Contents lists available at GrowingScience

International Journal of Data and Network Science

homepage: www.GrowingScience.com/ijds

Predictive data mining approaches in medical diagnosis: A review of some diseases prediction

Ramin Ghorbani[a*] and Rouzbeh Ghousi[a]

[a]*Department of Industrial Engineering, Iran University of Science and Technology, Tehran, Iran*

CHRONICLE

Article history:
Received: October 18, 2018
Received in revised format: December 20, 2018
Accepted: January 8, 2019
Available online:
January 8, 2019

Keywords:
Healthcare
Classification
Heart Disease
Breast Cancer
Diabetes Mellitus
Review

ABSTRACT

Due to the increasing technological advances in all fields, a considerable amount of data has been collected to be processed for different purposes. Data mining is the process of determining and analyzing hidden information from different perspectives to obtain useful knowledge. Data mining can have many various applications, one of them is in medical diagnosis. Today, many diseases are regarded as dangerous and deadly. Heart disease, breast cancer, and diabetes are among the most dangerous ones. This paper investigates 168 articles associated with the implementation of data mining for diagnosing such diseases. The study concentrates on 85 selected papers which have received more attention between 1997 and 2018. All algorithms, data mining models, and evaluation methods are thoroughly reviewed with special consideration. The study attempts to determine the most efficient data mining methods used for medical diagnosing purposes. Also, one of the other significant results of this study is the detection of research gaps in the application of data mining in health care.

1. Introduction

We live in a world where large volumes of data are collected every day and analyzing such data plays an essential role in business management (Han et al., 2011). In the past, traditional methods were used to analyze the data, which relied on manual operations. Data analysis using the traditional method was time-consuming and frustrating operations. Furthermore, it was impractical in many cases. Knowledge discovery is considered a significant challenge. The purpose of extracting knowledge is to discover useful knowledge, and data mining is one of the steps in knowledge discovery to obtain useful information. Data mining is the process of detecting and extracting hidden information, patterns and specific data connections of the prediction idea. Data mining is a new discipline with different applications known as one of the ten leading sciences influencing technology. Wherever the data exists, data mining is also meaningful, for instance: Market Basket Analysis, Education, Manufacturing Engineering, Customer Relationship Management, Fraud Detection, Intrusion Detection, Lie Detection, Customer Segmentation, Financial-Banking, Corporate Surveillance, Research Analysis, Criminal Investigation, Telecommunication, and Healthcare.

* Corresponding author. Tel.: +989135470588
E-mail address: ramin.ghorbani73@gmail.com (R. Ghorbani)

doi: 10.5267/j.ijdns.2019.1.003

Today, the healthcare industry generates large amounts of complex data on patients, hospital resources, diagnosis of diseases, electronic patient records and medical devices. More copious amounts of data are an essential resource for data mining. There is a vast potential in healthcare data mining applications, and some of the most critical applications in healthcare data mining are prediction and diagnosis, treatment effectiveness, healthcare management, fraud and abuse, customer relationship management, and the medical device industry (Koh & Tan, 2011). Choosing the wrong treatment for patients will not only waste time and money but also can cause adverse effects such as the death of patients. Therefore, a method for diagnosing and selecting the appropriate treatment is essential for patients. Data mining can help with the prediction and determination of the diseases in this area. In this study, concerning the importance of early detection, 168 articles on heart disease, breast cancer, and diabetes have been selected to review their performance in the field of prediction. After the initial review of these articles, 85 research is chosen for the analysis throughout 1997-2018. We hope the present study will be helpful for the future studies. The paper is prepared as follows: Section 2 explains knowledge discovery in databases and data mining concepts. Section 3 describes the research strategy used in this study. Section 4-6 thoroughly evaluate and report the review results of the heart diseases, breast cancer, and diabetes mellitus. Finally, the conclusion and future work recommendations are presented in section7.

2. Concepts

2.1. Knowledge Discovery in Databases

Knowledge discovery in databases (KDD) is the process of determining useful and helpful knowledge from the collection of the data. The steps of knowledge extraction are necessary to achieve essential knowledge, and blindly data mining can easily lead to meaningless patterns, which is very dangerous. Fig. 1 displays the knowledge discovery steps.

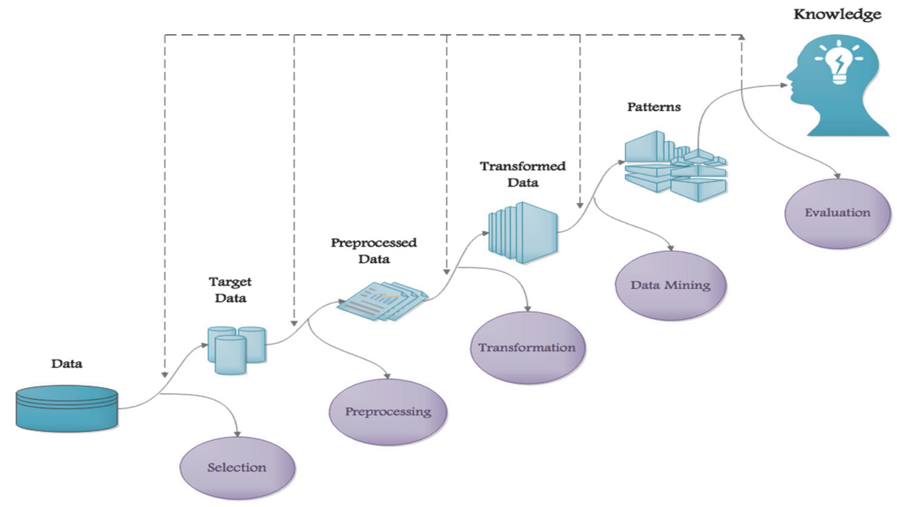

Fig. 1. Steps of knowledge discovery in databases

2.2 Data Mining

Data mining is one of the steps of knowledge discovery in a database as an effort to gather helpful information. Data mining is a new discipline with various applications known as one of the top ten sciences affecting technology. There are various major data mining techniques have been developing and applying in projects including classification, clustering, and association rules.

2.2.1. Classification

Classification is one of the data analysis techniques. Classification assigns items to target categories in a collection. This technique puts the same set of features into a class. Decision Tree, Bayesian Network, Rule-Based Classification, Artificial Neural Network, Support Vector Machine, Associative Classification, K-Nearest Neighbors, Genetic Algorithm, Rough Set Approach, and Fuzzy Set Classification are some of the classification methods.

2.2.2 Clustering

Clustering method finds clusters of data objects that are similar in some information to one another. In this technique, the members of a group are more like each other than to those in other groups. The clustering technique determines the classes and sets objects in each category, while in the classification techniques, objects are specified into predefined categories. K-Means, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Fuzzy Clustering, and Expected Maximization (EM) are some of the clustering methods.

2.2.3 Association Rules

Association rule technique discovers new relations between variables in a database. This technique focuses on finding frequent patterns among a collection of items.

3. Research Strategy

In this paper, 168 articles on heart disease, breast cancer, and diabetes were selected. After the initial searches, 85 research studies were selected for analysis and final examination between 1997 and 2018. Fig. 2 demonstrates the area and the number of these articles separately. These studies were identified by using the databases like IEEE Xplore, Google Scholar, Science Direct, and Springer Link.

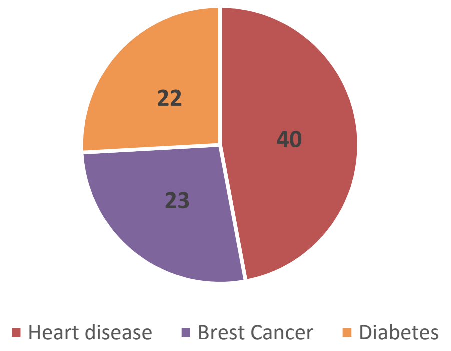

Fig. 2. The area and the number of selected articles

4. Heart Diseases

Cardiovascular or heart diseases are heart conditions that include diseased vessels, structural problems, and blood clots. Heart disease is so significant that many people have tried to investigate further for early diagnosis and effective treatment of cardiovascular diseases. Using data mining from information related to heart patients can create valuable knowledge to improve heart disease diagnosis. Studied research on heart disease is selected between 2008 and 2018. Among the 40 studied research, five articles have been studied in the form of a review paper, and 35 articles are associated with applications.

4.1 Literature Review

The application of data mining begins a new dimension to cardiovascular disease prediction. Several data mining techniques are used for identifying and extracting valuable information from the clinical dataset (Srinivas et al., 2010). Researchers investigated numerous ways to implement data mining in healthcare to achieve an accurate prediction accuracy.

Table 1
The overall review of the data mining techniques in heart diseases diagnosis

Articles	Classification	Clustering			Outlier detection	Association			Model Evaluation	Model Comparison	Ensemble methods			Hybrid	Optimization				Fuzzy approach	Decision Making Algorithms	Feature Selection
		Partitioning methods				Frequent pattern mining															
		Unspecific	K-Means	Fuzzy cluster		Apriori	Mafia	Weighted Association			Unspecific	Voting	Bagging		Particle Swarm Optimization	Ant Colony Optimization	Principal Component Analysis	Sequential Minimal Optimization			
Palaniappan and Awang (2008)	✓	✓							✓	✓											
Das et al. (2009)	✓								✓	✓	✓										
Tu et al. (2009)	✓								✓	✓			✓								
Rajkumar and Reena (2010)	✓								✓	✓											
Shouman et al. (2011)	✓								✓	✓		✓	✓								
Alizadehsani et al. (2012)	✓								✓	✓											✓
Bhatla and Jyoti (2012)	✓								✓	✓											
	✓										✓								✓		
	✓								✓	✓											
	✓								✓	✓											✓
	✓								✓	✓											
Vijiyarani and Sudha (2013)	✓								✓	✓											
Alizadehsani et al. (2013)	✓								✓	✓			✓					✓			✓
Jabbar et al. (2013)	✓	✓							✓	✓				✓							✓
Ratnakar et al. (2013)	✓								✓	✓				✓							
Masethe and Masethe (2014)	✓								✓	✓											
Kaur (2014)	✓			✓					✓	✓				✓							
Cinetha and Maheswari (2014)	✓								✓	✓				✓					✓		
Devi and Anto (2014)	✓								✓	✓	✓			✓					✓		
Venkatalakshmi and Shivsankar (2014)	✓								✓	✓											
Methaila et al. (2014)	✓					✓	✓	✓	✓	✓				✓							✓
Kim et al. (2015)	✓								✓	✓				✓					✓		
Verma et al. (2016)	✓		✓						✓	✓	✓			✓	✓				✓		✓
Verma and Srivastava (2016)	✓								✓	✓											
Kausar et al. (2016)	✓		✓						✓	✓	✓			✓			✓				
Baihaqi et al. (2016)	✓				✓				✓	✓	✓								✓		
Joshi et al. (2016)	✓				✓				✓	✓											
Malav et al. (2017)	✓		✓						✓	✓				✓							
Samuel et al. (2017)	✓								✓	✓				✓					✓	✓	✓
Al-Maqaleh and Abdullah (2017)	✓								✓	✓											
Dekamin and Sheibatolhamdi (2017)	✓		✓		✓				✓	✓											✓
Babu et al. (2017)	✓		✓				✓							✓							
Bhargava et al. (2017)	✓								✓	✓											
Singh et al. (2018a)	✓								✓	✓				✓							
Kulkarni et al. (2018)	✓								✓												
Shirwalkar et al. (2018)	✓		✓																		
Singh et al. (2018b)	✓								✓												
Wadhawan (2018)	✓		✓			✓	✓		✓		✓			✓							
Kurian and Lakshmi (2018)	✓								✓	✓		✓									

Table 1 exposes comprehensive information about the all implemented methods, and the concept of each paper is discussed as follows:

Palaniappan and Awang (2008) developed a prototype intelligent heart disease prediction system using data mining techniques, namely, Decision Trees, Naive Bayes, and Neural Network. They used the CRISP-DM methodology to build the mining models. Results showed that each method has its unique strength in realizing the objectives of the defined mining goals. Ensemble approaches, which use multiple data mining algorithms, have confirmed to be an effective technique of improving classification accuracy. Das et al. (2009) introduced a methodology for diagnosing of the heart disease. They propose a Neural Networks ensemble method using SAS base software by combining three independent Neural Networks models. They obtained 89.01% classification accuracy from this ensemble model. Tu et al. (2009) proposed the use of a Bagging algorithm to diagnose heart disease in patients. They compared the effectiveness of the Bagging algorithm with the Decision Tree algorithm. In the end, the results show that Bagging algorithm increases the accuracy and this algorithm has better performance and efficiency than the Decision Tree. Rajkumar and Reena (2010) used supervised machine learning algorithms such as Naive Bayes, K-Nearest Neighbor, and Decision List. The results are compared by Tanagra tool and confirm that the Naive Bayes algorithm has the best processing time and prediction accuracy. Shouman et al. (2011) recommended a model that outperforms Decision Tree J48, Voting and Bagging algorithm in the early prediction of heart disease. One of their results shows that applying the Voting algorithm increases the efficiency of the Decision Tree. Alizadehsani et al. (2012) attempted to find a way for specifying the lesioned vessel when there are not enough electrocardiogram changes. They processed with Decision Tree C4.5, Naive Bayes, and K-Nearest Neighbor algorithms and the highest achieved accuracy was related to the C4.5 algorithm.

Bhatla and Jyoti (2012) aimed at analyzing the various data mining techniques for heart disease prediction. For better understanding, each data mining technique has been shown separately in a different part, and various classifiers are employed in combination with different data mining techniques for heart disease prediction. Vijiyarani and Sudha (2013) analyzed the Decision Tree techniques namely, Decision Stump, Random Forest, and Logistic Model Tree (LMT) algorithm to investigate the experimental results that are related to the performance of these techniques for a heart disease dataset. The results show that the Decision Stump technique is a more reliable classifier than others. Alizadehsani et al. (2013) introduced a feature creation method to the dataset. They used data mining techniques, namely, Sequential Minimal Optimization (SMO), Support Vector Machine, Naive Bays, Neural Network and Bagging ensemble method. This study has measured the accuracy values by using ten-fold cross-validation. Jabbar et al. (2013) applied a K-Nearest Neighbor algorithm with feature subset selection. As a way to validate the proposed method, they tested other machine learning data sets and compared the proposed system with other data mining techniques. Ratnakar et al. (2013) discussed modeling techniques namely, Naive Bayes, Decision Tree with a Genetic algorithm optimization to predict the risk level of heart disease. The experimental results show that the Decision Tree is better than the Naive Bays technique.

Masethe and Masethe (2014) proposed different models based on 11 attributes. They applied the following algorithms such as Decision Tree J48, Bayesian Network, Naive Bayes, Simple Cart, and Reptree algorithm to classify and develop a model to diagnose heart attacks. The research results do not present a dramatic difference in the prediction when using different classification algorithms in data mining. Kaur (2014) provided an intelligent heart disease prediction system. In this research, the efficiency of heart disease system will enhance by using Classification Rules, Fuzzy-C Means clustering, and Genetic algorithm optimization. The studied dataset contains a total of 303 records, 14 attributes, and various parameters like accuracy, time, specificity and sensitivity are calculated. Cinetha and Maheswari (2014) suggested a Decision Support System which predicts the possibility of heart disease risk of patients for the next ten years using Fuzzy Logic and Decision Tree. This model predicts with 97.67% estimated accuracy. Devi and Anto (2014) proposed an evolutionary fuzzy expert system for the diagnosis of coronary artery disease based on a dataset with a total of 303 records and 14 attributes. Venkatalakshmi and Shivsankar (2014) executed a comparison of heart disease diagnosis with the help of Decision Tree and Naive Bayes. The results show that the accuracy of Naive Bayes and the Decision Tree is 85.03 % and 84.01 %.

Methaila et al. (2014) intended to use data mining classification techniques, namely, Decision Trees, Naive Bayes and Neural Network, along with Weighted Association Apriori algorithm and Mafia algorithm in heart disease prediction. The experimental outcomes show that applying a Genetic algorithm improves the prediction accuracy. Kim et al. (2015) introduced a prediction model of coronary heart disease by utilizing Fuzzy Logic and CART-based rule induction. The accuracy is 69.51%, and the results show that the proposed model improves prediction accuracy and sensitivity. Verma et al. (2016) presented a hybrid method for heart disease diagnosis, including risk factor identification using correlation-based feature subset selection with Particle Swarm Optimization search method and K-Means clustering algorithms. Also, supervised learning algorithms such as Multi-Layer Perceptron(MLP), Multinomial Logistic Regression, Fuzzy Unordered Rule induction algorithm, and C4.5 are used. Verma and Srivastava (2016) presented a Radial Basis Function (RBF) and Probabilistic Neural Network (PNN) and Decision Tree models to predict coronary heart disease cases. Results show that Neural Network models achieved the highest prediction accuracy and lowest miss-classification error rate as compared to other diagnostic models.

Kausar et al. (2016) combined supervised, and unsupervised learning methods namely Support Vector Machines and K-Means clustering for classification by adjusting their related parameters and measures. They also selected Principal Component Analysis (PCA) algorithm to reduce the attribute dimension. Baihaqi et al. (2016) compared the performance of C4.5, CART, and RIPPER as a fuzzy rules generator to be used on the fuzzy expert system. The combination of data mining and fuzzy expert systems have been successfully carried out in this research to diagnose coronary heart disease. Joshi et al. (2016) presented a Decision Tree-based classification technique for accurate heart disease prediction. The results determine that the accuracy of the proposed method is better than other methods that are discussed in this paper. Malav et al. (2017) suggested an efficient hybrid combination of K-Means clustering algorithm and Artificial Neural Network. They compared Naive Bays and K-Nearest Neighbor models with the hybrid method, and the hybrid approach gave a higher accuracy rate. Samuel et al. (2017) developed a fuzzy analytic hierarchy process technique that computes the global weights for the attributes based on their contribution. The performance of the newly suggested Decision Support System was evaluated by using 297 records and 13 attributes of heart disease patients.

Al-Maqaleh and Abdullah (2017) proposed an intelligent predictive system using classification techniques for heart disease diagnosis, namely, J48 Decision Tree, Naive Bayes, and Multi-Layer Perceptron Neural Network. The experimental results are evaluated by the common performance metrics like accuracy, F-measure, and ROC graph. Dekamin and Sheibatolhamdi (2017) provided a data preparation method based on clustering algorithms with higher efficiency and fewer errors. Naive Bayes, KNN, and Decision Tree are used for classification. According to the results, the proposed method is highly successful. Babu et al. (2017) provided a prototype heart disease diagnosis using data mining technique such as Genetic algorithm, K-Means algorithm, Mafia algorithm, and Decision Tree classification. The results show that Decision Tree has great efficiency after applying a Genetic algorithm. Bhargava et al. (2017) undertook an experiment on an application of mining algorithm CART to predict the heart attacks and to compare the best available method of prediction. They evaluated the performance of the CART algorithm by calculating the time taken, confusion matrix, f-measure, recall, precision, and prediction accuracy. Singh et al. (2018a) tried to devise out a model that gives a highly accurate prediction of heart disease. They have done a combination of Genetic and Naive Bayes technique. The Research developed a hybrid model of both these techniques using Python 3.6 platform. Kulkarni et al. (2018) used the Decision Tree classification algorithm to assess the events related to heart disease. Their work was mainly concerned with the development of a data mining model with the Random Forest classification algorithm. Also, their work was a kind of review paper, and they discussed some classifiers too. Shirwalkar et al. (2018) showed that each algorithm contains specific functions which are helpful to diagnose heart disease. Their work was a kind of review paper and focused on classification and prediction methods of data mining using Naive Bayes and improved K-Means algorithm.

Singh et al. (2018b) developed an effective heart disease prediction system using the Neural Network for predicting the risk level of heart disease. The obtained results have illustrated that the designed diagnostic system can effectively predict the risk level with 100% accuracy. Wadhawan (2018) developed a system prototype which can help determine and extract hidden knowledge related to heart disease. The proposed technique combines rule mining using Apriori algorithm and Mafia algorithm as well as classification using K-Nearest Neighbors algorithm to predict the heart diseases efficiently. Kurian and Lakshmi (2018) introduced an ensemble classifier approach that is the combination of three classifiers namely K-Nearest Neighbor algorithm, Decision Tree, Naive Bayes. The ensemble model can be used to give predictions with better accuracy than the individual classifiers.

Thenmozhi and Deepika (2014) proposed a review paper on various Decision Tree algorithms in classifying and predict heart disease. They studied different researches with some useful techniques. Patel et al. (2017) suggested a review paper. They described a prototype using data mining techniques mainly Naive Bayes and Weighted Associated classifier and entirely explained these two techniques. Shouman et al. (2012) offered a review paper that identifies gaps in the research on heart disease diagnosis. One of the results shows that hybrid data mining techniques have shown promising outcomes in the diagnosis of heart disease. Kumari and Godara (2011) recommended a review paper to review data mining classification techniques namely, Ripper Classifier, Decision Tree, Artificial Neural Networks, and Support Vector Machine. They compared these techniques through the lift chart, error rate, sensitivity, specificity, and accuracy. Kadi et al. (2017) proposed a systematic review that investigated the studies that were performed in cardiology using data mining techniques. Four hundred and seven papers from between 2000 and 2015 were identified, and finally, 149 studies were selected. The obtained results showed that hybrid approaches appear to be more interesting to researchers.

4.1.1. Classification Technique Analysis

The classification technique is one of the main data mining techniques used in all the studies. Table 2 and Fig. 3 compare the classification methods used in heart diseases diagnosis. The Decision Tree and the Bayesian Classifier method are utilized more than other methods.

Table 2
Comparison of the classification methods

Methods	Frequency of use	Usage percent
Decision Tree Algorithm	26	67%
Bayesian Classifier	16	41%
Artificial Neural Network	12	31%
Genetic Algorithm	8	21%
KNN	7	18%
Fuzzy set Approach	6	15%
Rule-Based	3	8%
Associative Classification	2	5%
SVM	2	5%
Logistic Regression	1	3%
Rough Set Approach	0	0%

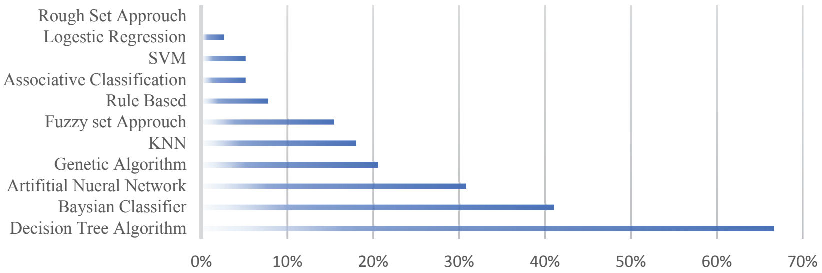

Fig. 3. Comparison of the classification methods

Table 3
The overall review of the classification methods

Article	Decision Tree Algorithm (DT)												K-Nearest Neighbors (KNN)	Bayesian		Support Vector Machine (SVM)	Rule based		Associative Classification	Artificial Neural Network					Genetic Algorithm	Rough Set Approach	Logistic Regression	Fuzzy Set Approach
	Unspecific	ID3	C5	CART	Rep tree	J48 (C4.5)	DT-Random (DTR)	LMT	DT-Forest (DTF)	DTR-Forest (DTRF)	DT stump	DT list		Naive Bays	Bayesian Network		IF-THEN	RIPPER		Unspecific	MLP	RBF	SOM	PNN				
Palaniappan and Awang (2008)	✓													✓						✓								
Das et al. (2009)																				✓								
Tu et al. (2009)						✓																						
Rajkumar and Reena (2010)												✓	✓	✓														
Shouman et al. (2011)						✓																						
Alizadehsani et al. (2012)						✓							✓	✓														
Bhatla and Jyoti (2012)																				✓								
																									✓			✓
												✓	✓	✓														
	✓													✓											✓			
	✓													✓						✓								
Vijiyarani and Sudha (2013)								✓		✓	✓																	
Alizadehsani et al. (2013)														✓		✓				✓								
Jabbar et al. (2013)													✓						✓									
Ratnakar et al. (2013)	✓													✓											✓			
Masethe and Masethe (2014)				✓	✓	✓								✓	✓													
Kaur (2014)																	✓								✓			
Cinetha and Maheswari (2014)	✓																											✓
Devi and Anto (2014)	✓																✓								✓			✓
Venkatalakshmi and Shivsankar (2014)	✓													✓														
Methaila et al. (2014)	✓													✓						✓					✓			
Kim et al. (2015)				✓															✓									✓
Verma et al. (2016)						✓															✓						✓	✓
Verma and Srivastava (2016)	✓																					✓		✓				
Kausar et al. (2016)																✓												
Baihaqi et al. (2016)				✓		✓												✓										✓
Joshi et al. (2016)	✓																											
Malav et al. (2017)																				✓								
Samuel et al. (2017)																				✓								
Al-Maqaleh and Abdullah (2017)						✓								✓							✓							
Dekamin and Sheibatolhamdi (2017)	✓												✓	✓														
Babu et al. (2017)	✓																								✓			
Bhargava et al. (2017)				✓																								
Singh et al. (2018a)														✓											✓			
Kulkarni et al. (2018)	✓									✓																		
Shirwalkar et al. (2018)														✓														
Singh et al. (2018b)																					✓							
Wadhawan (2018)													✓															
Kurian and Lakshmi (2018)	✓												✓	✓														

4.1.1.1. Decision Tree Method

Decision Tree algorithm is based on contingent possibilities. Decision Trees create rules, and a rule is a provisional statement that can easily be followed by humans and used within a database to recognize a set of records (Oracle, 2008). Unfortunately, some works of literature have not determined the name of the model used in the Decision Tree method.

Table 4
Comparison of the Decision Tree models

Models	Frequency of use	Usage percent
Unspecific Model	14	36%
C4.5 (J48)	7	18%
CART	4	10%
DTRF	2	5%
DT list	2	5%
Rep tree	1	3%
LMT	1	3%
DT stump	1	3%
ID3	0	0%
C5	0	0%
DTR	0	0%
DTF	0	0%

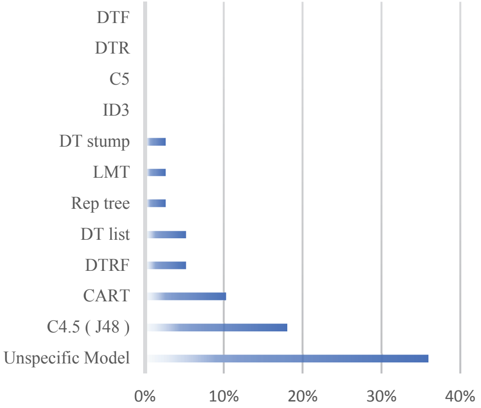

Fig. 4. Comparison of the Decision Tree models

4.1.1.2. *Artificial Neural Network Method*

Artificial Neural Network is an algorithm based on a biological Neural Network that is used to estimate or approximate functions depending on a large number of generally unknown inputs. (Oracle, 2008). Unfortunately, some works of literature have not determined the name of the model used in Artificial Neural Network method.

Table 5
Comparison of the Artificial Neural Network models

Models	Frequency of use	Usage percent
Unspecific Model	8	21%
MLP	3	8%
RBF	1	3%
PNN	1	3%
SOM	0	0%

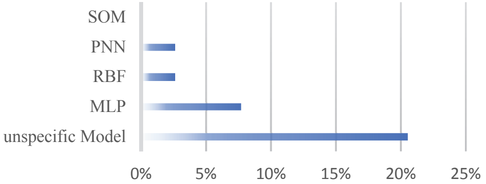

Fig. 5. Comparison of the Artificial Neural Network models

4.1.2. *Clustering Technique Analysis*

Clustering technique finds clusters of data objects that are similar in some senses to one another (Oracle, 2008). Table 6 and Fig. 6 compare the clustering methods in heart diseases diagnosis.

Table 6
Comparison of the clustering methods

Models	Frequency of use	Usage percent
K-Means	7	18%
Unspecific	2	5%
Fuzzy Cluster	1	3%

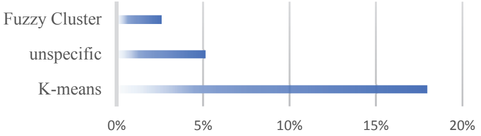

Fig. 6. Comparison of the clustering methods

4.1.3. *Evaluation Technique Analysis*

Evaluation methods determine the efficiency and performance of predictive models. These methods help to understand the quality of the model or any technique. Table7 and Fig. 7 compare the evaluation methods in heart diseases diagnosis.

Table 7
Comparison of the evaluation methods

Models	Frequency of use	Usage percent
Accuracy	35	90%
Sensitivity	17	44%
Confusion Matrix	12	31%
Specificity	12	31%
Time Taken	10	26%
ROC	8	21%
Precision	6	15%
Cross-Validation	6	15%
F-measure	4	10%
Performance Plot	1	3%
Lift Chart	1	3%
Classification Chart	1	3%

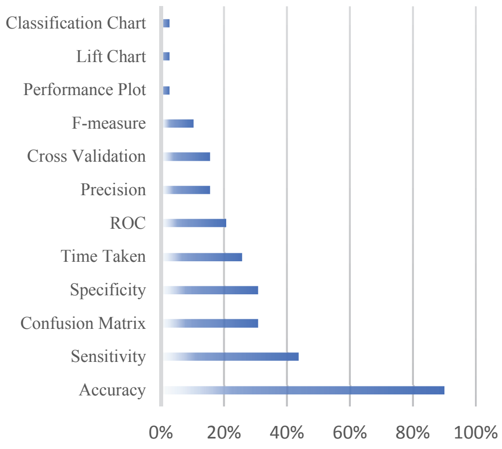

Fig. 7. Comparison of the evaluation methods

Table 8
The overall review of the evaluation methods

Article	Evaluation the Model											
	Confusion Matrix	performance Plot	lift Chart	accuracy	Time Taken	ROC	F-measure	classification chart	specificity	Sensitivity	Precision	Cross validation
Palaniappan and Awang (2008)	✓		✓									
Das et al. (2009)	✓			✓		✓		✓	✓	✓		
Tu et al. (2009)				✓					✓	✓		
Rajkumar and Reena (2010)				✓	✓							✓
Shouman et al. (2011)				✓					✓	✓		
Alizadehsani et al. (2012)	✓			✓		✓			✓	✓		✓
Bhatla and Jyoti (2012)	✓			✓								
				✓	✓							
				✓	✓							
				✓								
Vijiyarani and Sudha (2013)				✓	✓	✓						
Alizadehsani et al. (2013)	✓			✓		✓			✓	✓		✓
Jabbar et al. (2013)				✓								✓
Ratnakar et al. (2013)				✓	✓							
Masethe and Masethe (2014)	✓			✓	✓							✓
Kaur (2014)				✓	✓				✓	✓		
Cinetha and Maheswari (2014)				✓								
Devi and Anto (2014)	✓			✓					✓	✓		
Venkatalakshmi and Shivsankar (2014)				✓	✓							
Methaila et al. (2014)				✓								
Kim et al. (2015)	✓			✓		✓			✓	✓		
Verma et al. (2016)				✓								
Verma and Srivastava (2016)				✓								
Kausar et al. (2016)				✓								
Baihaqi et al. (2016)				✓			✓		✓	✓	✓	
Joshi et al. (2016)				✓								
Malav et al. (2017)				✓	✓							
Samuel et al. (2017)		✓		✓		✓			✓	✓		
Al-Maqaleh and Abdullah (2017)	✓			✓		✓	✓					
Dekamin and Sheibatolhamdi (2017)	✓			✓			✓		✓	✓	✓	
Babu et al. (2017)												
Bhargava et al. (2017)	✓			✓	✓		✓			✓	✓	✓
Singh et al. (2018a)				✓						✓	✓	
Kulkarni et al. (2018)				✓		✓				✓	✓	
Shirwalkar et al. (2018)												
Singh et al. (2018b)	✓			✓								
Wadhawan (2018)				✓						✓	✓	
Kurian and Lakshmi (2018)				✓					✓	✓		

5. Breast Cancer Diseases

Breast cancer forms in the breast cells and can occur in men and women, but it is much more common in women. Survival rates of breast cancer have increased, and the number of deaths associated with this disease is due to factors such as earlier detection (MayoClinic, 2018a). The studied research on breast cancer is selected between 1997 and 2018. Among the 23 studied research, three articles have been studied in the form of a review paper, and 20 articles are associated with applications.

5.1 Literature Review

The utilization of data mining opens a new dimension to breast cancer prediction. Many data mining techniques are used for recognizing and obtaining valuable information from the clinical dataset (Srinivas et al., 2010). Researchers studied various ways to implement data mining in healthcare to reach a perfect prediction accuracy.

Table 9
The overall review of the data mining techniques in breast cancer diagnosis

Articles	Classification	Clustering			Outlier detection	Association			Model Evaluation	Model Comparison	Ensemble methods			Hybrid	Optimization				Fuzzy approach	Decision Making Algorithms	Feature Selection
		Partitioning methods				Frequent pattern mining															
		Unspecific	K-Means	Fuzzy cluster		Apriori	Mafia	Weighted Association			Unspecific	Voting	Bagging		Particle Swarm Optimization	Ant Colony Optimization	Principal Component Analysis	Sequential Minimal Optimization			
Burke et al. (1997)	✓								✓	✓											
Kuo et al. (2001)	✓								✓	✓											
Hassanien and Ali (2004)	✓								✓	✓											
Bellaachia and Guven (2006)	✓								✓	✓											
Chang and Liou (2008)	✓								✓	✓									✓		
Sarvestani et al. (2010)	✓								✓	✓											
Anunciaçao et al. (2010)	✓							✓											✓		
Einipour (2011)	✓								✓	✓						✓			✓	✓	
Ghassem Pour et al. (2012)	✓		✓						✓	✓		✓							✓		
Rajesh and Anand (2012)	✓								✓	✓											
Raad et al. (2012)	✓								✓	✓											
Hota (2013)	✓								✓	✓	✓										
Yadav et al. (2013)	✓		✓						✓	✓									✓		
Sumbaly et al. (2014)	✓								✓												
Senturk and Kara (2014)	✓								✓	✓											
Joshi et al. (2014)	✓								✓	✓		✓	✓	✓							
Majali et al. (2014)	✓					✓				✓											
Coutinho and das (2017)				✓					✓										✓	✓	
Chaurasia et al. (2018)	✓								✓	✓											
Cherif (2018)	✓	✓							✓	✓									✓		

Table 9 reveals complete information about the all implemented methods, and the concept of each paper is reviewed as follows:

Burke et al. (1997) compared the prediction accuracy of the TNM staging system with Artificial Neural Network statistical models. The result of this paper shows that the prediction of the Artificial Neural Network was more accurate than the TNM staging system. Kuo et al. (2001) made a new system for the classification of breast cancers by using Decision Tree technique. Prediction accuracy, sensitivity, and specificity are some of the evaluation models that are used to estimate the performance of the proposed system. Hassanien and Ali (2004) presented a Rough Set method for generating classification rules. This study showed that the theory of Rough Sets seems to be a useful tool. Bellaachia and Guven (2006) offered an analysis of the prediction of survivability rate of breast cancer patients using data mining technique namely the Naive Bayes, Back-Propagated Neural Network, and the C4.5 Decision Tree algorithms. The results illustrated that the C4.5 algorithm is better in comparing other techniques. Chang and Liou (2008) gave a comparative study for predicting breast cancers. They used a Decision Tree, Neural Network, Genetic algorithm, and Logistic Regression to diagnosis the breast cancer. The results showed that the Decision Tree has the lowest prediction accuracy and the Logistic Regression model had a higher accuracy rate. Sarvestani et al. (2010) evaluated several Neural Network formations. The performance of the statistical Neural Network structures, RBF Network, General Regression Neural Network, and Probabilistic Neural Network are tested and investigated for breast cancer diagnosis problem. Anunciaçao et al. (2010) explored the applicability of Decision Trees. In their work; first, they made different association rules by default and then made one questionnaire based on that rules and important defined factors which can be related to cancer disease. Einipour (2011) proposed a model by the combination of Fuzzy Systems and Ant Colony Optimization algorithm. Conclusions showed that the proposed approach would be capable of classifying cancer instances with a high accuracy rate. Ghassem Pour et al. (2012) compared a model-based data mining technique with a Neural Network classification technique. This paper shows that adding an ensemble approach can improve the results. They also used evaluations model to compare the performance of these models to others. Rajesh and Anand (2012) applied a C4.5 classification algorithm to breast cancer dataset to classify patients. This paper also compared the performance of the C4.5 algorithm with other classification techniques. Raad et al. (2012) Proposed a Neural Network approach especially the MLP, and the RBF. A detailed comparison between these two models showed that the constructed model from the RBF Neural Network is much more efficient than other models based. Hota (2013) applied various intelligent techniques including Artificial Neural Network, Support Vector Machine, Bayesian Network, and Decision Tree to classify a data that is related to breast cancer health care with 699 records. Experimental results revealed that the accuracy rate of the ensemble model is better than a single individual model.

Yadav et al. (2013) prescribed a procedure that uses Support Vector Machines and Decision Tree to classify 100 breast cancer patients into two classes. Results showed that Support Vector Machine gives the 98% prediction accuracy. Sumbaly et al. (2014) presented a Decision Tree data mining technique for early detection of breast cancer using Weka tool. Experimental results confirm the effectiveness of the proposed model. Senturk and Kara (2014) applied seven algorithms including KNN, Decision Tree, Naive Bayes, Logistic Regression, MLP, Discriminant Analysis and Support Vector Machine for diagnosis of breast cancers. Also, this paper used evaluations model like accuracy to measure the performance of the models. Joshi et al. (2014) compared various classification rules to predict the best classifier. Authors claimed that they used 47 classification algorithms for recognizing healthy people from patients. Their experimental results showed that the results of approximately 13 techniques within those 47 applied techniques were same. Majali et al. (2014) presented a system to diagnosis cancer using Frequent Pattern Mining growth algorithm. Also, this research used the Decision Tree algorithm to predict the possibility of cancer. Coutinho and das (2017) presented new hybrid fuzzy clustering algorithms. This research used three kinds of fuzzy clustering, and the results obtained with the proposed hybrid methods indicate that it is possible to increase the performance of the conventional fuzzy clustering algorithms. Chaurasia et al. (2018) used three popular data mining algorithms namely Naive Bayes, RBF and Decision Tree J48 to develop the prediction models using a large dataset and the obtained results indicated that the Naive Bayes performed the best with a classification accuracy of 97.36%. Cherif (2018) investigated a novel approach for classification of breast cancers. It selected the most reliable attributes and then weights them according to their level of reliability. This research speeds up the performance of KNN by clustering method. Kharya (2012) recommended a review paper about applying different classification techniques for diagnosis of breast cancers. This paper studied different methods including DT, Bayesian Network, Logistic Regression, SVM, Naive Bayes, Association Rule Mining, and Artificial Neural Network. Shrivastava et al. (2013) gave an overview of the use of data mining techniques on breast cancer data. They observed that the Neural Network and Decision Tree approach mostly used by various researchers to create a predictive model. Oskouei et al. (2017) reviewed several types of research works for diagnosis, treatment or prognosis breast cancers. They studied 125 references and based on the results of this study, most of the research works are concerned about comparing the accuracy rate of data mining various algorithms or techniques.

5.1.1. Classification Technique Analysis

Table 10 and Fig. 8 compare the classification methods used in breast cancer diagnosis. The Decision Tree and the Artificial Neural Network are used more than other methods.

Table 10
Comparison of the classification methods

Methods	Frequency of use	Usage percent
Decision Tree Algorithm	13	65%
Artificial Neural Network	10	50%
Bayesian Classifier	5	25%
Logistic Regression	3	15%
SVM	3	15%
KNN	2	10%
Fuzzy Set Approach	1	5%
Genetic Algorithm	1	5%
Rough Set Approach	0	0%
Associative Classification	0	0%
Rule Based	0	0%

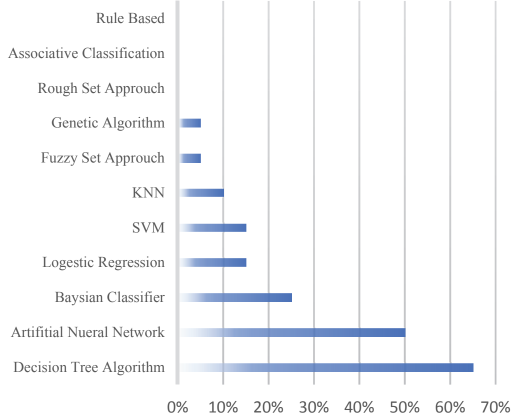

Fig. 8. Comparison of the classification methods

Table 11
The overall review of the classification methods

Article	Decision Tree Algorithm (DT)													Bayesian			Rule based			Artificial Neural Net-work								
	Unspecific	ID3	C5	CART	Rep tree	J48 (C4.5)	DT-Random (DTR)	LMT	DT-Forest (DTF)	DTR-Forest (DTRF)	DT stump	DT list	K-Nearest Neighbors (KNN)	Naive Bays	Bayesian Network	Support Vector Machine (SVM)	IF-THEN	RIPPER	Associative Classification	Unspecific	MLP	RBF	SOM	PNN	Genetic Algorithm	Rough Set Approach	Logistic Regression	Fuzzy Set Approach
Burke et al. (1997)																					✓	✓						
(Kuo et al., 2001)			✓																									
(Hassanien and Ali, 2004)		✓																										
(Bellaachia and Guven, 2006)	✓					✓								✓						✓	✓							
(Chang and Liou, 2008)	✓																			✓	✓				✓	✓	✓	✓
(Sarvestani et al., 2010)																					✓	✓	✓	✓	✓			
(Anunciaçao et al., 2010)	✓																											
(Einipour, 2011)																												✓
(Ghassem Pour et al., 2012)																				✓	✓							
(Rajesh and Anand, 2012)						✓																						
(Raad et al., 2012)																					✓	✓	✓					
(Hota, 2013)			✓	✓											✓	✓	✓			✓	✓							
(Yadav et al., 2013)	✓															✓												
(Sumbaly et al., 2014)						✓																						
(Senturk and Kara, 2014)	✓												✓	✓	✓	✓	✓				✓						✓	
(Joshi et al., 2014)					✓	✓	✓	✓		✓	✓	✓		✓	✓	✓					✓						✓	
(Majali et al., 2014)	✓																											
Coutinho and das (2017)																												
(Chaurasia et al., 2018)						✓								✓								✓						
(Cherif, 2018)													✓															

5.1.1.1. Decision Tree Method

Table 12 and Fig. 9 compare the Decision Tree models in classification. Unfortunately, some works of literature have not determined the name of the model used in the Decision Tree method.

Table 12
Comparison of the Decision Tree models

Models	Frequency of use	Usage percent
C4.5 (J 48)	5	25%
Unspecific Model	5	25%
C5	2	10%
ID3	1	5%
CART	1	5%
Rep tree	1	5%
DTR	1	5%
LMT	1	5%
DTRF	1	5%
DT stump	1	5%
DT list	1	5%
DTF	0	0%

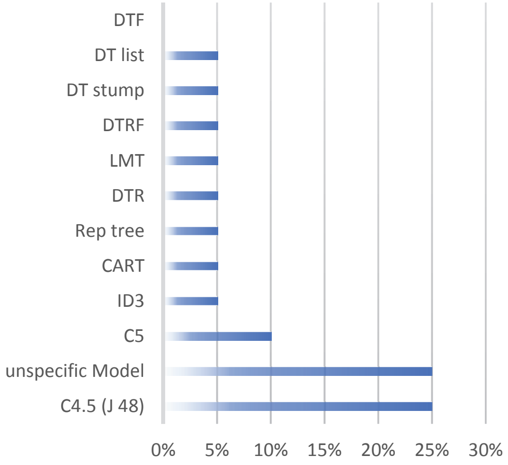

Fig. 9. Comparison of the Decision Tree models

5.1.1.2. Artificial Neural Network Method

Table 13 and Fig. 10 compare the Artificial Neural Network models. Unfortunately, some works of literature have not determined the name of the model used in the Artificial Neural Network method.

Table 13
Comparison of the Artificial Neural Network models

Models	Frequency of use	Usage percent
MLP	5	25%
Unspecific Model	4	20%
RBF	3	15%
SOM	1	5%
PNN	1	5%

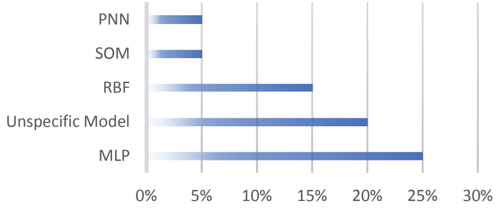

Fig. 10. Comparison of the Artificial Neural Network models

5.1.2. Clustering Technique Analysis

Table 14 and Fig. 11 compare different clustering methods in breast cancer diagnosis. Unfortunately, some works of literature have not determined the name of the method used in this technique.

Table 14
Comparison of the clustering methods

Models	Frequency of use	Usage percent
K-Means	2	10%
Unspecific	1	5%
Fuzzy cluster	1	5%

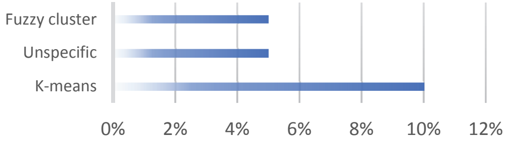

Fig. 11. Comparison of the clustering methods

5.1.3. Evaluation Technique Analysis

Table 15 and Fig. 12 compare the evaluation methods in breast cancer diagnosis. The prediction accuracy is obviously more common than other methods.

Table 15
Comparison of the evaluation methods

Models	Frequency of use	Usage percent
Accuracy	18	90%
Sensitivity	5	25%
Specificity	4	20%
Time Taken	2	10%
Confusion Matrix	1	5%
Cross Validation	1	5%
F-measure	1	5%
Precision	1	5%
Classification Chart	0	0%
Lift Chart	0	0%
Performance Plot	0	0%
ROC	0	0%

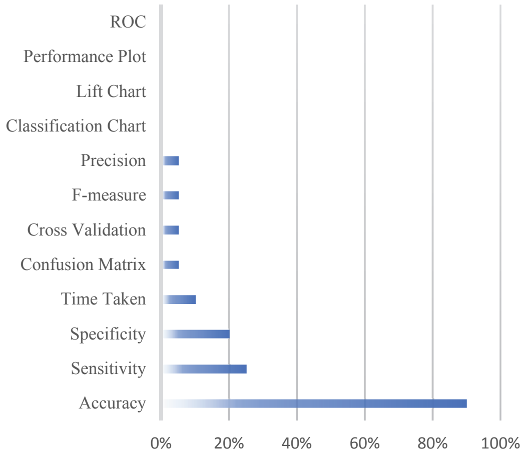

Fig. 12. Comparison of the evaluation methods

Table 16
The overall review of the evaluation method

Article	Evaluation the Model											
	Confusion Matrix	performance Plot	lift Chart	accuracy	Time Taken	ROC	F-measure	classification chart	specificity	Sensitivity	Precision	Cross validation
Burke et al. (1997)				✓								
Kuo et al. (2001)				✓					✓	✓		
Hassanien and Ali (2004)				✓								
Bellaachia and Guven (2006)				✓							✓	
Chang and Liou (2008)				✓								
Sarvestani et al. (2010)				✓	✓							
Anunciaçao et al. (2010)												
Einipour (2011)				✓								
Ghassem Pour et al. (2012)				✓					✓	✓		
Rajesh and Anand (2012)				✓								
Raad et al. (2012)				✓								
Hota (2013)				✓					✓	✓		
Yadav et al. (2013)	✓			✓								
Sumbaly et al. (2014)				✓								
Senturk and Kara (2014)				✓								
Joshi et al. (2014)				✓								
Majali et al. (2014)												
Coutinho and das (2017)				✓	✓							
Chaurasia et al. (2018)				✓					✓	✓		✓
Cherif (2018)				✓			✓					

6. Diabetes Disease

Diabetes mellitus refers to a group of diseases affecting the use of blood sugar or glucose in your body. Glucose is vital to your health, as it is an important energy source for the cells that makes up your muscles and tissues. Diabetes conditions include diabetes type1 and diabetes type2 (MayoClinic, 2018b). The studied research on diabetes mellitus is selected between 2013 and 2018. Among the 22 studied research, two articles have been studied in the form of a review paper and 20 articles are associated with applications.

6.1. Literature Review

The utilization of data mining reveals a new way to diabetes prediction. Many data mining techniques are used for identifying and collecting helpful knowledge from the clinical dataset (Srinivas et al., 2010). Researchers studied different approaches to implement data mining in healthcare to reach an excellent prediction accuracy.

Table 17
The overall review of the data mining techniques in diabetes diagnosis

Articles	Classification	Clustering			Outlier detection	Association			Model Evaluation	Model Comparison	Ensemble methods			Hybrid	Optimization				Fuzzy approach	Decision Making Algorithms	Feature Selection
		Partitioning methods				Frequent pattern mining															
		Unspecific	K-Means	Fuzzy cluster		Apriori	Mafia	Weighted Association			Unspecific	Voting	Bagging		Particle Swarm Optimization	Ant Colony Optimization	Principal Component Analysis	Sequential Minimal Optimization			
Meng et al. (2013)	✓								✓	✓											
Krati Saxena et al. (2014)	✓								✓												
Kandhasamy and Balamurali (2015)	✓								✓	✓											
kumar Dewangan and Agrawal (2015)	✓								✓	✓	✓									✓	
Santhanam and Padmavathi (2015)	✓		✓						✓	✓										✓	
Prajwala (2015)	✓								✓	✓											
Thirumal and Nagarajan (2015)	✓		✓						✓	✓										✓	
Perveen et al. (2016)	✓								✓	✓			✓		✓						
Shukla and Arora (2016)	✓								✓	✓					✓						
Meza-Palacios et al. (2016)	✓								✓	✓											✓
Garg et al. (2017)	✓								✓	✓									✓		
Xu et al. (2017)	✓		✓						✓	✓										✓	
Nilashi et al. (2017)	✓			✓					✓	✓										✓	
Khaleel et al. (2017)	✓								✓	✓		✓								✓	
Sambyal et al. (2018)	✓								✓	✓					✓					✓	
Lakshmi et al. (2018)	✓		✓						✓	✓										✓	
Das et al. (2018)	✓								✓	✓											
Wu et al. (2018)	✓		✓						✓	✓										✓	
Sisodia and Sisodia (2018)	✓								✓	✓											
Patil and Tamane (2018)	✓	✓							✓	✓										✓	

Table 17 exposes perfect information about the all implemented ways and methods, and the concept of each paper is reviewed as follows:

Meng et al. (2013) compared the performance of Artificial Neural Networks, Logistic Regression and Decision Tree C5 models for predicting diabetes. The results indicated that the C5 Decision Tree model performed best on classification accuracy. Krati Saxena et al. (2014) diagnosed diabetes mellitus using K-Nearest Neighbor algorithm with MATLAB software. The result is showing that as the value of K increases, accuracy rate and error rate will also increase. Kandhasamy and Balamurali (2015) compared machine learning classifiers namely J48 Decision Tree, KNN, and Random Forest, and SVM to classify patients with diabetes mellitus using eight essential attributes. kumar Dewangan and Agrawal (2015) attempted to make an ensemble hybrid model by combining Bayesian classification and multilayer perceptron techniques. The results show that hybrid models give higher accuracy than the individuals' model. Santhanam and Padmavathi (2015) used the K-Means method to remove the noisy data and Genetic algorithms to find the optimal set of features with Support Vector Machine as a classifier for classification. Prajwala (2015) discussed two classification algorithms namely Decision Trees and Random Forests considering 256 data samples. The experimental results show that the redistribution error rate of the Random Forest is less than the Decision Tree. Thirumal and Nagarajan (2015) proposed research that several data mining algorithms such as Naive Bayes, Decision Trees, K-Nearest Neighbor and Support Vector Machine algorithm have been discussed. The experimental results show that K-Nearest Neighbor provides lower accuracy compared to other algorithms.

Perveen et al. (2016) followed the Adaboost and Bagging ensemble techniques using the J48 Decision Tree as a base learner to classify patients with diabetes mellitus. This paper concluded that the overall performance of the Adaboost ensemble method is better than the bagging method. Shukla and Arora (2016) used Random Forest tree alongside information mining procedure scaled conjugate gradient to predict diabetes mellitus. This paper incorporates calculations of Random Forest tree and scaled conjugate gradient. diabetic is a life-threatening complication. Meza-Palacios et al. (2016) proposed the development of a fuzzy expert system that was a new and innovative proposal to help doctors. Garg et al. (2017) showed the comparison of different classification algorithms using Weka tool. These classification algorithms include Naive Bayes, Bayes Network, Decision Tree J48, Sequential Minimal Optimization (SMO)classifier, and Random Forest. The experimental results propose that SMO classifier has the best performance. Xu et al. (2017) proposed a prediction model based on a Random Forest. This method can significantly reduce the risk of disease by digging out a clear and understandable model for type2 diabetes from a medical database. The results show that using Random Forest can cause a better prediction accuracy. Nilashi et al. (2017) suggested a new system for diabetes prediction using clustering, noise removal, and prediction techniques. This research uses CART method to generate the fuzzy rules. Also, EM and PCA were used for clustering.

Khaleel et al. (2017) used One-Attribute-Rule algorithm to adjust the attributes weights and propose a new classification algorithm that improves the accuracy of the K-Nearest Neighbor algorithm. Sambyal et al. (2018) compared six different data mining algorithms. This system is trained and tested in Microsoft Azure, and the brilliant created system has been deployed as a web service using the python language. Lakshmi et al. (2018) introduced system use the Decision Tree and K-Nearest Neighbor algorithms, but there is not any information about the results. Das et al. (2018) studied Decision Tree J48 and Naive Bayesian techniques. This research will assist to propose a quicker and more efficient method for diagnosis of diabetes. Wu et al. (2018) recommended a hybrid model based on data mining techniques. They used the improved K-Means algorithm and the Logistic Regression algorithm that achieve higher accuracy of prediction. Sisodia and Sisodia (2018) designed a model which can prognosticate the likelihood of diabetes with maximum accuracy. This research is used three machine learning classification algorithms namely Decision Tree, Support Vector Machine algorithm and Naive Bayes to detect diabetes at early stages. Patil and Tamane (2018) used the combination of techniques such as feature selection with K-Nearest Neighbor and Naive Bayes approach to developing a predictive model. Joshi and Alehegn (2017) studied and reviewed various data mining techniques such as K-Nearest Neighbor, Naive Bayes, Random Forest, and J48. Rani and Kautish (2018) reviewed the most cited research papers of highest journals to investigate data mining techniques which are generally used to predict some chronic disease like diabetes.

6.2.1. Classification Technique analysis

Table 18 and Fig. 13 compare the classification methods in diabetes diagnosis. The Decision Tree, Bayesian Classifier, and K-Nearest Neighbors are more common than the other methods.

Table 18
Comparison of the classification methods

Methods	Frequency of use	Usage percent
Decision Tree Algorithm	14	70%
Bayesian Classifier	6	30%
KNN	6	30%
Artificial Neural Network	5	25%
SVM	5	25%
Logistic Regression	3	15%
Genetic Algorithm	2	10%
Fuzzy Set Approach	2	10%
Rule Based	0	0%
Associative Classification	0	0%
Rough Set Approach	0	0%

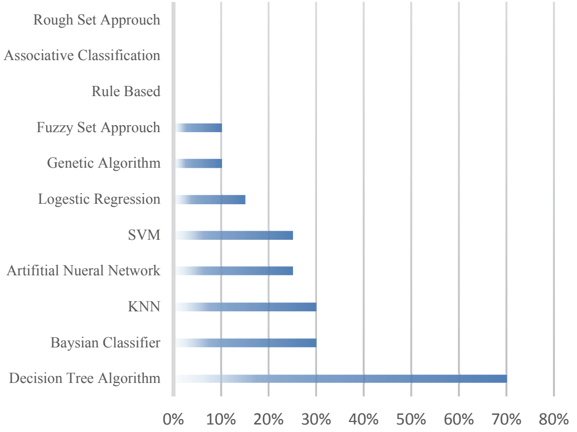

Fig. 13. Comparison of the classification methods

Table 19
The overall review of the classification methods

Article	Decision Tree Algorithm (DT)												K-Nearest Neighbors (KNN)	Bayesian		Support Vector Machine (SVM)	Rule based		Associative Classification	Artificial Neural Network					Genetic Algorithm	Rough Set Approach	Logistic Regression	Fuzzy Set Approach
	Unspecific	ID3	C5	CART	Rep tree	J48 (C4.5)	DT-Random (DTR)	LMT	DT-Forest (DTF)	DTR-Forest (DTRF)	DT stump	DT list		Naive Bays	Bayesian Network		IF-THEN	RIPPER		Unspecific	MLP	RBF	SOM	PNN				
Meng et al. (2013)			✓																	✓							✓	
Krati Saxena et al. (2014)													✓															
Kandhasamy and Balamurali (2015)						✓				✓			✓			✓												
kumar Dewangan & Agrawal (2015)						✓				✓					✓						✓							
Santhanam and Padmavathi (2015)																✓									✓			
Prajwala (2015)		✓								✓																		
Thirumal and Nagarajan (2015)						✓							✓	✓		✓												
Perveen et al. (2016)						✓																						
Shukla and Arora (2016)										✓										✓								
Meza-Palacios et al. (2016)																												✓
Garg et al. (2017)						✓				✓					✓						✓							
Xu et al. (2017)										✓							✓											
Nilashi et al. (2017)				✓																								✓
Khaleel et al. (2017)													✓				✓											
Sambyal et al. (2018)	✓								✓							✓				✓							✓	
Lakshmi et al. (2018)						✓							✓															
Das et al. (2018)						✓								✓														
Wu et al. (2018)																											✓	
Sisodia and Sisodia (2018)	✓													✓		✓												
Patil and Tamane (2018)													✓	✓											✓			

6.2.1.1. Decision Tree Method

Table 20 and Fig. 14 compare the Decision Tree classification models. Unfortunately, some works of literature have not determined the name of the model used in the Decision Tree method.

Table 20

Comparison of the Decision Tree models

Models	Frequency of use	Usage percent
C4.5(J48)	7	35%
DTRF	6	30%
unspecific Model	2	10%
ID3	1	5%
C5	1	5%
CART	1	5%
DTF	1	5%
Rep tree	0	0%
DTR	0	0%
LMT	0	0%
DT stump	0	0%
DT list	0	0%

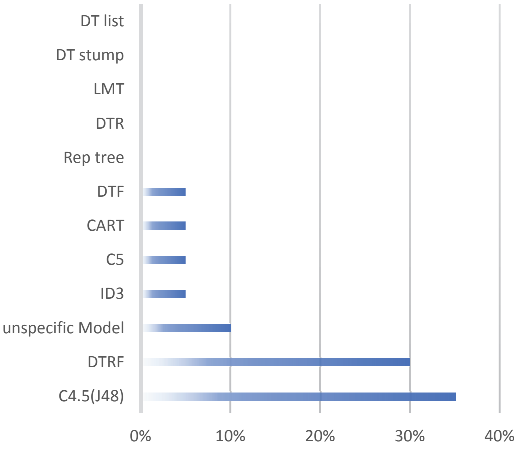

Fig. 14. Comparison of the Decision Tree models

6.2.1.2. *Artificial Neural Network Method*

Table 21 and Fig. 15 compare the Artificial Neural Network classification models. Unfortunately, some works of literature have not determined the name of the model used in this method.

Table 21

Comparison of the Artificial Neural Network models

Models	Frequency of use	Usage percent
Unspecific Model	3	15%
MLP	2	10%
RBF	0	0%
SOM	0	0%
PNN	0	0%

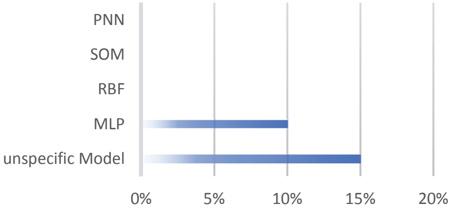

Fig. 15. Comparison of the Artificial Neural Network models

6.2.2. *Clustering Technique Analysis*

Table 22 and Fig. 16 compare the clustering methods in in diabetes diagnosis. Unfortunately, some works of literature have not determined the name of the model used in this technique.

Table 22

Comparison of the clustering methods

Models	Frequency of use	Usage percent
K-Means	5	25%
EM & PCA	1	5%
Unspecific	0	0%

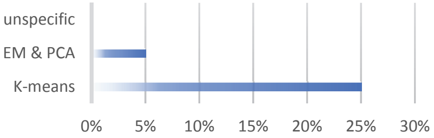

Fig. 16. Comparison of the clustering methods

6.2.3. *Evaluation Technique analysis*

Table 23 and Fig. 17 compare the evaluation methods in diabetes diagnosis. The prediction accuracy is more common than other methods.

Table 23
Comparison of the evaluation methods

Models	Frequency of use	Usage percent
Accuracy	18	90%
Sensitivity	12	60%
Specificity	8	40%
Confusion Matrix	5	25%
Precision	5	25%
ROC	4	20%
F-measure	3	15%
Cross Validation	3	15%
Time Taken	2	10%
Performance Plot	1	5%
Lift Chart	0	0%
Classification Chart	0	0%

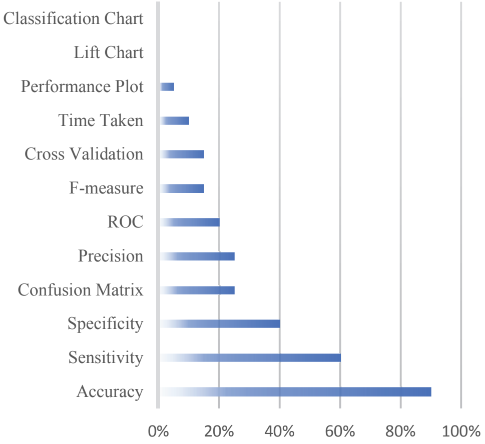

Fig. 17. Comparison of the evaluation methods

Table 24
The overall review of the evaluation methods

Article	Evaluation the Model											
	Confusion Matrix	Performance Plot	Lift Chart	Accuracy	Time Taken	ROC	F-measure	Classification Chart	Specificity	Sensitivity	Precision	Cross validation
Meng et al. (2013)				✓					✓	✓		
Krati Saxena et al. (2014)				✓					✓	✓		
Kandhasamy and Balamurali (2015)				✓					✓	✓		
kumar Dewangan and Agrawal (2015)				✓					✓	✓		
Santhanam and Padmavathi (2015)				✓					✓	✓		
Prajwala (2015)	✓				✓							
Thirumal and Nagarajan (2015)	✓			✓						✓	✓	✓
Perveen et al. (2016)				✓		✓						
Shukla and Arora (2016)		✓		✓					✓	✓		
Meza-Palacios et al. (2016)				✓								
Garg et al. (2017)				✓							✓	✓
Xu et al. (2017)				✓					✓	✓		
Nilashi et al. (2017)				✓								
Khaleel et al. (2017)				✓			✓		✓	✓	✓	
Sambyal et al. (2018)	✓			✓		✓						
Lakshmi et al. (2018)				✓								
Das et al. (2018)				✓	✓							
Wu et al. (2018)	✓			✓		✓	✓			✓	✓	✓
Sisodia and Sisodia (2018)	✓			✓		✓	✓				✓	
Patil and Tamane (2018)				✓								

7. Conclusion

This paper reviewed the predictive data mining approaches in heart disease, breast cancer, and diabetes diagnosis. The number of 168 articles associated with the implementation of data mining for medical diagnosis between 1997 and 2018 were identified. After the initial investigations, 85 empirical studies were selected for the final review. The obtained results reveal that a significant number of studies have used classification technique. Also, researchers have achieved better prediction accuracy results with hybrid and ensemble models. Furthermore, in most research, the performance of different data mining models is compared to each other. Comparison of the different clustering methods has appeared that K-Means clustering is the most common clustering method. Additionally, the Decision Tree algorithm, Bayesian Network, and Neural Network are three widely used classification methods based on the comparison of the different classification methods. Moreover, the most frequently used Decision Tree models are CART and C4.5, and for evaluating and comparing the models, prediction accuracy is widely used.

This paper recommends using large datasets to guarantee the performance of the prediction model. Further, model performance improvement techniques such as Ant Colony Optimization Algorithms and Particle Swarm Optimization are very little used, and it is better to use these techniques more. As mentioned, hybrid and ensemble models give better prediction accuracy results, so using these models are recommended in the future studies. Therefore, with regards to the mentioned notes about the research gaps and the use of predictive data mining approaches in medical diagnosis, new studies can be reached in this field.

References

Al-Maqaleh, B. M., & Abdullah, A. M. G. (2017). Intelligent predictive system using classification techniques for heart disease diagnosis. *International Journal of Computer Science Engineering (IJCSE)*, *6*(6), 145-151.

Alizadehsani, R., Habibi, J., Bahadorian, B., Mashayekhi, H., Ghandeharioun, A., Boghrati, R., & Sani, Z. A. (2012). Diagnosis of coronary arteries stenosis using data mining. *Journal of medical signals and sensors*, *2*(3), 153-159.

Alizadehsani, R., Habibi, J., Hosseini, M. J., Mashayekhi, H., Boghrati, R., Ghandeharioun, A., ... & Sani, Z. A. (2013). A data mining approach for diagnosis of coronary artery disease. *Computer methods and programs in biomedicine*, *111*(1), 52-61.

Anunciaçao, O., Gomes, B. C., Vinga, S., Gaspar, J., Oliveira, A. L., & Rueff, J. (2010). A data mining approach for the detection of high-risk breast cancer groups. In *Advances in Bioinformatics* (pp. 43-51). Springer, Berlin, Heidelberg.

Babu, S., Vivek, E. M., Famina, K. P., Fida, K., Aswathi, P., Shanid, M., & Hena, M. (2017, April). Heart disease diagnosis using data mining technique. In *Electronics, Communication and Aerospace Technology (ICECA), 2017 International conference of* (Vol. 1, pp. 750-753). IEEE.

Baihaqi, W. M., Setiawan, N. A., & Ardiyanto, I. (2016, August). Rule extraction for fuzzy expert system to diagnose coronary artery disease. In *Information Technology, Information Systems and Electrical Engineering (ICITISEE), International Conference on* (pp. 136-141). IEEE.

Bellaachia, A., & Guven, E. (2006). Predicting breast cancer survivability using data mining techniques. *Age*, *58*(13), 10-110.

Bhargava, N., Dayma, S., Kumar, A., & Singh, P. (2017, January). An approach for classification using simple CART algorithm in WEKA. In *Intelligent Systems and Control (ISCO), 2017 11th International Conference on* (pp. 212-216). IEEE.

Bhatla, N., & Jyoti, K. (2012). An analysis of heart disease prediction using different data mining techniques. *International Journal of Engineering*, *1*(8), 1-4.

Burke, H. B., Goodman, P. H., Rosen, D. B., Henson, D. E., Weinstein, J. N., Harrell Jr, F. E., ... & Bostwick, D. G. (1997). Artificial neural networks improve the accuracy of cancer survival prediction. *Cancer*, *79*(4), 857-862.

Chang, W. P., & Liou, D. M. (2008). Comparison of three data mining techniques with genetic algorithm in the analysis of breast cancer data. *J Telemed Telecare*, *9*(1), 26.

Chaurasia, V., Pal, S., & Tiwari, B. B. (2018). Prediction of benign and malignant breast cancer using data mining techniques. *Journal of Algorithms & Computational Technology*, *12*(2), 119-126.

Cherif, W. (2018). Optimization of K-NN algorithm by clustering and reliability coefficients: application to breast-cancer diagnosis. *Procedia Computer Science*, *127*, 293-299.

Cinetha, K., & Maheswari, P. U. (2014). Decision support system for precluding coronary heart disease (CHD) using fuzzy logic. *IJCST*, *2*(2), 2347-857.

Coutinho, P. H. S., & Thiago, P. (2017, November). Proposal of new hybrid fuzzy clustering algorithms—Application to breast cancer dataset. In *Computational Intelligence (LA-CCI), 2017 IEEE Latin American Conference on* (pp. 1-6). IEEE.

Das, H., Naik, B., & Behera, H. S. (2018). Classification of diabetes mellitus disease (DMD): A data mining (DM) approach. In *Progress in Computing, Analytics and Networking* (pp. 539-549). Springer, Singapore.

Das, R., Turkoglu, I., & Sengur, A. (2009). Effective diagnosis of heart disease through neural networks ensembles. *Expert systems with applications*, *36*(4), 7675-7680.

Dekamin, A., & Sheibatolhamdi, A. (2017). A data mining approach for coronary artery disease prediction in Iran. *Journal of Advanced Medical Sciences and Applied Technologies*, *3*(1), 29-38.

Devi, Y. N., & Anto, S. (2014). An evolutionary-fuzzy expert system for the diagnosis of coronary artery disease. *IJARCET), ISSN*, 2278-1323.

Einipour, A. (2011). A fuzzy-ACO method for detect breast cancer. *Global journal of health science*, *3*(2), 195-199.

Pour, S. G., Mc Leod, P., Verma, B., & Maeder, A. (2012). Comparing data mining with ensemble classification of breast cancer masses in digital mammograms. In *Second Australian Workshop on Artificial Intelligence in Health: AIH, 55-63.*

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

Hassanien, A. E., & Ali, J. M. (2004). Rough set approach for generation of classification rules of breast cancer data. *Informatica*, *15*(1), 23-38.

Hota, H. S. (2013). Diagnosis of breast cancer using intelligent techniques. *International Journal of Emerging Science and Engineering (IJESE)*, *1*(3), 45-53.

Jabbar, M. A., Deekshatulu, B. L., & Chandra, P. (2013). Heart disease classification using nearest neighbor classifier with feature subset selection. *Anale. Seria Informatica*, *11, 47-54*.

Joshi, A., Dangra, J., & Rawat, M. (2016). A decision tree based classification technique for accurate heart disease classification and prediction. *Int J Technol Res Manag*, *3*, 1-4.

Joshi, J., Doshi, R., & Patel, J. (2014). Diagnosis and prognosis breast cancer using classification rules. *International Journal of Engineering Research and General Science*, *2*(6), 315-323.

Joshi, R., & Alehegn, M. (2017). Analysis and prediction of diabetes diseases using machine learning algorithm: Ensemble approach. *IRJET, 4*(10), 426-435.

Kadi, I., Idri, A., & Fernandez-Aleman, J. L. (2017). Knowledge discovery in cardiology: A systematic literature review. *International journal of medical informatics*, *97*, 12-32.

Kandhasamy, J. P., & Balamurali, S. (2015). Performance analysis of classifier models to predict diabetes mellitus. *Procedia Computer Science*, *47*, 45-51.

Kaur, L. (2014). Predicting heart disease symptoms using fuzzy C-means clustering. *published in International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume*, *3*(12)*, 4232-4235*.

Kausar, N., Abdullah, A., Samir, B. B., Palaniappan, S., AlGhamdi, B. S., & Dey, N. (2016). Ensemble clustering algorithm with supervised classification of clinical data for early diagnosis of coronary artery disease. *Journal of Medical Imaging and Health Informatics*, *6*(1), 78-87.

Khaleel, A. H., Al-Suhail, G. A., & Hussan, B. M. (2017). A weighted voting of k-nearest neighbor algorithm for diabetes mellitus, *6*(1), 43-51.

Kharya, S. (2012). Using data mining techniques for diagnosis and prognosis of cancer disease. *arXiv preprint arXiv:1205.1923, 2*(2), 55-66.

Kim, J., Lee, J., & Lee, Y. (2015). Data-mining-based coronary heart disease risk prediction model using fuzzy logic and decision tree. *Healthcare informatics research*, *21*(3), 167-174.

Koh, H. C., & Tan, G. (2011). Data mining applications in healthcare. *Journal of healthcare information management*, *19*(2), 64-72.

Krati Saxena, D., Khan, Z., & Singh, S. (2014). Diagnosis of diabetes mellitus using k-nearest neighbor algorithm. *International Journal of Computer Science Trends and Technology (IJCST)*.

Kulkarni, S., Bhat, C. D., Patil, D., & Dara, J. (2018). Heart disease classification: A case study using machine learning and data mining. *International journal of computer trends and technology, 2*(4), 36-43.

Kumar Dewangan, A., & Agrawal, P. (2015). Classification of diabetes mellitus using machine learning techniques. *Int. J. Eng. Appl. Sci*, *2*(5), 145-148.

Kumari, M., & Godara, S. (2011). Review of data mining classification models in cardiovascular disease diagnosis. *International Journal of Computer Science and Technology*, *2*(2), 304-305.

Kuo, W. J., Chang, R. F., Chen, D. R., & Lee, C. C. (2001). Data mining with decision trees for diagnosis of breast tumor in medical ultrasonic images. *Breast cancer research and treatment*, *66*(1), 51-57.

Kurian, R. A., & Lakshmi, K. S. (2018). An ensemble classifier for the prediction of heart disease. *International Journal of Scientific Research in Computer Science, 3*(6), 25-31.

Lakshmi, K., Ahmed, D. I., & Kumar, G. S. (2018). A smart clinical decision support system to predict diabetes disease using classification techniques. *IJSRSET, 4*(1), 1520-1522.

Majali, J., Niranjan, R., Phatak, V., & Tadakhe, O. (2014). Data mining techniques for diagnosis and prognosis of breast cancer. *International Journal of Computer Science and Information Technologies (IJCSIT)*, *5*(5), 6487-6490.

Malav, A., Kadam, K., & Kamat, P. (2017). Prediction of heart disease using k-means and artificial neural network as Hybrid Approach to Improve Accuracy. *International Journal of Engineering and Technology*, *9*(4), 3081-3085.

Masethe, H. D., & Masethe, M. A. (2014, October). Prediction of heart disease using classification algorithms. In *Proceedings of the world Congress on Engineering and computer Science* (Vol. 2, pp. 22-24).

MayoClinic. (2018, November). Breast cancer. Retrieved from https://www.mayoclinic.org/diseases-conditions/breast-cancer/symptoms-causes/syc-20352470?utm_source=Google&utm_medium=abstract&utm_content=Breast-cancer&utm_campaign=Knowledge-panel

MayoClinic. (2018, August). Diabetes (diseases and conditions). Retrieved from https://www.mayoclinic.org/diseases-conditions/diabetes/symptoms-causes/syc-20371444

Meng, X. H., Huang, Y. X., Rao, D. P., Zhang, Q., & Liu, Q. (2013). Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. *The Kaohsiung journal of medical sciences*, *29*(2), 93-99.

Methaila, A., Kansal, P., Arya, H., & Kumar, P. (2014). Early heart disease prediction using data mining techniques. *Computer Science & Information Technology Journal*, 53-59.

Meza-Palacios, R., Aguilar-Lasserre, A. A., Ureña-Bogarín, E. L., Vázquez-Rodríguez, C. F., Posada-Gómez, R., & Trujillo-Mata, A. (2017). Development of a fuzzy expert system for the nephropathy control assessment in patients with type 2 diabetes mellitus. *Expert Systems with Applications*, *72*, 335-343.

Nilashi, M., bin Ibrahim, O., Ahmadi, H., & Shahmoradi, L. (2017). An analytical method for diseases prediction using machine learning techniques. *Computers & Chemical Engineering*, *106*, 212-223.

Oracle Database. (2008, July). Oracle data warehousing and business intelligence (data mining concepts). Retrieved from https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/toc.htm.

Oskouei, R. J., Kor, N. M., & Maleki, S. A. (2017). Data mining and medical world: breast cancers' diagnosis, treatment, prognosis and challenges. *American journal of cancer research*, *7*(3), 610.

Palaniappan, S., & Awang, R. (2008, March). Intelligent heart disease prediction system using data mining techniques. In *Computer Systems and Applications, 2008. AICCSA 2008. IEEE/ACS International Conference on* (pp. 108-115). IEEE.

Patel, A., Gandhi, S., Shetty, S., & Tekwani, B. (2017). Heart disease prediction using data mining. *International Research Journal of Engineering and Technology*, *4*(1), 1705-1707.

Patil, R. N., & Tamane, S. C. (2018). Upgrading the performance of KNN and naïve bayes in diabetes detection with genetic algorithm for feature selection. *International Journal of Scientific Research in Computer Science, 3*(1), 1371-1381.

Perveen, S., Shahbaz, M., Guergachi, A., & Keshavjee, K. (2016). Performance analysis of data mining classification techniques to predict diabetes. *Procedia Computer Science*, *82*, 115-121.

Prajwala, T. R. (2015). A comparative study on decision tree and random forest using R tool. *International journal of advanced research in computer and communication engineering*, *4*(1), 196-199.

Raad, A., Kalakech, A., & Ayache, M. (2012). Breast cancer classification using neural network approach: MLP and RBF. *networks*, *7*(8), 15-19.

Rajesh, K., & Anand, S. (2012). Analysis of SEER dataset for breast cancer diagnosis using C4.5 classification algorithm. *International Journal of Advanced Research in Computer and Communication Engineering*, *1*(2), 2278-1021.

Rajkumar, A., & Reena, G. S. (2010). Diagnosis of heart disease using datamining algorithm. *Global journal of computer science and technology*, *10*(10), 38-43.

Rani, S., & Kautish, S. (2018). Application of data mining techniques for prediction of diabetes-A review. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 3*(3), 1996-2004.

Ratnakar, S., Rajeswari, K., & Jacob, R. (2013). Prediction of heart disease using genetic algorithm for selection of optimal reduced set of attributes. *International Journal of Advanced Computational Engineering and Networking*, *1*(2), 51-55.

Sambyal, R. S., Javid, T., & Bansal, A. (2018). Performance analysis of data mining classification algorithms to Predict diabetes. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 4*(1), 56-63.

Samuel, O. W., Asogbon, G. M., Sangaiah, A. K., Fang, P., & Li, G. (2017). An integrated decision support system based on ANN and Fuzzy_AHP for heart failure risk prediction. *Expert Systems with Applications*, *68*, 163-172.

Santhanam, T., & Padmavathi, M. S. (2015). Application of K-means and genetic algorithms for dimension reduction by integrating SVM for diabetes diagnosis. *Procedia Computer Science*, *47*, 76-83.

Sarvestani, A. S., Safavi, A. A., Parandeh, N. M., & Salehi, M. (2010, October). Predicting breast cancer survivability using data mining techniques. In *Software technology and Engineering (ICSTE), 2010 2nd international Conference on* (Vol. 2, pp. V2-227). IEEE.

Senturk, Z. K., & Kara, R. (2014). Breast cancer diagnosis via data mining: performance analysis of seven different algorithms. *Computer Science & Engineering*, *4*(1), 35-46.

Shirwalkar, N., Gursalkar, S., Tak, T., & Kalshetti, A. (2018). Human heart disease prediction system using data mining techniques. International Journal of Innovations & Advancement in Computer Science, 7(3), 357-360.

Shouman, M., Turner, T., & Stocker, R. (2011, December). Using decision tree for diagnosing heart disease patients. In *Proceedings of the Ninth Australasian Data Mining Conference-Volume 121* (pp. 23-30). Australian Computer Society, Inc.

Shouman, M., Turner, T., & Stocker, R. (2012, March). Using data mining techniques in heart disease diagnosis and treatment. In *Electronics, Communications and Computers (JEC-ECC), 2012 Japan-Egypt Conference on* (pp. 173-177). IEEE.

Shrivastava, S. S., Sant, A., & Aharwal, R. P. (2013). An overview on data mining approach on breast cancer data. *International Journal of Advanced Computer Research*, *3*(4), 256-262.

Shukla, N., & Arora, M. (2016). Prediction of diabetes using neural network & random forest tree. *International Journal of Computer Sciences and Engineering*, *4*, 101-104.

Singh, N., Firozpur, P., & Jindal, S. (2018). Heart disease prediction system using hybrid technique of data mining algorithms. *International Journal of Advance Research, Ideas and Innovations in Technology, 4*(2), 982-987.

Singh, P., Singh, S., & Pandi-Jain, G. S. (2018). effective heart disease prediction system using data mining techniques. *International journal of nanomedicine*, *13*, 121-124.

Sisodia, D., & Sisodia, D. S. (2018). Prediction of diabetes using classification algorithms. *Procedia Computer Science*, *132*, 1578-1585.

Srinivas, K., Rani, B. K., & Govrdhan, A. (2010). Applications of data mining techniques in healthcare and prediction of heart attacks. *International Journal on Computer Science and Engineering (IJCSE)*, *2*(02), 250-255.

Sumbaly, R., Vishnusri, N., & Jeyalatha, S. (2014). Diagnosis of breast cancer using decision tree data mining technique. *International Journal of Computer Applications*, *98*(10).

Thenmozhi, K., & Deepika, P. (2014). Heart disease prediction using classification with different decision tree techniques. *International Journal of Engineering Research and General Science*, *2*(6), 6-11.

Thirumal, P. C., & Nagarajan, N. (2015). Utilization of data mining techniques for diagnosis of diabetes mellitus-a case study. *ARPN Journal of Engineering and Applied Science*, *10*(1), 8-13.

Tu, M. C., Shin, D., & Shin, D. (2009, October). Effective diagnosis of heart disease through bagging approach. In *Biomedical Engineering and Informatics, 2009. BMEI'09. 2nd International Conference on* (pp. 1-4). IEEE.

Venkatalakshmi, B., & Shivsankar, M. V. (2014). Heart disease diagnosis using predictive data mining. *International Journal of Innovative Research in Science, Engineering and Technology*, *3*(3), 1873-7.

Verma, L., Srivastava, S., & Negi, P. C. (2016). A hybrid data mining model to predict coronary artery disease cases using non-invasive clinical data. *Journal of medical systems*, *40*(7), 178-185.

Verma, L., & Srivastava, S. (2016). A data mining model for coronary artery disease detection using noninvasive clinical parameters. *Indian Journal of Science and Technology, 9*(48), 1-6.

Vijiyarani, S., & Sudha, S. (2013). An efficient classification tree technique for heart disease prediction. In *International Conference on Research Trends in Computer Technologies (ICRTCT-2013) Proceedings published in International Journal of Computer Applications (IJCA) (0975–8887)* (Vol. 201).

Wadhawan, R. (2018). Prediction of coronary heart disease using Apriori algorithm with data mining classification. *International Journal of Research in Science and Technology*, *3*(1), 1-15.

Wu, H., Yang, S., Huang, Z., He, J., & Wang, X. (2018). Type 2 diabetes mellitus prediction model based on data mining. *Informatics in Medicine Unlocked*, *10*, 100-107.

Xu, W., Zhang, J., Zhang, Q., & Wei, X. (2017, February). Risk prediction of type II diabetes based on random forest model. In *Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB), 2017 Third International Conference on* (pp. 382-386). IEEE.

Yadav, R., Khan, Z., & Saxena, H. (2013). Chemotherapy prediction of cancer patient by using data mining techniques. *International Journal of Computer Applications*, *76*(10), 28-31.