart disease
		ANN	77.40	77.40	21.7	N/A	
		C4.5	76.66	76.7	24	N/A	
		Hybrid-DT	78.14	78	22.9	N/A	
Tayefi et al. [21] (2017)	Prediction of coronary heart disease (CHD)		Acc (95%CI)	Sen (95%CI)	Spe (95%CI)	AUC (95%CI)	The 15 rules extracted through decision tree It was indicated that hs-CRP as a new biomarker is strongly associated with CHD
		DT (CART)	95.3	97.8	92.9	95.4	
Chen et al. [22] (2011)	Extraction the risk factors of preterm birth		Description				Multiple birth and hemorrhage during pregnancy are the top two risk factors
		DT	Classify the top 15 risk factors of preterm birth and extract rules with an accuracy of 80% or more				
		NN	Explore the top 15 risk factors of preterm birth with coefficients larger than 0.0300				
Khalilia et al. [23] (2011)	Prediction of diseases	**AUC measure for RF and SVM**					They listed the top four most important variable for the eight disease categories RF had a better result, the average AUC obtained across all disease was about 89.05% In general, the RF result with sub-sampling was improved
				RF		SVM	
		Breast cancer		0.9123		0.9063	
		Diabetes no complication		0.8791		0.8417	
		Diabetes with/complication		0.94317		0.9239	
		Hypertension		0.9003		0.8592	
		Coronary Atherosclerosis		0.9199		0.8973	
		Peripheral Atherosclerosis		0.9095		0.8972	
		Other Circulatory Diseases		0.7899		0.7591	
		Osteoporosis		0.87		0.867	

Table 1 (continued)

Data mining techniques in medicine								
Ahamad et al. [24] (2016)	Analysis of risk factors in Diabetes, Hypertension and Obesity	K-means		Clusters			SSE	Each cluster represented particular disease or the aggregation of obesity, hypertension and diabetes
				10			278. 57	
Wu et al. [25] (2018)	type 2 diabetes mellitus prediction model		Acc (%)	Sen (%)	Spe (%)		AUC (%)	The model attained a 3.04% higher accuracy of prediction
		K-means + LR	95.42	N/A	N/A		N/A	
		MLP	81.9	N/A	N/A		N/A	
		J48	76.7	N/A	N/A		N/A	
		LR	78.2	N/A	N/A		N/A	
		NB	74.9	N/A	N/A		N/A	
		CART	72.8	N/A	N/A		N/A	
Arslan et al. [26] (2016)	Prediction of Ischemic Stroke		Acc	Sen (%)	Spe (%)		AUC	The Acc and AUC values were with 95% CI SVM had the best predictive results and used radial basis function as the kernel function
		SVM	0.9789	N/A	N/A		0.9783	
		Penalized logistic regression	0.8947	N/A	N/A		0.8953	
Easton et al. [27] (2014)	Prediction of stroke mortality		1–7 day		8–93 days			Naïve Bayes was the best classifier
			Sen (%)	Spe (%)	Sen (%)		Spe (%)	
		Naïve Bayes	92.6	69.4	83.3		70.8	
		Logistic regression	88.9	59.9	81.5		51.0	
		DT	33.3	94.6	9.3		95.0	
Heydari et al. [28] (2012)	Classification of obesity		C (%)	Sen (%)	Spe (%)	AUC	Acc	Input of logistic regression: variables with P value less than 0.05 Input layer of neural network: 11 neurons The logistic regression and neural networks were similar in result
		Logistic regression	80.2	80.2	81.9	0.888	N/A	
		NN	81.2	79.7	83.7	0.884	N/A	

Table 1 (continued)

Data mining techniques in medicine							
Pochini et al. [29] (2014)	Analyzing risk factors of Overweight and Obesity	**Significant Variables for Overweight**					For preventing obesity students should have physical activity and eat breakfast, also avoid using tobacco
		Logistic Regression		DT			
		Breakfast		Breakfast			
		Physically Active		Fruit Juice			
		Breakfast		Sleep			
		Obesity results are available in the article					
Charlton et al. [30] (2014)	Analyzing risk factors of low fitness	Logistic Regression was used for cluster analyzing based on fitness and weight categories					3 main clusters: children at low risk (not obese, fit) , children 'visibly at risk' (overweight, unfit) and 'invisibly at risk' (unfit but not overweight)
Alizadehsani et al. [31] (2013)	Diagnosis of coronary artery disease		Method	Acc (%)	Sen (%)	Spe (%)	AUC	SVM was applied for feature selection Three features were created Selected features and three created features had best result
		Selected features	Naïve Bayes	75.51	67.59	95.40	N/A	
		and three created features	NN	88.11	91.20	80.46	N/A	
Adnan et al. [32] (2012)	Prediction of childhood obesity	Naive Bayes selected parameters and compared with the previously selected set					They used The new set increased the accuracy of childhood obesity prediction
Hossain et al. [33] (2018)	Prediction of obesity		Acc (%)	Sen (%)	Spe (%)		AUC	Naïve Bayes prediction model had the best prediction accuracy
		Naïve Bayes	98.4	N/A	N/A		N/A	
		Logistic Regression	99.2	N/A	N/A		N/A	

Table 1 (continued)

Data mining techniques in medicine					
Ilayaraja et al. [34] (2013)	Identifying frequent diseases	Apriori	Min-support (%) 0.35	Min-confidence(%) 0.9	Heart disease, Liver disease, Overweight, and smoking were most frequent among 29 diseases
Nahar et al. [35] (2013)	Analyzing risk factors of heart disease	AA Method Apriori Predictive Apriori Tertius		Requirement (%) confidence > 90 accuracy > 99 confirmation > 79	To select risk factors from rules, they considered the majority of factor appearance in rules and also the factors existed in rules with high confidence
Sharma et al. [36] (2017)	Mitigating the Increasing obesity	AA Fitness facility Living room Bedside	Average Support (%) 28.57 60.71 92.85	Average Confidence (%) 28.57 60.71 92.85	In the first two patterns, the number of exercise times has declined steadily, but the third pattern has been higher on average

Table 1 (continued)

Data mining techniques in medicine						
Ramezankhani et al. [37] (2015)	Analyzing risk factors of type 2 diabetes		AA			Rules extracted sample for females: IFG = yes, IGT = yes, BMI > = 30,waist to height > = 0.5 ⇒ Type 2 DM (support = 2.8, confidence = 75.0, lift = 6.6) Rules extracted sample for males: IGT = yes, IFG = yes, CHO to HDL ≥ 5.3, occupation status = employed, waist to hip > = 0.9 ⇒ Type 2 DM (support = 2.2, confidence = 65.1, lift = 6.2)
			Min- support (%)		Min-confidence (%)	
		Rules extracted for females using the Apriori	2		75	
		Rules extracted for males using the Apriori	1.8		65	
Salehnasab et al. [38] (2014)	Analyzing risk factors of high blood pressure	AA				The rules were analyzed from Microsoft BI and Clementine software
			Support (%)		Confidence (%)	
		Apriori	0.2		75	
Ordonez [39] (2006)	Prediction of heart disease		k	Min-support (%)	Min-confidence (%)	lift
		AA	4	1	70	1.2
		DT (C4.5)		Nodes		Acc (%)
		Phase 1		181		90
		Phase 2		83		77
		Phase 3		6		65

C correct classification, *Sen* sensitivity, *Spe* specificity, *PPV* positive predictive value, *NPV* negative predictive value, *AUC* area under the ROC curve, *Acc* accuracy, *CI* confidence intervals, *SSE* sum of squared error, *DT* decision tree, *RF* random forest, *NN* neural network, *SVM* support vector machine, *PNN* probabilistic neural network, *AA* association rule analysis, *DM* diabetes mellitus, $_k$ user-specified maximum item-set size, *CART* classification and regression tree

an improved version of the C4.5 and ID3 algorithms [40] was used for classifying risk factors related to preterm birth. The DT explored 17 rules to predict preterm birth, 10 of which had an accuracy of 80% or more.

Meng et al. compared some predictive models for predicting diabetes and prediabetes by using 12 risk factors [16]. In this study, the dataset consists of 1487 individuals. The output was a categorical variable of 0 and 1, where 0 means normal and 1 means diabetes or prediabetes. The input variables were 12 risk factors related to diabetes. In this study, they used the decision tree (C5.0) models and used entropy-based information gain as a measure to split data. As a result, the decision tree (C5.0) performed best among other tested methods with the accuracy, sensitivity, and specificity of C5.0 being 77.87%, 80.68%, and 75.13%, respectively.

Tesfaye et al. identified determinants of childhood mortality and developed a prediction model in Ethiopia [17]. A total of 11,654 records partitioned into "Alive" and "Dead" groups was used as a data set. In this study, J48 was deployed for constructing a prediction model, which is a simple C4.5 decision tree for classification. It can ignore missing values and create a binary tree. Breastfeeding, education of parents, birth interval, age of mother and low birth weight were the factors associated with child mortality. Different kinds of sub-models were developed in this study. The sub-model designed by applying decision tree algorithm with pruning had better sensitivity, AUC and Positive Predictive Value (PPV) among other sub-models. This sub-model had a 94.3% sensitivity, 92.4% specificity, 94.8% AUC, 92.2% PPV and 94.5% NPV.

Similar to the mentioned articles, there are other researches that used the decision tree as a method for prediction and classification of stroke mortality, hypertension and hyperlipidemia, childhood obesity, breast cancer and heart disease [18–20, 27, 41–43].

Also, decision tree is utilized for analyzing diseases risk factors like diabetes, obesity and low fitness [29, 30, 44].

Random forest (RF)

Random forest is an ensemble learning method for classification and regression. An ensemble method combines several base classifiers with the aim of creating a classification model. The ensemble declares a class prediction according to the votes of the base classifiers. It could be more accurate than a single classifier. In a random forest, each of the base classifiers is a decision tree, and the set of decision trees builds the forest [11].

The random forest creates forest mostly trained with the bootstrap aggregation like "bagging" method which combines learning models to increase the overall result. The input samples are passed down through each of the constituent decision trees and each of them votes for the corresponding class. RF prediction is the majority of the votes. These decision trees aggregation help to deal with noises and has more robustness in the face of outliers than a single decision tree output [11].

In addition, RF can be used for ranking the variable importance scores by measuring the increase in prediction error [11] and it is comparatively faster than other algorithms.

Khalilia et al. used the random forest and some other methods for disease prediction. In this study, they considered breast cancer, diabetes, hypertension, coronary atherosclerosis, peripheral atherosclerosis, other circulatory diseases, and osteoporosis to construct prediction models. One of the challenges in this research was dealing with the imbalanced data. They utilized the Healthcare Cost and Utilization Project (HCUP) dataset consisting of about 8 million records of clinical data with 262 features. To apply the random forest, the repeated random sub-sampling method was used for preparing training data. RF uses the Gini measure for splitting the data with each tree being responsible for a different set of variables to extract rules. RF is efficient to deal with missing values and can represent the variable importance. The average AUC across all disease was about 89.05% by RF which was better than previous works [23].

Furthermore, random forest was used for predicting antibody incompatible kidney transplantation [15].

K-means clustering

K-means clustering algorithms is one of the most common unsupervised machine learning algorithms. A cluster is a set of data samples which have certain similarities. The samples within a cluster are similar but dissimilar to samples in other clusters [11]. K-means clustering aims to split the whole data set to k clusters each having a centroid which is the mean value of samples in that cluster. The algorithm assigns samples to each cluster based on the Euclidean distance between the sample and centroid. The k-means algorithm iteratively improves the position of the centroids. For each cluster in the new iteration, it computes the new the centroid by calculating the mean value of samples within each cluster, and the iterations continue until the assignment is stable [11].

In classification, we want to know the correct label of at least a part of the data set, but these labels are unavailable in many cases or it is costly to achieve these labels. In contrast, clustering does not need any label and it can partition the whole data set to a meaningful subset without having any prior knowledge. However, it is computationally expensive for large datasets.

Ahamad et al. attempted to analyze risk factors of obesity, hypertension, and diabetes by using clustering data mining technique. They applied k-means for a data set of 1046

patients, each record consisting of 19 variables. To select the number of clusters, they used trial and error methods and decided to use 10 clusters for study. K-means helps for extracting interesting patterns of risk factors and each cluster represented a particular disease or the aggregation of obesity, hypertension, and diabetes [24].

Furthermore, k-means was employed in another study for analyzing risk factors of diabetes and clustering observation in type 2 diabetes mellitus prediction model [25, 44].

Support vector machine

A Support Vector Machine (SVM) is a supervised machine learning method for classification and regression analysis and novelty detection [14]. The objective of the support vector machine algorithm for two-classes problems is to find an optimal hyperplane with maximum distance to the closest samples of the two classes. A set of these closest samples to optimal hyperplane is called a support vector. This optimal hyperplane provides a linear classifier. Also, the support vector machine is used for non-linear problems. In such cases, non-linear kernels transform the original attribute space to a higher dimensional space [11].

Support Vector Machines are popular in medicine, especially in high-dimensional multivariate problems used specifically to deal with the problem of imbalanced data, than other classification algorithms. The disadvantage is that it is difficult to interpret the results by domain experts and the training is slow. Furthermore, its computational cost is prohibitively high, especially in high-dimensional data. However, for creating highly accurate models less prone to overfitting than other methods, SVM is one of the best choices [11].

Arslan et al. aimed to test different data mining methods for predicting ischemic stroke [26]. They worked on 80 patients and 112 healthy individuals and used 17 different predictors.

SVM and some other methods were used in this paper. The radial basis function (RBF) was selected as the kernel function. To evaluate the model, the tenfold cross validation resampling method was employed.

Consequently, the accuracy and AUC values with Confidence Intervals of 95% were 0.9789 and 0.9783 respectively, for SVM. Therefore, SVM obtained the best predictive performance among other methods.

Acharya et al. applied data mining approaches for breast lesion classification. To diagnosis breast cancer, they used 156 patients' records including 73 with malignant and 83 with benign cancer. The data mining techniques reduced the time and improved the accuracy of diagnosis. Also, the data mining techniques ranked the parameters according to their importance. SVM was one of the classification methods which was adopted in this study and achieved an acceptable value for accuracy, sensitivity, and specificity [18].

In some other researches, they used SVM for the diagnosis of coronary artery disease, detecting early childhood obesity, constructing the obesity prediction model, breast cancer diagnosis and predicting heart diseases [19, 31, 43, 45, 46].

Logistic regression

Logistic regression is a statistical technique for regression analysis and classification. Through the given features of data it computes a linear combination for the input data and passes through the logistic function. One of its application is representing the occurrence or non-occurrence of some events, which can be used for predicting diseases. It is easy to implement and works well for large datasets [14] and also it is very good for low latency applications.

Easton et al. designed prediction models for post-stroke by logistic regression and other data mining methods. Post-stroke mortality was studied in two-time phases: very short term and short/intermediate term. The prediction models were obtained from obvious risk factors that exist in UK Glucose Insulin in Stroke Trial dataset. This dataset includes 933 clinical records collected during 3 months. Logistic regression was one of the prediction methods used for building the model. It had 88.9% sensitivity and 59.9% specificity for very short term mortality and 81.5% sensitivity and 51.0% specificity for short/intermediate term. The results of logistic regression were acceptable but not ideal [27].

Heydari et al. used logistic regression to detect obesity among 414 patients. Among all risk factors related to obesity, 4 significant factors were measured by p-value applied in logistic regression which had 80.2% correct classification, 80.2% sensitivity, 81.9% specificity and AUC = 0.888 [28].

Pochini et al. attempted to find related risk factors to obesity and overweight. They applied logistic regression to detect risk factors and protective factors for handling obesity and overweight rate. They performed logistic regression on the data of 15,425 students and found that for preventing obesity students should have physical activity, eat breakfast, and avoid tobacco [29].

Charlton et al. aimed to cluster risk factors for predicting future heart disease and diabetes among adolescents with low fitness. This study examines 1147 students in 3 main clusters: children at low risk (not obese, fit), children 'visibly at risk' (overweight, unfit) and 'invisibly at risk' (unfit but not overweight). Logistic regression was used as a method for clustering risk factors with Confidence Intervals of 95% reported for heart disease and diabetes risk [30].

Logistic regression is a popular method for detection and prediction of various diseases widely applied for discovering risk factors for instance for detecting drug-drug interactions [47]. To mention some examples of predictive models we

can refer to the prediction of ischemic stroke, child mortality, diabetes or prediabetes, hypertension and hyperlipidemia, type-2 diabetes and heart disease [16, 17, 19, 25, 26, 41, 48]. It is also employed to analyze risk factors of the disease, for example in predicting risk factors of obesity among middle-aged people and preschool children [33, 49]. as well as evaluating the importance of selected features to predict obesity [43].

Neural networks

A neural network is a network of artificial neurons that processes information like biological neurons mostly used for classification and prediction. A neural network consists of connected input/output units in which each connection has a weight associated with it. To predict the correct class label of input data, the neural network adjusts the weights for the learning to happen [11].

The most popular neural network algorithm is backpropagation which performs learning on a multilayer feed-forward neural network. This algorithm iteratively learns a set of weights for the prediction of the class label of data samples. A multilayer feed-forward neural network includes an input layer, one or more hidden layers, and an output layer [11].

This method is useful in clinical medicine because of its good predictive performance and modeling complex issues and non-linear problems in comparison with logistic regression and Naïve Bayes. It can cope with noisy data and be used for continuous value inputs and outputs, unlike decision trees.

The neural network is suitable for problems with little knowledge of the associations between attributes and classes [11]. But, there are some negative points like being costly (computational), demanding long training time, and being uninterpretable for domain experts and sensitive to initial parameters [11].

As previously mentioned, Chen et al. used data mining to explore risk factors of preterm birth based on 455 samples [22]. The neural network mined 15 most important factors out of original factors. Afterwards, a C5.0 decision tree was employed for classifying risk factor related to preterm birth.

Alizadehsani et al. endeavored to diagnose coronary artery disease with the assistance of data mining methods. They employed neural network and some other methods on 303 patients with 54 features. They also used information gain and confidence to discover the importance of the features. Confidence is calculated as follow:

$$\text{Confidence} = P(y|x) = \frac{P(x \cap y)}{P(x)} \tag{8}$$

For feature selection, SVM was applied. For better detection of coronary artery disease, three features were derived from original features. The results of this study were grouped into 4 classes: (1) All features without three created features with 85.43% accuracy, 90.28% sensitivity, and 73.56% specificity; (2) All features and three created features with 87.11% accuracy, 91.67% sensitivity, and 75.86% specificity; (3) Selected features without three created features with 87.13% accuracy, 90.28% sensitivity and 79.31% specificity; and (4) Selected features and three created features with 88.11% accuracy, 91.20% sensitivity, and 80.46% specificity for the neural network. As can be seen, feature selection and feature creation achieved a better result [31].

As we mentioned previously, Heydari et al. desired to detect obesity in 414 patients. Among all risk factors related to obesity, 4 significant factors were measured by p-value applied in neural network. Neural network had 81.2% correct classification, 79.7% sensitivity, 83.7% specificity and AUC = 0.884. High classification accuracy made neural network a powerful method for detecting and predicting diseases [28].

The neural network is a powerful method for predicting and classifying diseases like breast cancer, diabetes or prediabetes and prediction heart disease [16, 18–20].

Deep Learning

Deep learning is a subset of machine learning in artificial intelligence but it is different from traditional machine learning. They have similarities in training, testing data and finding an optimized model. The differences between deep learning and traditional artificial neural networks (ANNs) lie in the number of hidden layers, the connections of hidden layers and learning meaningful abstractions from the inputs. The designing of layers in deep learning is learned from data that is both unstructured and unlabeled not by developers [50]. Some advantages like utilizing the unstructured data and new methods for preventing overfitting allow deep learning to solve a significant number of tasks [51]. Deep learning applications include object detection in images, speech recognition and natural language understanding, medicine, etc. [51].

Choi et al. aimed to predict the initial diagnosis of heart failure by a recurrent neural network (RNN). They used electronic health records on 3884 incident HF cases and 28 903 controls with 12- to 18-month observation window of cases and controls. The area under the curve (AUC) for the RNN model was 0.777 and they compared this achievement with other methods like logistic regression, multilayer perceptron (MLP) with 1 hidden layer, SVM and K-nearest neighbor (KNN). The results showed RNN as a deep learning method which improved the performance of models for the detection of incident heart failure [52].

Yala et al. studied 88 994 consecutive screening mammograms in 39 571 women to improved Breast Cancer Risk

Prediction. They adopted three models to assess breast cancer risk: a risk-factor-based logistic regression model (RF-LR) that used traditional risk factors, a DL model (image-only DL) that used mammograms alone, and a hybrid DL model that used both traditional risk factors and mammograms. The results represented that Hybrid DL achieved significantly higher AUC (0.70) than the other two models [53].

Furthermore, deep learning was employed for predicting lung cancer, classification of skin cancer, classification of pulmonary tuberculosis and prognosis and guiding therapy in adult congenital heart disease [54–57].

Naive Bayes

Naive Bayes is a simple machine learning classifier which is the most accurate in comparison with all other classifiers [11]. with natural and efficient handling of missing data. By having the models implemented the Bayesian classifier researchers achieve good prediction accuracy. Naive Bayes is a probabilistic classifier using the Maximum a Posteriori decision rule. Given a classification problem represented by a vector $X=(x_1,\ldots,x_n)$ representing some n independent variables [12]. It assigns to this problem probabilities:

$$p(C_k|x_1,\ldots,x_n) \tag{9}$$

for each of K classes C_k.

Using Bayesian probability:

$$posterior = \frac{prior \times likelihood}{evidence} \tag{10}$$

The conditional probability can be decomposed as:

$$p(C_k|x) = \frac{p(C_k)p(x|C_k)}{p(x)} \tag{11}$$

Naive Bayes' computational efficiency aids in the classification of high-dimensional data points and large datasets.

Naive Bayes conveniently handles classification by simply choosing the C_i that has the largest probability given the data point's features [12].

$$y = argmax_{C_i} p\,(C_i)\prod_{j=1}^{n} p\,(x_j|C_i) \tag{12}$$

As discussed in logistic regression section, Easton et al. desired to predict post-stroke by data mining methods. They studied post-stroke mortality in two-time phases: very short term and short/intermediate term. The dataset includes 933 clinical records collected for 3 months. Naive Bayes had 92.6% sensitivity and 69.4% specificity for very short term mortality, 83.3% sensitivity and 70.8% specificity for short/intermediate term. Compared to other method results, Naive Bayes achieved high accuracy in post-stroke prediction [27].

Adnan et al. discovered parameters related to childhood obesity which were suitable to construct a prediction model, and extracted obesity risk factors from articles, books, etc. Also, they grouped these factors into three categories: children's attributes, lifestyle, and Family or environmental factors. They used Naive Bayes to select parameters and compare them with the previously selected set. The new set enhanced the accuracy of childhood obesity prediction [32].

Hossain et al. identified the major obesity risk factors in middle-aged people by investigating 259 patients. From 19 features, 7 significant features were selected used in data mining prediction models. Naïve Bayes prediction model generates the best prediction accuracy (99.2% accuracy) according to selected features [33]. In addition, Naïve Bayes is employed in other studies like the prediction of coronary artery disease, prediction and classification of obesity in children and prediction of heart disease [19, 31, 42, 45].

Association rule mining

Association rule mining is the process of finding if–then rules for discovering an association between variables in large datasets. This method is widely adopted in marketing but found its way to other fields and has significant outcomes. Given a set of transactions $D = \{T_1,\ldots,T_n\}$ and a set of items $I = \{i_1,\ldots,i_n\}$, any transaction T in D is a set of items in I. An association rule has two parts: (1) left-hand side (LHS) called an antecedent, and (2) right-hand side (RHS) called a consequent that are subsets of a transaction T in D with the form of the rules being $x \Rightarrow y$. Association rule mining starts searching frequent items in datasets by using some algorithms to find frequent item sets [11].

Apriori is an example of this bunch of algorithms aimed to construct frequent item sets, having at least a user-specified threshold called minimum support. The golden point in Apriori is that an item set X of length k is frequent if and only if every subset of X, having length k—1, is also frequent. This point helps to reduce searching space and finding frequent subsets in less time. Combination of these subsets constructs rules. Also, support and confidence are the criteria that represent the performance of rules. Support is an indicator of how frequently the items appear in the data. Confidence indicates the number of times the if–then statements are found true [11]. These criteria are calculated as follows:

$$\text{Support} = P(x \cap y) \tag{13}$$

$$\text{Confidence} = P(y|x) = \frac{P(x \cap y)}{P(x)} \tag{14}$$

In medicine, association rule mining can be used for analyzing disease risk factors and extraction of if–then rules for better decision making.

Ilayaraja et al. used apriori to discover frequent diseases that helped to make the clinical decision. The dataset was collected from different geographical locations in 12 months of the year 2012 and consisted of 1246 clinical records. The minimum support and minimum confidence were 0.35 and 0.9, respectively. As a result of the apriori algorithm, heart disease, liver disease, overweight, and smoking were the most frequent among 29 diseases [34].

Nahar et al. identified the related factors to heart disease in males and females by using association rules. For generating rules, three algorithms were employed: Apriori, Predictive Apriori and Tertius. Apriori is previously explained, Predictive Apriori uses accuracy instead of confidence and Tertius algorithm finds the most confirmed hypotheses. The study was conducted in two phases: (1) detecting disease factors and health factors, and (2) extracting rules according to gender. The three algorithms reported the results of confidence > 90%, accuracy > 99% and confirmation > 79%, respectively. Only the " healthy " or "sick" class came in the right-hand side of the extracted rules. In the case of the absence of a rule for these conditions, the rules that "health" or "sick" class came in their left-hand side were extracted. In the first phase, Apriori, Predictive Apriori and Tertius detected different features. In general, males are at a higher risk of heart disease. In the second phase, they used Apriori to extract rules in males and females. To select risk factors from rules, they considered the majority of factors in rules and also the factors with high confidence [35].

Sharma et al. attempted to detect ways to decrease the spread of obesity via association rules. This study focused on physical exercise and sleeping as risk factors. Association rules algorithm was used for setting up the exercise pattern to control obesity among adults. Therefore, physical exercise and sleeping were considered as a frequent item set. In this study, three exercise patterns were tested: (1) fitness facility; (2) living room and housing exercising equipment; and (3) bedside. The support and confidence were calculated by counting the number of times each person exercises in three months. In the first two patterns, the frequency of exercise steadily declined, but the third pattern has been higher on average [36].

Ramezankhani et al. extracted risk factors for type 2 diabetes with association rules, and investigated 6647 clinical records. Equal frequency binning method was employed for categorizing variables did not have a clinical cutoff. In addition to the support and confidence for evaluating the rules lift measure were employed in this study.

The lift equation is:

$$Lift = \frac{conf(x \rightarrow y)}{conf(\theta \rightarrow y)} = \frac{supp(X \cup Y)/supp(X)}{supp(X)/supp(\theta)} \quad (15)$$

where supp(θ) is the number of records in the dataset. The relationship between the two variables is measured by lift. With the initialization of support and confidence, no rule was found for men, so they lowered the amount set to find some rules. As a result, some new risk factors were discovered in comparison with previous works [37].

Salehnasab et al. studied the variables related to high blood pressure. They applied association rules on 1000 patients. To identify frequent itemset apriori was employed. They used feature selection for determining the importance of variables. The rules were analyzed by Microsoft BI and Clementine software [38].

Ordonez compared association rules and decision trees to predict heart disease, where the concept of search constraints was proposed. Search constraints found the association rules that were important only in medicine. Support, confidence and lift were regarded as performance indicators. Association rules and decision trees were applied to 655 patients with 25 features.

The following search constraint was included:

(1) Maximum itemset size which helped to find simpler rules.
(2) Limit for variable placement in the right or left side of rules which assured researchers to find only the prediction rules where the variables of the disease come to the right of the rule.
(3) Grouping similar variables which helped to find the explicit rules.

In comparison, for diseases with multiple target variables, association rules were more appropriate than decision trees. The decision tree extracted rules that are simple with little reliability and most variables were split differently from that in medicine. Also, the extracted rules were related to a small set of patients. As a result, the paper concluded that association rules work with parameters that can be controlled by the user. Association rules generates rules with higher confidence and more relevance to a larger set of patients [39].

In some of the other studies, association rules were used for the diagnosis of coronary heart disease and analyzing the risk factors for diabetes [31, 44, 58, 59].

Table 1 showed that the data mininig techniques in medicine and Supplement Table 1 described that the summary of Table 1 with the advantages and disadvantages of each methodes.

Conclusion and future works

In this article, we surveyed the most common medical data mining techniques and summarized the results from the reviewed articles. Data mining is widely applied in diagnosis, prognosis, risk factors analysis of various diseases,

clinical decision making, etc. Data mining algorithms provide highly accurate predictive and classification models. We can utilize data mining techniques to make medical decision with less clinical experiments. Clinical experiments are expensive and the data may be affected by measurement errors or missing values commonly encountered by data mining [60].

Most papers apply classification methods like decision tree, artificial neural network and support vector machine to predict various diseases [61, 62].

There is a large number of articles that study specific diseases and use data mining methods to model diseases. The authors of [13] concluded that association rules analysis discovered important rules in metabolic syndrome, and the decision tree was a robust method for classifying metabolic syndrome and determining risk factors.

Most studies which employed data mining for analyzing diabetes adopted classification methods like decision trees and k-means as clustering methods [63]. Among the various data mining classifiers, the decision tree was the best classifier with 93.62% accuracy in breast cancer. Also, the Bayesian network was successfully utilized for breast cancer prognosis and diagnosis [64].

In some papers, neural networks achieved an accuracy of 100% in the prediction of heart disease. In addition, the decision tree performed well with 99.62% accuracy by using 15 attributes [65].

Most researches applied decision trees and logistic regression for classification in different diseases. The neural network was also employed in some papers. Other classification methods like support vector machine, Naïve Bayes and random forest are less frequently used. With the purpose of clustering, k-means successfully grouped original datasets. In descriptive models, association rules have a great role to discover hidden patterns and extract important rules. Association rules helped to determine disease risk factors and detect the association between different attributes.

As future work, researchers can utilize association rules in another kind of diseases and compare with standard medical results and extract interesting and specific relationships and patterns. They can also combine data sets to discover the association between different diseases.

Currently, multi-diseases predictive models offer much information to physicians and scientists about the detection of common risk factors and their impact on various diseases; now, there are a few articles in this area which need more exploration.

Furthermore, working with continuous data can also provide researchers with more information and more accurate models.

Integrating machine learning and data mining techniques is helpful in examining all aspects of a medical issue and generates more detailed solutions [66].

Future predictive and descriptive models in the field of medicine should have higher accuracy, greater comprehensiveness, and faster speed.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s40200-021-00884-2.

Authors' contributions All authors read and approved the final paper.

Declarations

Conflict of interest The authors declare they have no conflicting financial interests.

References

1. Fayyad U, Piatetsky-Shapiro G, Smyth P. Knowledge Discovery and Data Mining: Towards a Unifying Framework. Proceeding KDD'96 Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. 1996:82–8.
2. M. Durairaj VR. Data Mining Applications In Healthcare Sector: A Study. Int J Sci Technol Res. 2013;2(10).
3. Song C-W, Jung H, Chung K. Development of a medical big-data mining process using topic modeling. Cluster Computing. 2017.
4. Gonzalez GH, Tahsin T, Goodale BC, Greene AC, Greene CS. Recent advances and emerging applications in text and data mining for biomedical discovery. Brief Bioinform. 2016;17(1):33–42.
5. Salazar J, Espinoza C, Mindiola A, Bermudez V. Data mining and endocrine diseases: a new way to classify? Arch Med Res. 2018;49(3):213–5.
6. YoussefAgha AH, Lohrmann DK, Jayawardene WP. Use of data mining to reveal body mass index (BMI): patterns among Pennsylvania schoolchildren, pre-k to grade 12. J Sch Health. 2013;83(2):85–92.
7. Hosseini M, Ataei N, Aghamohammadi A, Yousefifard M, Taslimi S, Ataei F. The relation of body mass index and blood pressure in Iranian children and adolescents aged 7–18 years old. Iran J Public Health. 2010;39(4):126–34.
8. Hand D. Statistics and data mining: intersecting disciplines. ACM SIGKDD Explorations Newsl. 1999;1(1):16–9.
9. Makridakis S, Spiliotis E, Assimakopoulos V. Statistical and Machine Learning forecasting methods: Concerns and ways forward. PLoS One. 2018;13(3):e0194889.
10. Kantardzic M. Data Mining: Concepts, Models, Methods, and Algorithms. second ed: Wiley-IEEE Press; 2011.
11. Han J, Kamber M, Pei J. Data Mining: Concepts and Techniques. third ed: Morgan Kaufmann; 2011.
12. Mitchell T. Machine Learning. first ed: McGraw-Hill Education; 1997.
13. Worachartcheewan A, Schaduangrat N, Prachayasittikul V, Nantasenamat C. Data mining for the identification of metabolic syndrome status. EXCLI J. 2018;17:72–88.
14. Bishop C. Pattern Recognition and Machine Learning. first ed: Springer-Verlag New York; 2006. 738 p.
15. Shaikhina T, Lowe D, Daga S, Briggs D, Higgins R, Khovanova N. Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation. Biomed Signal Process Control. 2017.
16. Meng XH, Huang YX, Rao DP, Zhang Q, Liu Q. Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. Kaohsiung J Med Sci. 2013;29(2):93–9.

17. Tesfaye B, Atique S, Elias N, Dibaba L, Shabbir SA, Kebede M. Determinants and development of a web-based child mortality prediction model in resource-limited settings: a data mining approach. Comput Methods Programs Biomed. 2017;140:45–51.
18. Acharya UR, Ng WL, Rahmat K, Sudarshan VK, Koh JEW, Tan JH, et al. Data mining framework for breast lesion classification in shear wave ultrasound: a hybrid feature paradigm. Biomed Signal Process Control. 2017;33:400–10.
19. Mohammad Shafenoor Amin YKC, Varathan KD. Identification of significant features and data mining techniques in predicting heart disease. Telematics Inform. 2019;36:82–93.
20. Srabanti Maji SA. Decision tree algorithms for prediction of heart disease. Inform Commun Technol Compet Strat. 2019;40:447-54.
21. Tayefi M, Tajfard M, Saffar S, Hanachi P, Amirabadizadeh AR, Esmaeily H, et al. hs-CRP is strongly associated with coronary heart disease (CHD): A data mining approach using decision tree algorithm. Comput Methods Programs Biomed. 2017;141:105–9.
22. Chen H-Y, Chuang C-H, Yang Y-J, Wu T-P. Exploring the risk factors of preterm birth using data mining. Expert Syst Appl. 2011;38(5):5384–7.
23. Khalilia, Chakraborty S, Popescu M. Predicting disease risks from highly imbalanced data using random forest. BMC Med Inform Decis Mak 2011;11.
24. Ahamad M, Ahmed M, Uddin M. Clustering as Data Mining Technique in Risk factors Analysis of Diabetes, Hypertension and Obesity. Eur J Eng Res Sci. 2016;1.
25. Han Wu SY, Zhangqin Huang, Jian He, Xiaoyi Wang (2018) Type 2 diabetes mellitus prediction model based on data mining. Inform Med Unlock.;10:100–7.
26. Arslan AK, Colak C, Sarihan ME. Different medical data mining approaches based prediction of ischemic stroke. Comput Methods Programs Biomed. 2016;130:87–92.
27. Easton JF, Stephens CR, Angelova M. Risk factors and prediction of very short term versus short/intermediate term post-stroke mortality: a data mining approach. Comput Biol Med. 2014;54:199–210.
28. Heydari ST, Ayatollahi SM, Zare N. Comparison of artificial neural networks with logistic regression for detection of obesity. J Med Syst. 2012;36(4):2449–54.
29. Pochini A, Wu Y, Hu G. Data Mining for Lifestyle Risk Factors Associated with Overweight and Obesity among Adolescents. 2014 IIAI 3rd Int Conf Adv Appl Inform; 2014. p. 883–8.
30. Charlton R, Gravenor M, Rees A, Knox G. Factors associated with low fitness in adolescents – A mixed methods study. BMC Public Health. 2014;14.
31. 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(1):52–61.
32. Adnan M, Husain W, Abdul Rashid N. Parameter Identification and Selection for Childhood Obesity Prediction Using Data Mining. 2nd International Conference on Management and Artificial Intelligence2012.
33. Hossaina R, Mahmud S, Hossin M. PRMT: predicting risk factor of obesity among middle-aged people using data mining techniques. Inter Conf Comput Intell Data Sci. 2018:1068–76.
34. Ilayaraja M, Meyyappan T. Mining medical data to identify frequent diseases using Apriori Algorithm. Proceedings of the 2013 Int Conf Pattern Recog Inform Mobile Eng 2013.
35. Nahar J, Imam T, Tickle KS, Chen Y-PP. Association rule mining to detect factors which contribute to heart disease in males and females. Expert Syst Appl. 2013;40(4):1086–93.
36. Sharma S. Concept of association rule of data mining assists mitigating the increasing obesity. Int J Inf Ret Res. 2017;7(2):1–18.
37. Ramezankhani A, Pournik O, Shahrabi J, Azizi F, Hadaegh F. An application of association rule mining to extract risk pattern for type 2 diabetes using tehran lipid and glucose study database. Int J Endocrinol Metab. 2015;13(2):e25389.
38. Salehnasab C, Jahandideh F, Ahmadzadeh M. Use association rules to study the relation between variables that affect high blood pressure. Acta HealthMedica. 2017;2(1)
39. Ordonez C. Comparing Association Rules and Decision Trees for Disease Prediction. Proceeding HIKM '06 Proceedings of the international workshop on Healthcare information and knowledge management. 2006:17–24.
40. Quinlan J. Induction of decision trees. Mach Learn. 1986;1:81–106.
41. Chang C-D, Wang C-C, Jiang BC. Using data mining techniques for multi-diseases prediction modeling of hypertension and hyperlipidemia by common risk factors. Expert Syst Appl. 2011;38(5):5507–13.
42. Adnan MHM, Husain W, Rashid NAA. A Framework for Childhood Obesity Classifications and Predictions using NBtree. 2011 7th International Conference on Information Technology in Asia; 12–13 July 2011: IEEE; 2011.
43. Wang HY, Chang SC, Lin WY, Chen CH, Chiang SH, Huang KY, et al. Machine Learning-Based Method for Obesity Risk Evaluation Using Single-Nucleotide Polymorphisms Derived from Next-Generation Sequencing. J Comput Biol. 2018;25(12):1347–60.
44. Vijayalakshmi N, Jenifer T. An analysis of risk factors for diabetes using data mining approach. Int J Comput Sci Mob Comput. 2017;6(7):166–72.
45. Lingren T, Thaker V, Brady C, Namjou B, Kennebeck S, Bickel J, et al. Developing an algorithm to detect early childhood obesity in Two Tertiary Pediatric Medical Centers. Appl Clin Inform. 2016;7(3):693–706.
46. Haifeng Wang BZ, Yoon SW, Ko HS. A support vector machine-based ensemble algorithm for breast cancer diagnosis. Eur J Oper Res. 2018;267(2):687–99.
47. Vilar S, Friedman C, Hripcsak G. Detection of drug-drug interactions through data mining studies using clinical sources, scientific literature and social media. Brief Bioinform. 2018;19(5):863–77.
48. Razavian N, Blecker S, Schmidt AM, Smith-McLallen A, Nigam S, Sontag D. Population-level prediction of type 2 diabetes from claims data and analysis of risk factors. Big Data. 2015;3(4):277–87.
49. Dev DA, McBride BA, Fiese BH, Jones BL, Cho H, Behalf Of The Strong Kids Research T. Risk factors for overweight/obesity in preschool children: an ecological approach. Child Obes. 2013;9(5):399–408.
50. DE Rumelhart HG, Williams RJ. Learning representations by back-propagating errors. Nature. 1986;323:533–6.
51. Riccardo Miotto FW, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. Brief Bioinform. 2018;19(6):1236–46.
52. Edward Choi AS, Stewart Walter F, Sun J. Using recurrent neural network models for early detection of heart failure onset. J Am Med Inform Assoc. 2017;24(2):361–70.
53. Adam Yala CL, Schuster Tal, Portnoi Tally, Barzilay Regina. A deep learning mammography-based model for improved breast cancer risk prediction. Radiology. 2019;292(1):60–6.
54. Diego Ardila APK, Bharadwaj S, Choi B. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nat Med. 2019;25:954–61.
55. Andre Esteva BK, Novoa R A, Ko J, Swetter SM. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542:115–8.
56. Lakhani PSB. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. Radiology. 2017;284(2):574–82.
57. Gerhard-Paul Diller AK, Babu-Narayan SV. Machine learning algorithms estimating prognosis and guiding therapy in adult

congenital heart disease: data froma single tertiary centre including 10 019 patients. Eur Heart J. 2019;40(13):1069–77.
58. Ordonez C, Omiecinski E, de Braal L, Santana CA, Ezquerra N, Taboada JA, et al. Mining constrained association rules to predict heart disease. Proceedings 2001 IEEE Int Conf Data Min; 2001. p. 433–40.
59. Ordonez C. Association rule discovery with the train and test approach for heart disease prediction. IEEE Trans Inf Technol Biomed. 2006;10.
60. Bellazzi R, Zupan B. Predictive data mining in clinical medicine: current issues and guidelines. Int J Med Inform. 2008;77(2):81–97.
61. Jothi N, Rashid NAA, Husain W. Data mining in healthcare – a review. Procedia Computer Science. 2015;72:306–13.
62. RaminGhorbani RG. Predictive data mining approaches in medical diagnosis: a review of some diseases prediction. Int J Data Net Sci. 2019;3:47–70.
63. Marinov i, Mosa M, Yoo l. Data-mining technologies for diabetes: a systematic review. J Diabetes Sci Technol. 2011;5(6(
64. Kharya S. Using data mining techniques for diagnosis and prognosis of cancer disease. Int J Comput Sci, Eng Inf Technol. 2012;2(2):55–66.
65. Kaur B, Singh W. Review on Heart Disease Prediction System using Data Mining Techniques. Int J Recent Innov Trends Comput Commun. 2014;2(10(
66. Muhammad Noman Sohail RJ, Muhammad Musa Uba. A Comprehensive Looks at Data Mining Techniques Contributing to Medical Data Growth: A Survey of Researcher Reviews. Recent Dev Intell Comput, Commun Devices. 2019;752:21–6.