Expert Systems with Applications 41 (2014) 4434–4463

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Review

Knowledge discovery in medicine: Current issue and future trend

Nura Esfandiari [a], Mohammad Reza Babavalian [a,*], Amir-Masoud Eftekhari Moghadam [a], Vahid Kashani Tabar [b]

[a] Faculty of Computer and Information Technology, Qazvin Branch, Islamic Azad University, Qazvin, Iran
[b] Trauma Research Center, Kashan University of Medical Sciences, Kashan, Iran

ARTICLE INFO

Keywords:
Data mining application
Medical data mining
Medicine
Disease
Data mining algorithms

ABSTRACT

Data mining is a powerful method to extract knowledge from data. Raw data faces various challenges that make traditional method improper for knowledge extraction. Data mining is supposed to be able to handle various data types in all formats. Relevance of this paper is emphasized by the fact that data mining is an object of research in different areas. In this paper, we review previous works in the context of knowledge extraction from medical data. The main idea in this paper is to describe key papers and provide some guidelines to help medical practitioners. Medical data mining is a multidisciplinary field with contribution of medicine and data mining. Due to this fact, previous works should be classified to cover all users' requirements from various fields. Because of this, we have studied papers with the aim of extracting knowledge from structural medical data published between 1999 and 2013. We clarify medical data mining and its main goals. Therefore, each paper is studied based on the six medical tasks: screening, diagnosis, treatment, prognosis, monitoring and management. In each task, five data mining approaches are considered: classification, regression, clustering, association and hybrid. At the end of each task, a brief summarization and discussion are stated. A standard framework according to CRISP-DM is additionally adapted to manage all activities. As a discussion, current issue and future trend are mentioned. The amount of the works published in this scope is substantial and it is impossible to discuss all of them on a single work. We hope this paper will make it possible to explore previous works and identify interesting areas for future research.

1. Introduction

The growth of information storage technology has led to generate huge amount of raw data that considers two aspects. These aspects are algorithm development and rise of modern storage equipment. Valuable knowledge can be obtained by these raw data. In early 1990s, knowledge discovery from data (KDD) term was used with the aim of knowledge extraction from database (Piatetsky-Shapiro & Frawley, 1991): "*Knowledge discovery is the nontrivial extraction of implicit, previously unknown, and potentially useful information from data*". Desirable features for extracted knowledge are reasonable time complexity, comprehensibility, accuracy and useful result. This kind of extracted knowledge can be used as a new knowledge. Data mining was originally considered as synonym of KDD (Fayyad, Piatetsky-Shapiro, & Smyth, 1996).

Recent researches have shown that application of data mining in several fields is growing such as CRM (Ngai, Xiu, & Chau, 2009), education (Romero & Ventura, 2010; Romero, Ventura, & García, 2008), clinical medicine (Bellazzi & Zupan, 2008), financial fraud detection (Kirkos, Spathis, & Manolopoulos, 2007; Ngai, Hu, Wong, Chen, & Sun, 2011), intrusion detection (Pietraszek & Tanner, 2005) and genetic data analyzing (Jiang, Tang, & Zhang, 2004). We expect its application in other fields to increase in the same manner. In a one hand, medicine, plays a great rule on human life, and on the other hand, need of automation for knowledge extraction and impossibility of manual processing, application of data mining in medicine has become a great issue. As we will see in Section 3, research on medical data mining is growing fast. Recently, application of data mining in medicine and healthcare is most widely used by data mining developers and academic researchers compared to the other fields. The rapid growth of medical data mining in the recent years represents the kick-off medical data mining.

The goals of collecting diverse medical data from various resources are to assist physicians, improve public health and support patients. All activities in medicine can be divided into six tasks: screening, diagnosis, treatment, prognosis, monitoring and management. These tasks start from a patient with a hidden disease without any symptoms, and lead to better management of the resources to improve social health care services.

* Corresponding author. Tel.: +98 9131633805.
E-mail addresses: n.esfandiary@qiau.ac.ir (N. Esfandiari), m.babavalian@qiau.ac.ir (M.R. Babavalian), eftekhari@qiau.ac.ir (Amir-Masoud Eftekhari Moghadam), vahidkashanitabar@gmail.com (V.K. Tabar).

http://dx.doi.org/10.1016/j.eswa.2014.01.011
0957-4174/

Some of the medical data mining trend anticipated here. Application of data mining in some disease is more common than others. High epidemic and mortality rate, expensive test, time consuming and requirement of special experience are expressed as the main reasons. Even though the heart disease has most mortality, number of works published on cancer is significant. This is due to the diversity, increased patients and global concerns over the time. Screening, early and accurate diagnosis of risky disease such as cancer, heart disease and diabetes is more common. This issue is progressed by developing efficient algorithms and novel technologies such as microarray data. A decision system can be used to treat in several conditions such as emergency situation, shortage of physicians and decrease human errors. The successful application and proven reliability of data mining in other tasks such as diagnosis is effective in advancing this issue. In several fields such as ICU and post surgery, we need continuously care. Then monitoring is known as another tendency. At last, the management issues such as bed management and scheduling can contribute to the improvement of medical services. By a data mining view, decision tree, artificial neural network and support vector machine (SVM) algorithms are the most popular. That is because of the simplicity and comprehensibility of decision tree, popularity and ability of artificial neural network in general model extraction and efficiency of SVM. Several challenges such as huge data; integration of text, structural data, signal and image; web and image mining; and integration with the hospital workflow wait to be tackled.

The purpose of this paper is to review medical data mining applications. We will examine similar papers in the context of review and survey to identify and understand previously presented approaches. We try to present a review study by completing previous works and integrating medical and data mining issues at the same time. For this reason, 291 papers have studied that published between 1999 and 2013 in 90 different medical, data mining and biomedical engineering journals. These studies have been used to provide medical data mining definitions, challenges, goals and to make a review. Our review is based on the six medical tasks: screening, diagnosis, treatment, prognosis, monitoring and management. In each task, related papers according to the data mining algorithms are presented. Data mining algorithms are classified as classification, regression, clustering, association rule mining and hybrid approach.

The relevance of this paper is strengthened by the fact that medical data mining is also an object of research in different communities. These communities include data mining, medicine, pattern recognition, machine learning, statistics and management. It is even more important to make present comprehensive review that covers different approaches, even those with little work carried out on. Proposed review examines each previous paper by medicine and data mining point of view.

It is quite hard to distinguish between data mining and aforementioned fields. That is why we have chosen papers aiming knowledge extraction from medical data. The number of works published on medical data mining is substantial and it is impossible to discuss all of them on a single work. Each of electronic records, medical stored events, text data, results of tests, medical images and signals and web data, can be known as the medical data. All these data types should be taken into account when implementing an appropriate algorithm for extracting knowledge. In this work, only structural data has been considered. Therefore, other types of data such as text are not considered. We hope this work will make it possible to develop a deep understanding of various approaches.

The rest of this paper is organized as follows: The next section will consider motivation for this study. Data mining concepts and medical data mining definitions and goals, main challenges, research trend and adaptive standard framework are expressed in Section 3. Sections 4, described 291 papers based on six medical tasks. In Section 5, discussion of the reviewed papers is provided to shape the current status and highlighted important issues. We will conclude and recommend for future work in final section.

2. Motivation of this study

In this section, the method and materials of this study are described. The method is referring to the procedure of collecting and gathering related papers. In the material, two subjects are tailored: glance of the prior surveys in the field of knowledge extraction from medical data, and determine the scope of this study.

As a result, 291 papers associated with application of data mining in medicine were selected that have been published between 1999 and 2013. These papers will be used in Section 3 to provide basic concepts such as definitions and goals and in Section 4 to present review. Also, the gathered collection is examined to obtain highlight issues, some facts and figures and investigate finding.

2.1. Workflow applied for extraction of medical data mining papers

The workflow used to provide this survey contains four processes: *broad search*, *refine search*, *extract basic concepts* and *analyze*. An overview of the mentioned workflow is drawn in Fig. 1.

In *broad search* initial papers that associated with the application of data mining were extracted in medicine. We selected papers aiming knowledge extraction from medicine with "data mining" term appeared in their title, abstract and keywords. To select papers two ways were taken: search in four scientific databases ELS., IEEE*xplore*, Springer and Pubmed and, search in well-known journals in data mining, bioinformatics and medicine.

Raw papers that extracted from broad search process pass to the *refine search*. In this process, abstract and introduction of each paper was studied. Some of the papers tend completely to the medicine domain and the weight of data mining was little. Thus these papers ignored. Finally, 291 papers published in 81 journals between 1999 and 2013 were remained.

Obtained papers from previous process were used to provide medical data mining definition and goals and, made a review. *Extract definition and goals* process attempt to pluralize and unify the different definition and goals given in various papers. In *analysis based medical tasks* process each paper placed in one of the six medical tasks. Moreover, in each task analyzing based on data mining view is carried out. Finally, *current issue* is discussed and *future trend* is drawn.

2.2. Previous reviews in the context of knowledge extraction in medicine

In this section previous reviews associated with extraction knowledge in medicine are outlined. We can consider each of these review papers from two points of view: algorithm and medicine. In algorithm view, some of the papers focused on a certain algorithm such as neural network and others considered a discipline of AI such as machine learning. From medicine point of view, certain disease (such as urology) or group of disease (such as cardiovascular diseases) or part of medicine tasks (such as diagnosis) has been studied. The goal of this sub-section is to better understanding of medical knowledge extraction background. For this reason, twelve review papers are studied.

In Itchhaporia, Snow, Almassy, and Oetgen (1996) application of neural network cardiovascular disease is reviewed. For this purpose cardiovascular is divided into coronary artery disease, acute myocardial infarction, electrocardiography, arrhythmia identification, arrhythmia localization, cardiac image analysis and cardiac drug dosing. In the conclusion, three main disadvantages of neural net-

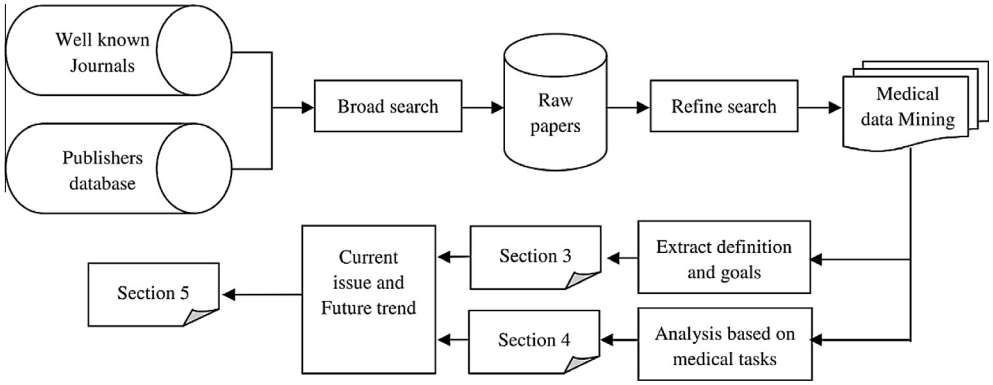

Fig. 1. Workflow to collect related papers with the aim of provide this survey.

works are studied. Authors believe that neural networks cannot replace expertise and traditional statistical approaches, but it can enhance performance. In a similar way, application of neural networks in oncology, critical care and cardiology for diagnosis, outcome prediction, radiology and monitoring are studied (Lisboa, 2002). He concludes that most of the papers have been appeared in diagnosis and prognosis. He believes that the role of computer in medicine will be extended in future. Also in Anagnostou, Remzi, Lykourinas, and Djavan (2003), application of neural network in diagnosis, screening, staging and progression of prostate cancer in urological oncology domain is reviewed. Authors try to investigate two issues: benefit of neural network in clinical decision making such as urology domain, and a comparison between neural network and statistical techniques. Finally, they conclude that Superiority of this method compared with traditional statistical approaches cannot be confirmed in all conditions.

Applications of six types of evolutionary algorithms are discussed in Peña-Reyes and Sipper (2000). Genetic algorithm, genetic programming, evolutionary strategies, evolutionary programming, classifiers and hybrid systems have been mentioned in this paper. Medicine has been divided into diagnosis, prognosis, imaging, signal processing, planning and scheduling. This review shows that a genetic algorithm is widely used in this field.

Abbod, von Keyserlingk, Linkens, and Mahfouf (2001) made a review study about application of fuzzy technology in medical domain until 1998. For this purpose, medical domain was divided into nine main categories that each category divided into subcategory: conservative medicine, invasive medicine, regionally defined medical disciplines, neuro-medicine, image and signal processing, laboratory, basic science, nursing, healthcare and oriental medicine. As a conclusion three main tips can be pointed: (1) more attention fields: internal medicine, anesthesia, radiology, electrophysiology, pharmacokinetics, and neuromedicine; (2) less attention fields: surgical disciplines, dental medicine, general practice and nursing; (3) growing fields: medical reasoning and decision support sciences. In a further research, previous works was considered by a fuzzy control and monitoring point of view (Mahfouf, Abbod, & Linkens, 2001). In this review seven fuzzy techniques (basic controller, rule based open loop, rule based closed loop, self learning, model based and adaptive, hybrid systems, hierarchical systems) and three fuzzy algorithms (fuzzy clustering, fuzzy classification, fuzzy modeling) were introduced for solving the control problems in a medical domain.

Applications of smart and adaptive engineering systems in diagnosis, therapy and imaging domain have been reviewed (Abbod, Linkens, Mahfouf, & Dounias, 2002). In this study five major categories were considered: emergency and intensive care unit (ICU), general medicine, surgical medicine, pathology and medical imaging. As a result, ANN known as a most focused algorithm.

Application of knowledge based systems, intelligent computing systems and combination of them in diagnosis, treatment and planning are reviewed from 1970 to 2008 (Pandey & Mishra, 2009). In this paper rule-based reasoning, case-based reasoning and model-based reasoning known as knowledge based systems whereas genetic algorithm, artificial neural network and fuzzy logic known as intelligent computing methods. The conclusion of this review has two points: (1) most of the methods are used in diagnosis and less of them in planning, (2) three most of the used algorithms are case-based reasoning + rule based reasoning.

Yardimci (2009), provides a review of application of soft computing in medicine. For this purpose, fuzzy logic, artificial neural network, genetic algorithm and combinations of them considered as soft computing from 2000 to 2008. He believes that fuzzy + neural network is significantly used and using neural network + genetic algorithm is rising. This study covers basic science, diagnostic science, clinical and surgical disciplines.

Recently, three review papers were published in the machine learning and data mining field. Make a prediction model by classification algorithms in medicine, especially in genomics and forecasting outcome were studied in Bellazzi and Zupan (2008). In this papers only focused on classification algorithms such as decision tree, decision rule, logistic regression, neural network, naïve bayes, bayesian network, support vector machine and *k*-nearest neighbor. This study tries to provide a comprehensive framework to organize application of data mining in medicine. Application of data mining algorithms in healthcare and biomedicine is discovered in Yoo et al. (2012). In this paper three data mining tasks are selected and then application of each task in medicine is reviewed. For this purpose, a brief discussion about each task and their advantage/disadvantage is presented. The main medical aspects that mentioned are: prediction health costs, prognosis and diagnosis, extract hidden knowledge from biomedicine data, discover relationship among diseases and among drugs. This paper has been finished with three data mining problems in medicine: set and calibrate parameters of algorithms, accuracy of data mining is not reliable yet and, lack of data mining package for medical domain. In Wagholikar, Sundararajan, and Deshpande (2012) reviews application of more than eight modeling techniques in diagnosis. In this study more than ten diseases were selected. Finally, authors conclude that application of these techniques in gastroenterology, oncology and cardiovascular are more than the others diseases.

2.3. Scope of this study

The study of review papers in previous section shows that application of knowledge extraction from medicine is started from last decade and therefore it is located in a teenage period. Growth of the published papers in wide verity of journals, spread data min-

ing to apply in a large part of the medicine fields and increase tendency of journals to publish papers in this area in recent years reveal the importance of updating the previous researches. On the other hand, previous reviews focused on a certain part of data mining and medicine.

In this study only the sample of representative papers are selected that published in journals and we cannot claim that all the papers are examined. The chosen papers have been published during the 1999–2013. This study is constraint to structural data. Therefore text and similar data are not covered.

3. Data mining and medical data mining: basic concepts, definition and goals

In this section, basic concepts of data mining will be described. First, data mining process divided into three phase: data pre-processing, data modeling, and data post-processing. In each phase a brief explanation of well-known and important methods are given. Then, tried to provide a definition for medical data mining and elicit main goals based on literature. Understanding of data types and challenges are discussed in later. Finally, an adaptive standard framework to formulize usage of data mining in medicine will be introduced. This section makes it possible to achieve a comprehensive definition of medical data mining, goals and challenges, various data and available data sets. In the following, previous trends will be illustrated.

3.1. Data mining

Data mining can be considered as KDD with three steps: data pre-processing, data modeling and data post-processing (Fayyad et al., 1996). Nowadays, data mining is stated as a multidisciplinary area. The goal of data pre-processing is to prepare raw data for mining. In data modeling step, relationships between various data are discovered to extract a pattern. Extracted pattern should be evaluated in data post-processing step to be verified and stated as knowledge. We can also use background knowledge to clarify the problem and verify extracted knowledge (Tan, Steinbach, & Kumar, 2006).

3.1.1. Data pre-processing

Pre-processing step refers to data preparation and obtain a general overview of data. In practice between 60% and 90% of project time is normally spent for data understanding and preparation (Chapman et al., 2000). This shows the importance of data pre-processing step. In Fig. 2 we have presented main approaches of data pre-processing. In the following, examples in medicine application are given for each of them.

3.1.1.1. Data description and summarization. Statistical methods such as mode, median, mean, etc. (Frize, Ennett, Stevenson, & Trigg, 2001; Ghazavi & Liao, 2008; Li, Fang, Lai, & Hu, 2009; Li et al., 2004; Shah & Kusiak, 2007) and a hybrid of genetic algorithm + correlation (Razavi et al., 2005; Shah, Kusiak, & O'Donnell, 2006) can be considered in this approach.

3.1.1.2. Data cleaning. Noise is the random part of error and should be eliminated (Apiletti, Baralis, Bruno, & Cerquitelli, 2009; Huang, Chen, & Lee, 2007; Ramon et al., 2007). Outlier data have a different behavior from other data and should be detected (Combes, Meskens, Rivat, & Vandamme, 2008; Su, Yang, Hsu, & Chiu, 2006; Thongkam, Xu, Zhang, & Huang, 2009). Missing value means that some data cells remain empty. Elimination, estimation and ignorance can solve this problem (Chen, Hou, & Chuang, 2010; Yang, Lin, Chen, & Shi, 2009).

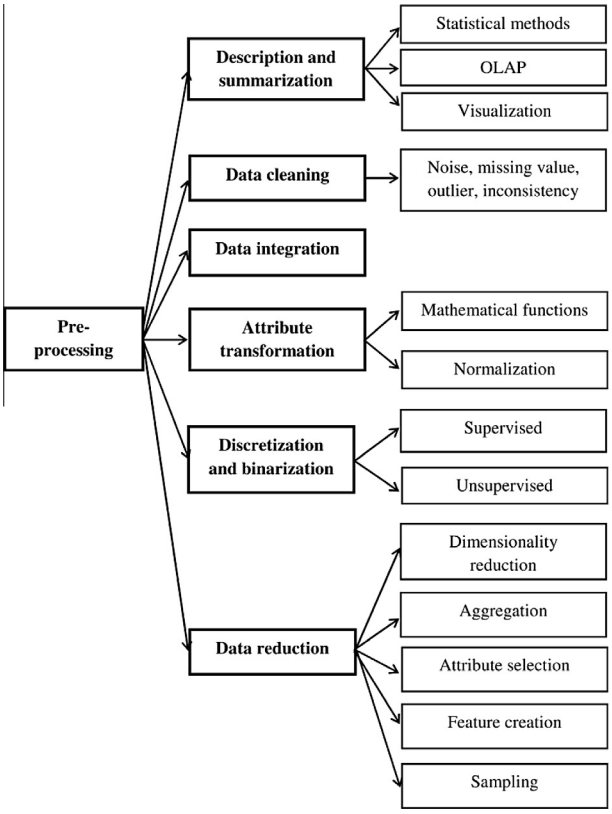

Fig. 2. Main data pre-processing approaches.

3.1.1.3. Data integration. Integrating two or more data sets to create single data set (Delen, Oztekin, & Kong, 2010; Hautemanière, Florentin, Hartemann, & Hunter, 2011; Santos, Malheiros, Cavalheiro, & de Oliveira, 2013).

3.1.1.4. Attribute transformation. Converting all values of a special variable to the desired scale or value (Chang, Yeh, & Huang, 2010; Papachristoudis, Diplaris, & Mitkas, 2010; Subasi, 2012; Zhou et al., 2010).

3.1.1.5. Discretization and binarization. Preparing continuous data for classification and association algorithms (Bertolazzi, Felici, Festa, & Lancia, 2008; Cheng, 2012; Wiggins, Saad, Litt, & Vachtsevanos, 2008).

3.1.1.6. Data reduction. Reducing records or columns to achieve a simpler and more interpretable model, decrease time and memory, eliminate irrelevant features and avoid curse of dimensionality. Several methods were used for data reduction in literature such as feature subset selection (Bontempi, 2007; Filipovic, Ivanovic, Krstajic, & Kojic, 2011; Hsu & Ho, 2004), clustering (Au, Chan, Wong, & Wang, 2005; Guenther, Mueller, Preuss, Kruse, & Sabel, 2009; Pechenizkiy, Tsymbal, & Puuronen, 2006; Polat, 2012), sampling (Taft et al., 2009) and aggregation (Imamura et al., 2007; Lamma et al., 2006). Genetic algorithm (Chang & Chen, 2009), regression (Samanta et al., 2009) and artificial neural network sensitivity analysis (Yan, Zheng, Jiang, Peng, & Xiao, 2008) can also be used for feature selection. In medical data mining, expert knowledge can help to select appropriate features (Alonso, Caraça-Valente, Martínez, & Montes, 2003; Nahar, Imam, Tickle, & Chen, 2013b). In addition, fuzzy gain ratio (Dai & Xu, 2013) inspired by gain ratio, and hybrid forward selection (Shilaskar & Ghatol, 2013) were used to improve feature selection process. Martis, Ach-

arya, Mandana, Ray, and Chakraborty (2012) investigated application of principal component analysis for diagnosis of cardiac health. In this paper, five types of ECG beats were classified based on arrhythmia database. Performance evaluation showed that proposed approach is ready to be used in clinical programs.

3.1.2. Data modeling

All the tasks in data modeling step can be divided into *predictive* and *descriptive* categories. Predictive algorithms are divided into *classification* and *regression* based on type of target variable (discrete or continuous). Descriptive algorithms are also classified as *clustering* and *association rule* mining (Tan et al., 2006). In Section 4, six medical tasks will are considered and then in each of them, five data mining approach (classification, regression, clustering, association rule and hybrid) will be studied.

Classification refers to supervised methods that determine target class value of unseen data. The process of classification is shown in Fig. 3. In classification, data is divided into training and test sets used for learning and validation, respectively. We have described most popular algorithms in medical data mining in Table 1. These algorithms are the most used in literatures and also known as popular based on kdnuggets pool (http://www.kdnuggets.com/polls/2011/algorithms-analytics-data-mining.html). Performance evaluation of classifiers can be measured by hold-out, random sub-sampling, cross-validation and bootstrap. Among which, cross validation is the most common.

Regression analysis is a statistical technique that estimates and predicts relations between variables. Instances of regression algorithms are simple linear, multiple linear, fuzzy and logistic. In data mining, regression is used to predict unseen data based on continuous training data. In this approach, behavior of dependent variable y is explored by independent variables x.

Data clustering consists of grouping and collecting a set of objects into similar classes. In data clustering process, objects in the same cluster are similar to each other while objects in different clusters are dissimilar. Data clustering can be seen as *grouping* or *compression* problem (Han, Kamber, & Pei, 2011). Most popular data clustering methods are described in Table 2.

Many fields such as business enterprise collect large data from daily operations. Sequential data such as purchased items in a market are common and popular. Association rule mining is a method for exploring sequential data to discover relationships between large transactional data. The result of this analysis is in the form of association rules or frequent items. In Table 3, most popular association algorithms are shown. Performance evaluation of discovered rules was done considering various criteria such as support and confidence. In Section 4, papers are reviewed based on six medical tasks that on each of them, five data modeling approach are considered.

3.1.3. Data post-processing

In this step, visualization and evaluation of extracted knowledge is taken into account. Visualization is stated as the essential matter in data mining. It is considered as an advantage for an algorithm. There are various predictive and descriptive algorithms for knowledge extraction. This means that, different performance

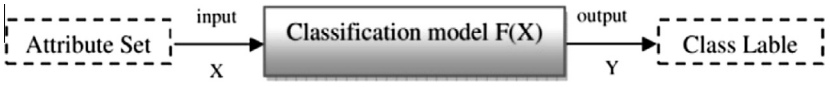

Fig. 3. Process of classification.

Table 1
Most popular classification algorithms in medical data mining based on the literature and kdnuggets pool.

Algorithm	Advantage	Disadvantage	Characteristic
DT	Non-parametric, interpretable, resistant to noise and replication	Separation line parallel to axis x, y; sensitive to the inconsistent data	Eager approach; Greedy; recursive; partitioning; stable
ANN	Diagonal separation line, popular in the other fields, ability to complex relation, resistant to replication	Black box, parametric, sensitive to the noise and missing value; increase time by increase hidden layers	Eager approach; multi-layer network with at least one hidden layer
Rule-based	Interpretable; resistant to noise and imbalance data	Separation line parallel to axis x, y	Eager approach; produce *if…then* rules; partitioning
SVM	Diagonal separation line; appropriate for high dimensional data and little training data	Black box; parametric	Eager approach; mathematical based; unstable; optimization; global minimum
NB	Resistant to noise, missing value and irrelevant features	Accuracy degraded by correlated attribute; required to determine initial probability	Eager approach; statistical based; non-deterministic
KNN	Simple; flexible; arbitrary decision boundaries	Sensitive to noise and replication; parametric	Lazy approach; instance based; required similarity measurement; prediction based on local data

Table 2
Most popular data clustering methods based on the literature.

Algorithm	Advantage	Disadvantage	Characteristic
K-means	Simple, fast, popular	Parametric, susceptible initial value, inappropriate for data different in size and density, different results in each run, sensitive to noise	Optimization problem, prototype-based, partitioning problem, center-based
Hierarchical	Non-parametric, less susceptible of initial value	Time and space complexity, sensitive to noise	Graph-based, prototype-based, bottom-up
DBSCAN	Resistant to noise, handle arbitrary density and size.	Time and space complexity	Density-based, non-complete, partitioning problem
Fuzzy *c*-means	Same as *K*-means	Same as *K*-means	Same as *K*-means, determining membership of each object to the clusters.

Table 3
Most popular association rules methods based on literature.

Algorithm	Advantage	Disadvantage	Characteristic
Apriori	Popular, simple	Time and I/O complexity, reviewing entire database at each stage, searching in all variables	Using prior knowledge, iterative approach
DIC	Decrease I/O complexity	Sensitive to data homogeneity	Dynamic, Retrieving lost patterns by moving forward, investigating the specified distance of transactions
DHP	Reducing the number of candidate patterns	Relation between runtime and database size, Collision problem in the hash table	Using hash table
Eclat	Decreasing I/O complexity, exploring large length patterns, and discovering all sequential objects	Space complexity, inappropriate for large data	Bottom-up approach, using lattice-theoretic
D-CLUB	Removing the empty bits, reduce time and space complexity, self-adaptive	–	Appropriate for parallel process and distributed database, dynamic, differential optimization

evaluation methods are required. In general, performance evaluation methods can be divided into *single scalar* and *graphical* (Prati, Batista, & Monard, 2011). Accuracy, sensitivity and specificity are classified in the first group. The main characteristic of this group is simplicity in implementation but less efficiency in covering various aspects of evaluation. On the other group, ROC Curve, Cost-Line and Lift are considered. Methods in this group have a complex implementation but make good sense. We have shown popular performance evaluation methods in Table 4 based on literature. In below some of works in post-processing in medicine are seen.

Cao, Maloney, and Brusic (2008) proposed a system for rapid extraction of useful information about cancer by visualization and summarization. This system uses clinical trial registries and analyzed data associated with cancer vaccine trials. Results of this approach are extracted as key data about cancer vaccine trials and can be used for developing vaccines in future.

Voznuka et al. (2004) provided a reporting system for thoracic surgery. This reporting system dynamically generates two types of reports based on users including physicians, administrative staff, and patients. Results are visualized as tabular and charts that report about morbidity and mortality of diseases in patients. Lavrač et al. (2007) proposed a novel decision support system based on visualization to manage health care resources. Visualization can be used to facilitate knowledge management and produce interpretable model (Yanqing et al., 2011). Some authors have used simulations of physician learning with the help of their experience (Becker, 2001). In several papers, proposed model was evaluated based on expert opinions (Alonso, Caraça-Valente, González, & Montes, 2002; Shusaku, 2000).

Recently, application of data mining is substantial. Fig. 4, illustrated top ten usage of data mining in different branches of science up to 2012, by polling (http://www.kdnuggets.com/polls/2012/where-applied-analytics-data-mining.html).

Table 4
Most popular performance evaluation methods.

Algorithm	Characteristic
ROC graph	Comparing performance of two or more predictive models, independent from class distribution
Cost-lines	Evaluating error rate based on the different costs
ROC curve	Able to rank positive and negative samples, independent from class distribution
Precision-recall curve	Evaluation and ranking each sample based on positive class
Lift graph	Identifying relations between true positive rate and profanity of positive classification
Reliability diagram	Investigating probability of true calibration of model
Attributed diagram	Identifying regions of model that degrade performance compared with reference models with constant performance
Discrimination diagram	Showing discrimination between each class prediction
Accuracy	Rate of correct classification
Sensitivity	Rate of true positive classification
Specificity	Rate of true negative classification
AUC	Area under the ROC curve

3.2. Medical data mining

Today, a massive amount of medical data in various formats and data types is stored in databases. That is a result of development of technology and storage equipment. For this reason, it is impossible to process data by physicians using traditional techniques. Based on Fig. 4, it is clear that using data mining in medicine is inevitable. Medical data mining organization presented in this paper can be seen in Fig. 5. Definition and goals of medical data mining are given in this Section. We have provided data mining goals as considering time and accuracy importance, decision support system necessity and non-trivial knowledge extraction. Important factors such as data and its challenges affect extracted knowledge. The main challenge regarding data is that wrong input will lead to wrong output.

3.2.1. Definition and goals

Medical data mining definitions in the literature are different depending on the author's view. Some of papers have focused on data mining algorithms, while others concentrated on medical domain and diseases. According to literature, following definition is

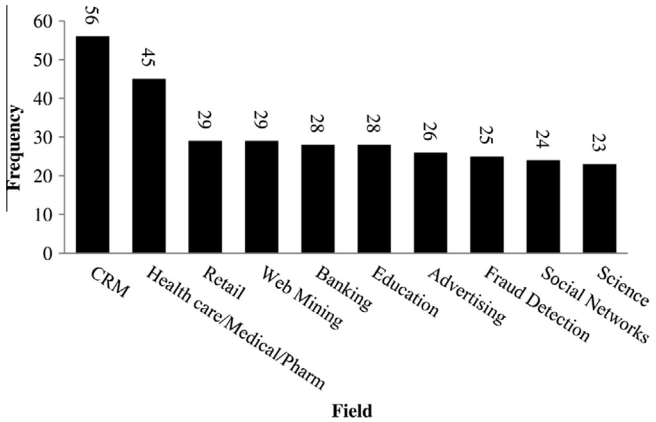

Fig. 4. Frequency of data mining applications in different fields up to 2012 based on kdnuggets pool.

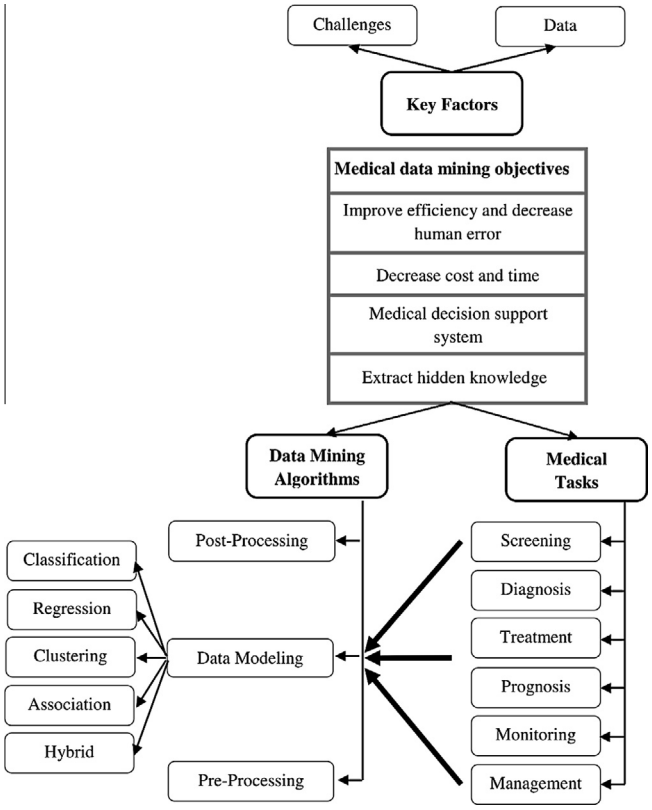

Fig. 5. Medical data mining organization of our work.

introduced: *Extraction of implicit, potentially useful and novel information from medical data to improve accuracy, decrease time and cost, construct decision support system with the aim of health promotion.*

This definition contains three parts:

(1) Data mining: same as what it was defined in Section 1 that extraction of implicit and useful knowledge are taken into account.
(2) Medical nature: using medical data and applying extracted model to medical domain.
(3) Goals: with the study of various papers, four goals can be explored as follow:
   (3a) *Improving efficiency and decreasing human error*: Suitable for particular diseases such as heart disease and cancer, in which accuracy is an important factor.
      - Early diagnosis of cerebrovascular disease is becoming more important due to the high mortality and morbidity rate and treatment cost (Yeh, Cheng, & Chen, 2011).
      - Microarray data can be used in diagnosis to increase accuracy and decrease time. Selecting appropriate genes is the main challenge that affects efficiency of extracted knowledge (Lee & Leu, 2011), e.g. in tumor detection (Dai & Xu, 2013).
      - Medical data mining has made it possible to construct a prediction model for type 2 diabetes that is more accurate than previous models (Patil, Joshi, & Toshniwal, 2010).
      - Breast cancer survivability suffers data problems such as outliers and skews. This leads to the need of data pre-processing to improve performance (Thongkam et al., 2009).

   (3b) *Decrease in time and cost*: Suitable for tests with time complexity and high attributes.
      - Given that diagnosis of tuberculosis by laboratory tests is time consuming, we can do temporary treatment by data mining systems until test results are provided (Uçar & Karahoca, 2011).
      - One of the main causes of blindness is glaucoma. Several expensive and complex methods, such as optical coherence tomography, Scanning Laser Polarimetry (SLP) and Heidelberg Retina Tomography (HRT) scanning methods are used for glaucoma diagnosis. An automatic diagnosis system can be made for this purpose by the use of data mining (Mookiah, Rajendra Acharya, Lim, Petznick, & Suri, 2012).
      - Medical data mining is used to build a web-based support system for early diagnosis and prevention of typhoid fever (Samuel, Omisore, & Ojokoh, 2013).

Alzheimer is a progressive disease that affects mental functions. Functional brain imaging method known as Single-Photon Emission Computed Tomography (SPECT) is used for diagnosis by manual processing and visual observation. Data mining techniques are used to deal with the cost problem (Illán et al., 2011).

   (3c) *Medical decision support system*: Used to automation of several process. This is can be used for inexperienced physician and critical situations such as flood, earthquake.
      - Building a smart system by six prediction algorithms and evaluate them to make head injury mortality estimation (Sut & Simsek, 2011).
      - Classification of vehicle accident patients based on severity of injury (Scheetz, Zhang, & Kolassa, 2009).
      - Determining hypovilemic state in trauma patients during their transportation to hospital using vital signs (Chen, McKenna, Reisner, Gribok, & Reifman, 2008).
      - Diagnosis and prognosis of chronic diseases by using a combination of data mining and case based reasoning (Huang et al., 2007).
   (3d) *Knowledge extraction*: Main features of this goal are extracting relations between variables, identifying risk factors and exploring novel knowledge.
      - Extracting new hypothesis about adverse drug effects by logging patient's comments posted on the message boards and comparing them with existing hypothesis (Benton et al., 2011).
      - Producing efficient medical rules for experts with association rule mining and pruning them (Mansingh, Osei-Bryson, & Reichgelt, 2011).
      - Exploring patterns from Alzheimer's microarray data by sequential pattern mining and using visualization techniques to gain a better knowledge understanding (Sallaberry, Pecheur, Bringay, Roche, & Teisseire, 2011).

Percentage of each medical data mining goals as detailed in Fig. 6.

3.2.2. Medical data

Medical data is one of the most difficult data type for mining. This type of data is generated by several sources such as medical examinations, imaging and tests. We need to select an appropriate data mining algorithm regarding the type of data. Using various types of data for diagnosis and treatment requires standardization

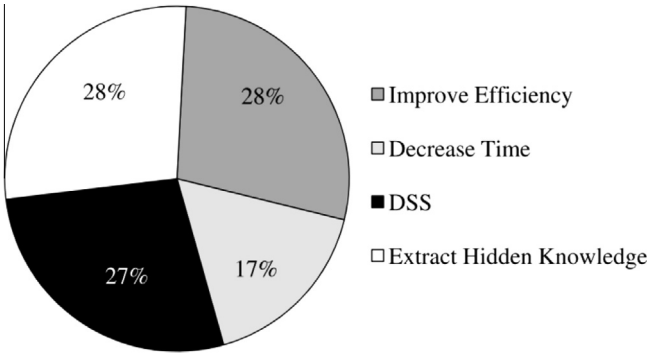

Fig. 6. Percentage of medical data mining goals.

and integration. Based on Lavrač (1999) medical data can be divided into clinical and temporal. These data are collected with the aim of helping researchers in the fields of screening, diagnosis, treatment, prognosis, monitoring and patient management.

Different types of data are presented in the following:

(1) Clinical data with the nature of text and qualitative format.
(2) Trial data with the nature of numeric and quantitative format.
(3) Image data such as MRI and Radiology.
(4) Ultrasound data such as Echo and Sonography.
(5) Sequential or time series data with unique time stamp.
(6) Signal data such as EEG and ECG that have the same characteristics as time series data.
(7) Genetic, microarray and protein data that have small records and large variables.

In this study only structural data are considered and others types such as text are not considered. In Fig. 7, frequency of using each data types in literature are shown. Based on this figure, numeric and microarray data are most widely used.

The main characteristic of microarray data is small number of records (Human) and large number of variables (Genes). This is led to curse of dimensionality phenomenon. Feature selection can be used to provide dimensionality reduction and irrelevant gene elimination (Li, Yang, Sablok, Fan, & Zhou, 2013). SVM is the most frequently used method when facing this kind of data and has proven high efficiency. Size of microarray data used in literature in terms of record and column is shown in Fig. 8. According to Fig. 8, average number of records and columns are 94 and 8420, respectively. Microarray data are almost used in cancer screening and discovering important genes.

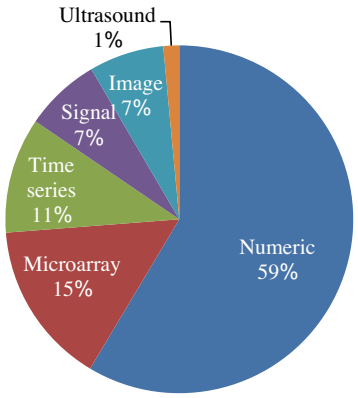

Fig. 7. Percentage of medical data types that used by literature.

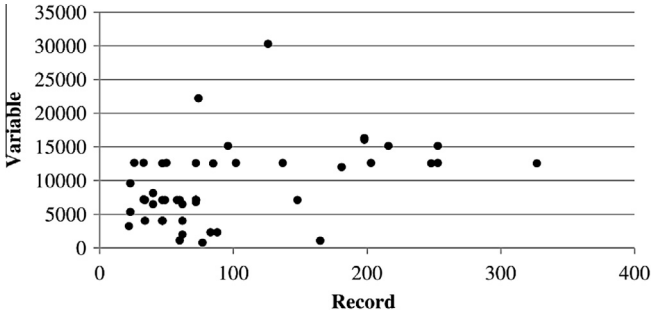

Fig. 8. Size of microarray data.

Sequential and time series data are considered of transactional type, in which correlation between data is the most important feature. It is possible to process these data by association rule mining (Laxminarayan, Ruiz, & Moonis, 2006) and clustering (Altiparmak, Erdal, & Trost, 2006). Signal, image, and ultrasound require signal processing (Bourien, Bellanger, Bartolomei, Chauvel, & Wendling, 2004), image processing and image mining (Cilla, Martinez, Pena, Marti, & nez, 2012; Sacha, Cios, & Goodenday, 2000), and voice processing in data pre-processing phase. Hence, papers associated with mentioned data types, can also be classified in other domains except data mining. As Fig. 7 shows, numeric data is the most frequently used data type due to the available data sets and adaptation with data mining. In Fig. 9, sizes of numeric datasets are presented. For this purpose, number of records in data set is considered as Low, Medium or Large if the amount of records is less than 1000, between 1000 and 100,000 or more than 1,000,000, respectively. In the same way, number of columns was named Low, Medium or Large by the value of less than 20, between 20 and 50 or more than 50, respectively.

In Fig. 9, most works (47%) have been carried out on a small number of records and variables. Little effort has been devoted to large scale data in terms of records or variables because of lack of available large data sets and scalability problem. Popular data sets that available on the web, based on literatures can be seen in Table 5.

Data suffer some problems such as noise and missing value as common challenges in all fields. It is exposed to some other problems exclusively important in medicine. Medical data challenges are classified in Fig. 10.

Main challenge of provision of a unique medical knowledge extraction framework is lack of appropriate collection and transmission standards. Despite the fact that there are standards such as ICD-10 for disease information integration (WHO). Some of general challenges such as skewness and high dimensionality of microarray data are more significant in medicine. Medical data are often poorly characterized mathematically compared to other domains. Given that medical data contain personal information and are prone to abuse, ethical and legal constraints should be taken into consideration in their case. Berman (2002) tries to address confidentially issue.

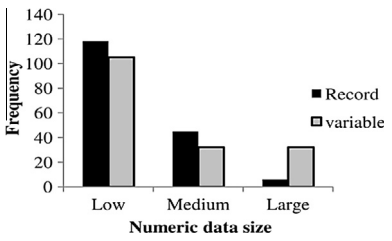

Fig. 9. Size of numeric data sets.

Table 5
Available data sets that used in literature.

Organization	Disease	Data Type	Site
University of California, Irvine, School of Information and Computer Sciences: (UCI)	Parkinson, heart, arrhythmia, liver, breast cancer, thyroid, diabetes, lung cancer and survival of surgery for breast cancer	Numerical data	http://archive.ics.uci.edu/ml/datasets/
BROAD Institute	cancers	Microarray data	http://www.broadinstitute.org/
ADNI Organization: Alzheimer's Disease Neuroimaging Initiative.	Alzheimer's disease	Numeric, genetic, image data	http://adni.loni.ucla.edu/data-samples/mri/
Agency for Healthcare Research and Quality	Healthcare	Numerical data	http://www.ahrq.gov/research/data/index.html
Health Informatics Lab, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences	Cardiovascular	Numerical data	http://www.healthinformaticslab.org/supp/index.php
National Human Genome Research Institute	Cancer	Microarray data	http://research.nhgri.nih.gov/microarray/Supplement/
The Health and Social Care Information Centre, England (NHS)	Cancer, child and adolescent mental health services, heart and diabetes diseases	Numerical & image data	http://www.ic.nhs.uk/collectingdata
Princeton University Gene Expression Project	Cancer	Microarray data	http://genomics-pubs.princeton.edu/oncology/
Nanyang Technological University, Singapore	Leukemia, breast cancer, central nervous system, colon tumor, ovarian cancer, prostate cancer, lung cancer, lymphoma diseases.	Numerical data	http://levis.tongji.edu.cn/gzli/data/mirror-kentridge.html
The UHN Microarray Centre (UHNMAC), North America	Cancer	Microarray data	http://www.microarrays.ca/resources/databases.html
Surveillance Epidemiology and End Result, U.S. National Cancer Institutes	Cancer	Numerical data	http://www.seer.cancer.gov
Healthcare Cost and Utilization Project (HCUP). Agency for Healthcare Research and Quality	Health care	Numerical data	http://www.hcup-us.ahrq.gov/databases.jsp
Pepe Lab, Fred Hutchinson Cancer Research Center.	Cancer	Microarray and numeric data	http://labs.fhcrc.org/pepe/dabs/datasets.html
BIO GPS	Cancer	Microarray and image data	http://biogps.org/dataset
The Stanford volume data archive	Head diseases	Image data	http://graphics.stanford.edu/data/voldata/
Department of Biostatistics, Vanderbilt University	Heart disease, meningitis, diabetes, muscular dystrophy, cancers, hypertension	Numerical data	http://biostat.mc.vanderbilt.edu/twiki/bin/view/Main/DataSets?CGISESSID=10713f6d891653ddcbb7ddbdd9cffb79
Seizure Prediction Project Freiburg, University of Freiburg	Brain disease	EEG data	http://epilepsy.uni-freiburg.de/freiburg-seizure-prediction-project
Supplemental data for classification, subtype discovery and prediction of outcome in pediatric lymphoblastic leukemia by gene expression profiling, st.Jude childern Research hospital	Lymphoblastic leukemia disease	Microarray data	http://www.stjuderesearch.org/site/data/ALL1/
Meyerson laboratory at the Dana-Farber Cancer Institute	Cancers	Microarray data	http://research4.dfci.harvard.edu/meyersonlab/New_MM_site_datasets.html
PhysioNet: Components of a New Research Resource for Complex Physiologic Signals	ICU	ECG and numeric data	http://physionet.org/physiobank/database/mimicdb/
DDSM: Digital Database for Screening Mammography, University of South Florida	Breast cancer	Image data	http://marathon.csee.usf.edu/Mammography/Database.html
World Health Organization, United Nations system	Health care	Numerical data	http://www.who.int/research/en/

In medicine, human is primarily considered as a patient and secondarily as a research resource, therefore collecting data in medicine is different from statistical domain (Cios & William Moore, 2002). Patel et al. (2009) has given an insight on model building challenges such as lack of standards and confidentially issue. A medical intelligent system is required to integrate various fields such as speech understanding, computer vision system and gesture tracking.

3.2.3. Medical nature challenges

Two underlying challenges can be taken into account in addition to mentioned challenges.

(1) *High risk*: medicine can be known as high risk due to the association with human life. Medical activities outcome is death or life and this is different from engineering view such as optimization. Mistake or failure can lead to punishment. A large portion of governments' expenditure is spent on health care. Among all sciences, medicine requires the longest education.
(2) *Interactions with a wide variety of users*: knowledge should be extracted by user type, so that each user has particular goals:
   - Patients: get rid of the disease and doing a better life.
   - Physician: early diagnosis, selecting appropriate treatment and reducing disease effects.

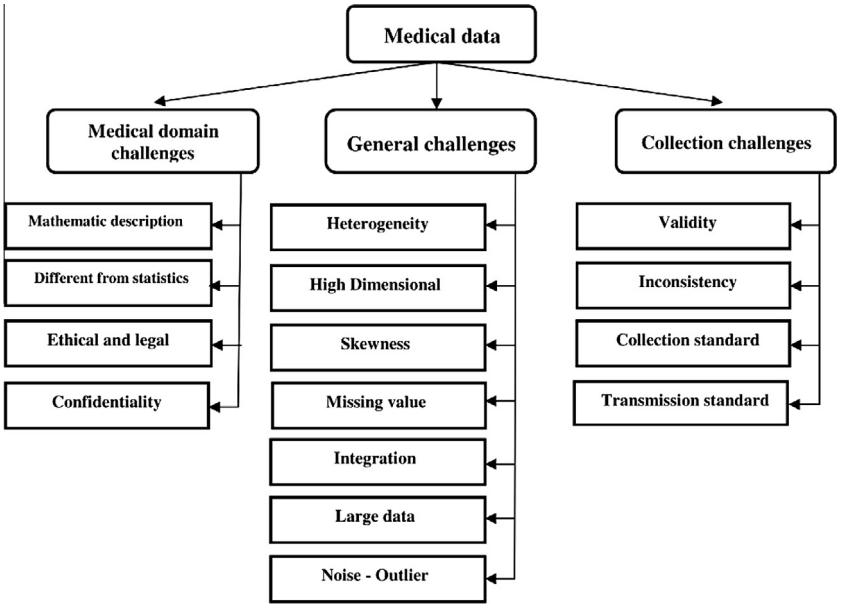

Fig. 10. Medical data challenges.

- Nurses: cooperating with physicians to improve patient care and monitoring treatment process.
- Health care system: improving quality of services and community health.

3.2.4. Medical data mining adaptive standard framework

Based on a standard framework, we can extract and evaluate knowledge from medical data in a uniform way. Cross Industry Standard Process for Data Mining (CRISP-DM) (Chapman et al., 2000) that known as a standard model was adapted for medical data mining in this section. Based on Fig. 11, this framework contains six steps:

(1) *Problem understanding*: initially, project objectives, requirements and constraints are determined from medical perspective and then converted to data mining problem. Project plan and validation strategy are produced as the last phase in this step.
(2) *Data understanding*: data are collected from extensive sources and an overview of them is achieved.
(3) *Data pre-processing*: description and summarization, data cleaning, integration, attribute transformation, discrimination, binarization and data reduction are performed in this step as detailed in Section 3.1.1.
(4) *Data processing*: algorithms that solve data mining problem in step 1, generate different models. In this step the best model is selected. Stepping-back to the previous step may be required if some algorithms need special data pre-processing. We will address data processing issue in Section 3.1.2.

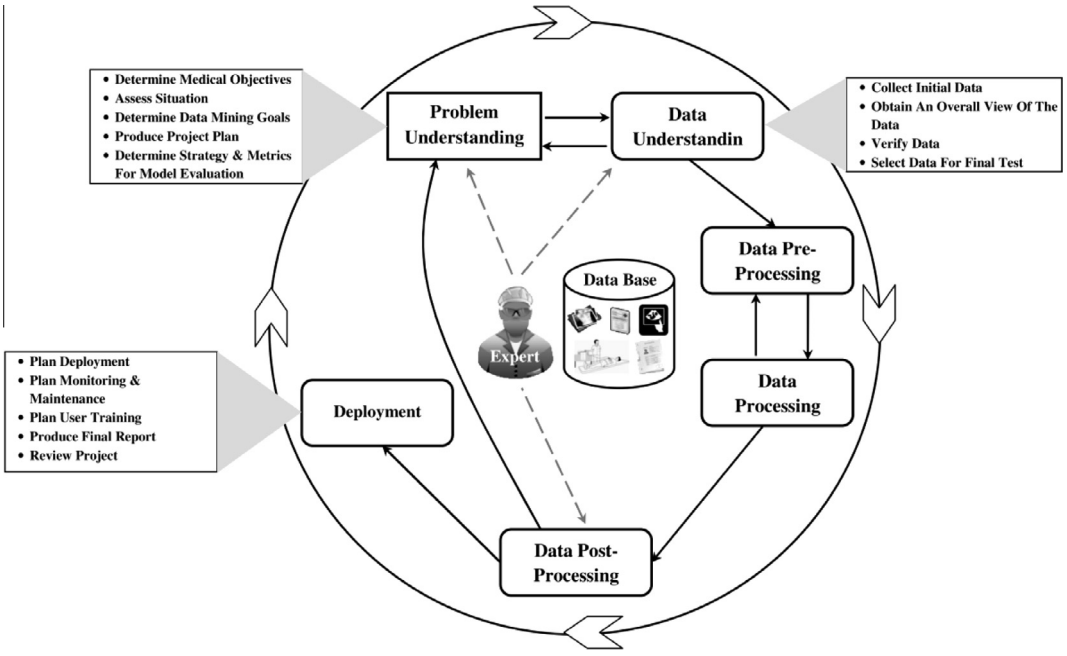

Fig. 11. Medical data mining adaptive standard framework.

(5) *Data post-processing*: validity of the model is assessed to meet medical objectives that may be needed to re-examine overall knowledge extraction process. Visualizations and interpretable models along with expert help contribute to model verification. Evaluation techniques will be described in Section 3.1.3.
(6) *Deployment*: prepared model should be useful for medical purposes. For example, if we have considered data mining system as embedded, it should be placed in the main system.

According to the adaptive standard framework, medical expert has an important role in problem understanding, data understanding and data post-processing steps in following phases as mentioned:

- Determining medical goals.
- Problem definition from medical view.
- Presenting basic information about problem.
- Presenting metrics.
- Explaining data and their importance.
- Assessing data validity.
- Collecting data for final test.
- Verification of extracted knowledge by background knowledge.
- Making suggestions for correcting steps and activities.

3.2.5. Medical data mining trend

The goal of this Section is to draw medical data mining trend during 1999 through 2013 and comparing it with data mining trend. The idea behind this section is to prove that application of data mining is hot topic and researcher's tendency is rising. On the other hand, by comparing data mining with medical data mining trend we can show that rising medical data mining tendency is motivated by the rising in data mining applications.

In Fig. 12, medical data mining trend is presented and compared with data mining. It must be noted that data mining trend is depicted with 216 papers until 2013 based on another surveys (Liao, Chu, & Hsiao, 2012) and medical data mining trend is drawn by literature.

Amount of research conducted in the field of medical data mining has been rising similar to data mining trend, particularly between years 2005 and 2010. The reason for this can be as follow:

- Recent developments of bioinformatics and subsequently, drawing more attention to data mining applications in microarray and genetic.
- Success of data mining in other areas such as CRM.
- Promotion of data mining from an academic science to common efficiency analytic tools (Piatetsky-Shapiro, 2007).

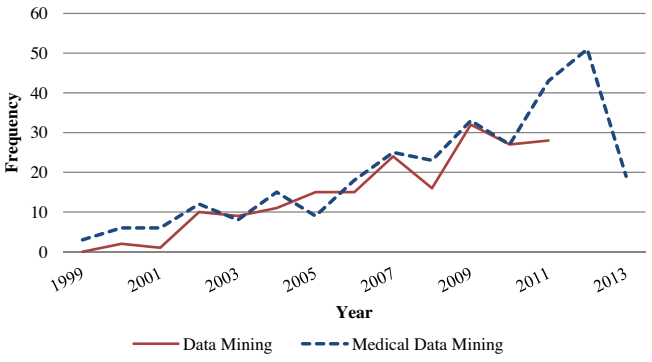

Fig. 12. Medical data mining trend compared with data mining.

Table 6
Journals with the highest frequency publication in medical data mining.

Journal
Expert Systems with Applications
Journal of Medical Systems
Artificial Intelligence in Medicine
IEEE Transactions on Information Technology in Biomedicine
Information Sciences
Applied Soft Computing
Journal of Biomedical Informatics
Computers in Biology and Medicine
IEEE Transactions on Biomedical Engineering
Computer Methods and Programs in Biomedicine
Knowledge-Based Systems
IEEE Engineering in Medicine and Biology
Decision Support Systems

Fig. 12 only cover a three month of 2013. In Table 6, list of journals with the highest number of publications based on literature in this field are presented.

4. Review and analysis

All activities in medicine can be divided into six domains that we know as "medical tasks": screening, diagnosis, treatment, prognosis, monitoring and management. Some of papers under review can be assigned to multiple tasks, for example in *"exploring effects of smoking as a heart disease risk factor"* problem, *"stopping smoking"* and *"smoking status of patient"* can be considered as treatment and diagnosis tasks, respectively. Recently, most effort has been devoted to the diagnosis and moving treatment forward in time. The goal of this Section is to study each paper from medical tasks perspective and then categorize them based on data mining algorithms. In each task, a brief summarization and discussion can be shown.

Among various disease, there are some top mortal diseases in the world (WHO, 2013) such as: heart, cancer and diabetes. Nowadays, application of data mining in these diseases is more common. This is because of high epidemic, requirement of special experience and early detection, expensive and time consuming tests.

4.1. Screening

Screening consists of identifying an unrecognized disease before appearance of its symptoms or signs. Screening is performed on persons who are apparently in good health. Activities are similar in screening and diagnostic processes. The difference is that in the latter, symptoms have already appeared and examined person is in an uncomfortable condition. We can prevent progress of diseases more efficiently by earlier detection.

4.1.1. Classification

Mammography can be used as a screening tool to detect breast cancer at early stage. The efficiency of mammography is limited to 60–70%. It is possible to make a prediction model for detecting breast masses by ensemble approach (Luo & Cheng, 2012), SVM and artificial neural network (Ramos-Pollán et al., 2012) based on mammography data. Early diagnosis of pancreatic cancer based on ensemble approach (Ge & Wong, 2008), identification of cancer-related objects to diagnose various cancers by graph theory (Pospisil, Iyer, Adelstein, & Kassis, 2006). In Roman et al. (2012), colorectal cancer was diagnosed early by decision tree and artificial neural network. Prostate cancer can be screened with artificial neural network and SVM (Çınar, Engin, Engin, & Ziya Ateşçi, 2009).

Chen et al. (2008) provides a real time decision system to investigate hypovolemic state in trauma patients when they are trans-

ferred to the hospital. This system uses basic sequence of vital-signs as time-series input for ensemble learning. The AUC of this approach is 0.76. Less computational requirements allows this system to be used as a continuous monitoring system of trauma patients.

Some works in this scope are as follows:

Using artificial neural network and decision tree to determine preterm birth risk factors (Chen, Chuang, Yang, & Wu, 2011); Combining ensemble classifier and neighborhood rough set as a tool for tumor classification and early prediction of cancer (Wang, Li, Zhang, Gui, & Huang, 2010); Determination of significant biomarkers in conversion of Hepatitis B to liver cancer by evolutionary rule-based classifier (Kwong-Sak et al., 2011); Making a comparison study on artificial neural network, self organization map, naïve bayes, decision tree and SVM to early diagnosis of breast cancer (Nahar, Imam, Tickle, Shawkat Ali, & Chen, 2012); Screening and developing early-warning systems for ailments such as hypertension, diabetes, cardiovascular, liver and renal disease by *K*-nearest neighbor (Jen, Wang, Jiang, Chu, & Chen, 2012); Exploring risk factors associated with diabetes and pre-diabetes and comparing performance of logistic regression, artificial neural network and decision tree (Meng, Huang, Rao, Zhang, & Liu, 2013); Analyzing accuracy of logistic regression, association rule, decision tree, Bayesian classifiers, artificial neural network and SVM for prediction of childhood obesity (Zhang, Tjortjis, et al., 2009); control diabetes by decision tree and naïve Bayes (Huang, McCullagh, Black, & Harper, 2007); overcome some drawbacks such as time complexity and curse of dimensionality by singular value decomposition combined with SVM (Simek et al., 2004).

4.1.2. Regression

Chang, Wang, and Jiang (2011) proposed a two-phase analysis system to predict hypertension and hyperlipidemia using common risk factors. First, risk factors of each disease were explored separately by six data mining algorithms. Then, voting methods were used to extract common risk factors. Finally, a system was made to predict hypertension and hyperlipidemia simultaneously by Multivariate Adaptive Regression. Briones and Dinu, explored risk factors associated with Alzheimer's disease by logistic regression and random forest (Briones & Dinu, 2012). Relations between vaccination and risk of preterm birth can be Discover by regression algorithm (Orozova-Bekkevold, Jensen, Stensballe, & Olsen, 2007).

4.1.3. Clustering

Data clustering is known as a very important method in gene expression data for extracting underlying information. Engreitz, Daigle, Marshall, and Altman (2010) tried to extract information from transcriptional modules in microarray data for acute myelogenous leukemia. In this paper, independent component analysis and two step pipeline can be used for dimensionality reduction and data normalization, respectively. Mueller et al. (2005) explored potential protein biomarkers to identify individuals at high risk of bladder cancer by supervised fuzzy clustering. Also colon cancer have been Screened by DBSCAN (Antonelli et al., 2012).

4.1.4. Association rule analysis

It is possible to make a prediction model for detecting breast masses by association rule (Mohanty, Senapati, & Lenka, 2012) based on mammography data; Early prediction of required hospitalization for hemodialysis patients by apriori (Yeh, Wu, & Tsao, 2011) and Predicting acute myocardial infarction in young patients (Lee, Ryu, Bashir, Bae, & Ryu, 2013).

4.1.5. Hybrid approaches

Yamaguchi, Kaseda, Yamazaki, and Kobayashi (2006) proposed a model to predict instability of glucose level in the next days by stepwise method and cluster analysis. In this study, time-series data was collected for more than five months from four patients with the type 1 diabetes. Variables with positive correlation with blood glucose were chosen. Practical results show that variable selection is very effective in this criterion.

Plant et al. (2010) developed a new framework by combining SVM, naïve bayes and voting feature intervals. This approach explores patterns that describe the conversion from Mild Cognitive Impairment to Alzheimer's disease. Important features were extracted based on DBSCAN algorithm from MRI images.

Pattern mining can be discovered compound-risk factors of leukemia cancer (Jinyan & Qiang, 2007); brain, prostate, breast and lung cancer can also be screened by genetic programming with microarray data (Paul & Iba, 2009).

Combination of genetic programming, genetic algorithm and decision tree cab be used for cardiovascular diseases (Podgorelec, Kokol, Stiglic, Heričko, & Rozman, 2005); Genes associated with multi-factoring diseases are Exploring by *K*-means + genetic hybrid algorithm (Jourdan, Dhaenens, Talbi, & Gallina, 2002); SVM + case base reasoning are Deploying to screen macular degeneration (Hijazi, Coenen, & Zheng, 2012).

4.1.6. Summarization and discussion

In Table 7, review papers in this task are summarized due to the data mining algorithm and disease. Other disease referring to cases that only have one paper and contains: childhood obesity, hepatitis, hypertension, dialysis and trauma. Stacked bar graph of papers associated with screening based on data mining algorithms and disease is depicted in Fig. 13. The disease that seen in this figure are elicit from Table 7. We can show that classification is the most widely used than others. Also cancer has more attention. With respect to Section 3.2.2 microarray data almost used in cancer screening that data reduction algorithms are more suitable. Therefore in Fig. 13 we show that, clustering and hybrid algorithm are used in cancer more than other disease. In Alzheimer only predictive approach is used.

4.2. Diagnosis

Diagnosis procedure is the attempt to identify the nature of a disease. Results of this procedure are the basis of selecting an appropriate treatment method. Clinical and Para-clinical information such as laboratory, imaging and microarray can be able to help to the diagnosis process. The most important goals of data mining applications in diagnosis are accurate detection, early diagnosis and risk factors identification.

4.2.1. Classification

Diagnosis of cerebrovascular is complicated and requires reliable knowledge and expensive tests. Yeh, Cheng, et al. (2011) made an accurate predictive model to diagnose cerebrovascular by decision tree, Bayesian classifier and back-propagation neural network. They concluded that using decision tree provides more accuracy.

Coronary Artery Disease (CAD) known as a most frequently cause of death in the world. In this disease, coronary arteries are narrowed and blood flow rate is reduced. Sudden death may occur in this case. It is clear that CAD should be detected and treated in an early stage. Some authors extracted prediction models to diagnose CAD based on multi-label learning (Liu, Li, Wang, & Wang, 2010) and ensemble approaches (Mandal & Sairam, 2012). Short-term heart rate variability system was proposed for chronic heart failure (Pecchia, Melillo, Sansone, & Bracale, 2011).

In Barakat, Bradley, and Barakat (2010) and Jin, Tang, and Zhang (2007)) the authors have tried to develop a system based on SVM for accuracy improvement. This system can be used for early diagnosis of diabetes, heart and liver diseases. In the first paper,

Table 7
Medical data mining papers associated with screening.

Disease	Classification	Regression	Clustering	Association rules	Hybrid
Diabetes	Huang et al. (2007) and Meng et al. (2013)				Jourdan et al. (2002)
Thyroid	Simek et al. (2004)			Li, Fu, and Fahey (2009)	
Cancer	Çınar et al. (2009), Wang et al. (2010), Roman et al. (2012), Nahar et al. (2012), Lee and Leu (2011), Li et al. (2013), Li, Fang, et al. (2009), Yoon et al. (2008), Bontempi (2007), Pospisil et al. (2006), Ge and Wong (2008), Luo and Cheng (2012) and Ramos-Pollán et al. (2012)		Mueller et al. (2005), Engreitz et al. (2010) and Antonelli et al. (2012)	Mohanty et al. (2012)	Ghazavi and Liao (2008), Shah and Kusiak (2007) and Paul and Iba (2009)
Alzheimer	Illán et al. (2011) and Ramírez et al. (2013)	Briones and Dinu (2012)			Plant et al. (2010)
Preterm birth	Chen et al. (2011)	Orozova-Bekkevold et al. (2007)			
Heart	Jen et al. (2012)				Podgorelec et al. (2005)
Others	Zhang, Tjortjis, et al. (2009), Kwong-Sak et al. (2011) and Chen et al. (2008)			Lee et al. (2013) and Yeh, Cheng, et al. ('2011)	Hijazi et al. (2012) and Chang et al. (2011)

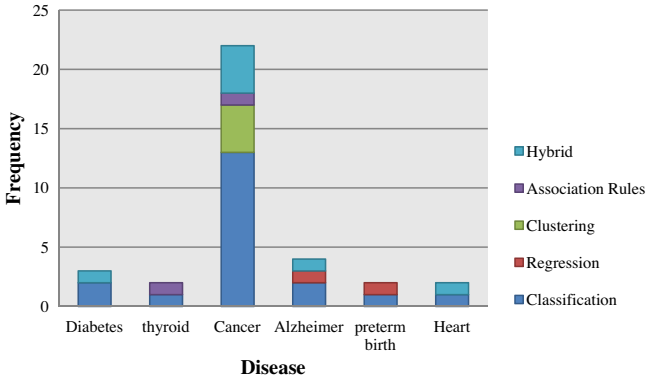

Fig. 13. Frequency of published papers in screening.

initially, sub-sampling and *K*-means were performed to overcome the problem of imbalance class distribution and to decrease number of negative examples. This step led to production of a model that is robust to skew data and is more sensitive to false positive examples. In the next step, SVM models were generated based on varying the misclassification cost. Finally, SQRex- SVM was applied to extract best rules from each model. Also artificial neural network can be used for diagnosis diabetes (Temurtas, Yumusak, & Temurtas, 2009).

Ensemble SVM classifier (Samuel et al., 2013) has been used for early diagnosis of Alzheimer's Disease. The use of NMF- SVM (Non-negative Matrix Factorization) (Padilla et al., 2012), has been used for early diagnosis of Alzheimer's Disease. Sleep Apnea Syndrome can be recognized by SVM (Al-Angari & Sahakian, 2012). In Riganello, Candelieri, Quintieri, Conforti, and Dolce (2010), Tsumoto (1998) and Tsumoto (2004), they have made prediction models based on classification algorithms for brain processes in vegetative state and diagnosis of headache, meningitis and cerebrovascular diseases.

Cancer diagnosis by proteomic data leads to increase in accuracy and decrease in time. Li et al. (2004) proposed ovarian cancer diagnosis system. This approach building a prediction model by SVM. As another instance, in Abbass (2002), Cruz-Ramírez, Acosta-Mesa, Carrillo-Calvet, and Barrientos-Martínez (2009), Kuo, Chang, Chen, and Lee (2001), and Perner (2002) breast cancer was diagnosed by naïve bayes, decision tree and multi objective artificial neural network. It has been shown that, the best ratio for partitioning data into training and test is 80–20% (Polat & Güneş, 2007). Breast cancer can be diagnosed by Least Square SVM (Polat & Güneş, 2007) and decision tree (Azar & El-Metwally, 2012). Hepatocellular Carcinoma (HCC) is known as the most dangerous cancer due to not being diagnosed until advanced tumor stages. Luk et al. (2007) tries to obtain profiling data and identify significant patterns to provide discrimination between HCC and non-malignant liver tissues. This approach performs based on artificial neural network and decision tree.

A system was developed for diagnosis of liver diseases by ensemble classifier Wang, Ma, and Liu (2009) and combined case base reasoning + CART algorithm Lin (2009). SVM has been used for diagnosis of polycythemia vera (Kantardzic, Djulbegovic, & Hamdan, 2002) and artificial neural network for glaucoma (Mookiah et al., 2012). It is also possible to diagnose neonatal jaundice (Ferreira, Oliveira, & Freitas, 2012) and pressure ulcer development in surgical patients (Su, Wang, Chen, & Chen, 2012) by classification algorithms.

Fakih and Das (2006)) proposed a methodology to provide a robust mechanism for diagnosis. This methodology called *LEAD* that uses rough set, utility theory and reinforcement learning.

In time-series problem, network-based SVM and support feature machine can be seen as optimization approaches. Chaovalitwongse, Pottenger, Shouyi, Ya-Ju, and Iasemidis (2011) claims that, epilepsy can be predicted and diagnosed by this approach.

4.2.2. Regression

Su et al. (2006) provides a novel early diagnosis system for diabetes to reduce treatment expenditures and improve patients' health. This is done with the help of logistic regression, artificial neural network, decision tree, and rough set based on anthropometrical body surface scanning data.

Making a decision support system for diagnosis of dementia is a complicated task due to its complexity and lack of comprehensive tools. Mazzocco and Hussain (2012) improved efficiency of bayesian network by logistic regression. Sut and Simsek (2011) provided a comparison study on six regression trees for mortality prediction of head injury. Results show that boosted tree classifiers and regression have a higher performance.

4.2.3. Clustering

Unsupervised clustering methods can be applied to diagnosis of mental health problems such as schizophrenia (Diederich, Al-Ajmi,

& Yellowlees, 2007). This can be done by identification of potential hidden structures in cognitive performance (Silver & Shmoish, 2008). Fuzzy clustering can be used to diagnose thyroid (Straszecka, 2006). Combination of decision tree with hierarchical clustering (Roy Walker et al., 2004) has been used for early diagnosis of Alzheimer's Disease. Hirano, Sun, and Tsumoto (2004) presents a comparison study on different clustering algorithms for diagnosis of meningoencephalitis.

4.2.4. Association rule analysis

Association rule (Chaves, Ramírez, Górriz, & Puntonet, 2012), has been used for early diagnosis of Alzheimer's Disease, detected Cardiac arrhythmia and ischemic episode automatically (Exarchos, Papaloukas, Fotiadis, & Michalis, 2006), diagnosis contributing heart disease factors (Nahar, Imam, Tickle, & Chen, 2013a) and extract diagnostic rules for lung cancer (Qiang et al., 2007).

Imberman, Domanski, and Thompson (2002) tries to find indications for requirement of computed tomography (CT) scanning in minor head trauma patients. In this paper, rules are initially extracted by association analysis and weighted by the expert user. Then, Probabilistic Interestingness Measure (PIM) is applied to introduce rules based on event dependency.

4.2.5. Hybrid approaches

Some attempts use hybrid approaches for diagnosis diseases as follow:

Vatankhah, Asadpour, and Fazel-Rezai (2012) and Bojarczuk, Lopes, and Freitas (2000) try to address pain identification by adaptive neuro fuzzy inference system + SVM hybrid and a combination of decision tree and genetic programming. Roychowdhury, Pratihar, Bose, Sankaranarayanan, and Sudhahar (2004) diagnosed jaundice and pneumonia based on an expert system and genetic fuzzy algorithm. Magoulas, Plagianakos, and Vrahatis (2004) investigated computer-assisted colonoscopy image diagnosis by evolutionary algorithm + artificial neural network. Used fuzzy *K*-nearest neighbor to early diagnosis of Parkinson's disease and reduce its effects (Chen et al., 2013), Sleep Apnea Syndrome can be recognized by fuzzy algorithm (Kwiatkowska, Atkins, Ayas, & Ryan, 2007) and SVM (Al-Angari & Sahakian, 2012). Early detection of hepatitis by temporal abstraction Ho, Nguyen, Kawasaki, Le, and Takabayashi (2007), Genetic programming has been combined with image processing techniques to detect lung abnormalities in an early stage (Choi & Choi, 2012), diagnosis colon, lymphoma and leukemia cancer by hybrid fuzzy + GA (Schaefer & Nakashima, 2010) and prostate, lung and ovarian cancer by PCA, DT and SVM algorithms (Lee, Rodriguez, & Madabhushi, 2008). Diagnosed Heart disease by fuzzy genetic algorithm (Lahsasna, Ainon, Zainuddin, & Bulgiba, 2012) and rotation forest ensemble (Karabulut & İbrikçi, 2012). diagnosed breast cancer by isotonic separation technique (Ryu, Chandrasekaran, & Jacob, 2007), probabilistic greedy heuristic (Kohli, Krishnamurti, & Jedidi, 2006) and hybrid approach (artificial neural network + SVM + fuzzy) (Hassanien & Kim, 2012).

In Patil et al. (2010), the authors have proposed a novel system to predict type 2 diabetes by hybrid methods. First, *K*-means has been applied to validate chosen classes. Finally, model has been built based on C4.5. In Jin et al. (2007), features were transformed into high feature space by fuzzy methods and then SVM was used to construct the model for diagnosis Diabetes, liver and Heart disease.

Asymptomatic carotid stenosis is considered as an important factor of stroke. It has several risk factors such as smoking, hypertension, diabetes, cardiac diseases and physical inactivity. Bilge, Bozkurt, and Durmaz (2013) discovered rules between these risk factors and asymptomatic carotid stenosis by hybrid approach of genetic algorithm and regression.

A fuzzy rule based system was proposed for diagnosis of coronary artery disease (Tsipouras et al., 2008). In this approach, a set of crisp rules was extracted from induced decision tree. These crisp rules were converted into fuzzy rules. Performance was increased by optimization of fuzzy parameters in this work. Fuzzy systems can also be combined with artificial neural network (Kahramanli & Allahverdi, 2008) and PSO + decision tree (Muthukaruppan & Er, 2012) to improve accuracy.

A diagnostic system for breast cancer was also proposed by integrating artificial neural network with Multivariate Adaptive Regression (Chou, Lee, Shao, & Chen, 2004). In Chen and Hsu (2006), effective genes and significant indicators associated with breast cancer were discovered by hybrid of genetic algorithm + decision tree.

Uçar and Karahoca (2011) predicted presence of Mycobacterium tuberculosis on patients by means of adaptive neuro fuzzy inference system. This would help early diagnosis of tuberculosis. Normally, it is a very time consuming process. Their proposed approach outperforms partial decision tree and multi-layer perceptron.

An accurate determination of Intra-Cranial Pressure (ICP) system by data mining algorithm and nonlinear mapping function, are presented in Sunghan et al. (2012) with the aim of decreasing costs of skull disease. Wongseree, Chaiyaratana, Vichittumaros, Winichagoon, and Fucharoen (2007) used a combination of artificial neural network and genetic programming to diagnose Thalassaemia. Chaochang et al. (2007) built a diagnostic model for Hypertension by a hybrid of genetic algorithm, apriori and decision tree. They have claimed that their proposed system is more accurate.

Some attempts have been made to develop a hybrid system for decreasing human error and increasing accuracy of extracted knowledge. In Chuang (2011), CART algorithm was used to predict presence of a disease. Then case base reasoning was added to classify type of the disease. naïve bayes can also be combined with evolutionary algorithms to improve validity of extracted knowledge (Raymer, Doom, Kuhn, & Punch, 2003). In Fan, Chang, Lin, and Hsieh (2011) a new hybrid system was proposed for medical data classification. In this approach, first, case-based clustering was applied in data pre-processing and then a model was extracted by a combination of fuzzy decision tree and genetic algorithm. Finally, a set of fuzzy decision rules is generated for each cluster. Pham and Triantaphyllou (2009) proposed a new meta-heuristic approach called *Homogeneity-Based Algorithm* to explore medical data. This approach is classified as one of the optimization algorithms and is combined with classification methods to improve their performance. In a similar way, swarm optimization (Yeh, 2012) can be used to improve performance of classification.

4.2.6. Summarization and discussion

In Table 8, review papers in this task are summarized due to the data mining algorithm and disease. Other disease referring to cases that only have one paper and contains: Polycythemia Vera, Trauma, headache, colonoscopic, Thalassaemia, Brain Injury, Cerebrovascular, Glaucoma, Parkinson, Fever, Mycobacterium tuberculosis, Dermatology (skin), physiotherapy, congenital malformation, Hepatitis, Pressure Ulcer, Jaundice and Rheumatic. Stacked bar graph of papers associated with screening based on data mining algorithms and disease is depicted in Fig. 14. The disease that seen in this figure are elicit from Table 8. We can show that classification and hybrid is the most widely used than others. Also cancer, heart and diabetes more frequent, respectively. In general, diagnosis has more attention and covering more disease than other tasks.

4.3. Treatments

All activities are done to remedy the health problem after disease detection. The main goals of treatment are regaining health

Table 8
Medical data mining papers associated with diagnosis.

Disease	Classification	Regression	Clustering	Association rules	Hybrid
Cancer	Perner (2002), Abbass (2002), Li et al. (2004), Luk et al. (2007), Polat and Güneş (2007), Cruz-Ramírez et al. (2009), Dai and Xu (2013), Warren Liao (2011), Samanta et al. (2009), Peterson and Ringnér (2003), Au et al. (2005), Shenghuo, Dingding, Kai, Tao, and Yihong (2010), Kuo et al. (2001), Ozcift (2012) and Azar and El-Metwally (2012)		Shaik and Yeasin (2009)		Chou et al. (2004), Chen and Hsu (2006), Kohli et al. (2006), Ryu et al. (2007), Fan et al. (2011), Hassanien and Kim (2012) and Schaefer and Nakashima (2010)
Alzheimer	Padilla et al. (2012)		Roy Walker et al. (2004)	Chaves et al. (2012)	
Meningitis	Pechenizkiy et al. (2006)		Hirano et al. (2004)		
Thyroid			Straszecka (2006)		Yeh (2012)
Mental			Diederich et al. (2007) and Silver and Shmoish (2008)		
Diabetes	Temurtas et al. (2009), Su et al. (2006) and Barakat et al. (2010)	Su et al. (2006)			Jin et al. (2007), Kahramanli and Allahverdi (2008), Pham and Triantaphyllou (2009), Patil et al. (2010) and Raymer et al. (2003)
Liver	Wang et al. (2009), Lin (2009)				Chuang (2011)
Heart	Nahar et al. (2013b), Martis et al. (2012), Šajn and Kukar (2011), Shilaskar and Ghatol (2013), Wiggins et al. (2008), Sacha, Cios, and Goodenday (2000), Cilla et al. (2012), Pecchia, Melillo, Sansone, et al. (2011), Liu et al. (2010), Karabulut and İbrikçi (2012) and Mandal and Sairam (2012)	Cilla et al. (2012) and Filipovic et al. (2011)		Nahar et al. (2013a), Exarchos et al. (2006) and Imamura et al. (2007)	Muthukaruppan and Er (2012), Bilge et al. (2013), Tsipouras et al. (2008), Karabulut and İbrikçi (2012) and Lahsasna et al. (2012)
Pulmonary	Fakih and Das (2006)				Choi and Choi (2012)
Pain					Vatankhah et al. (2012) and Bojarczuk et al. (2000)
Hypertension					Hsu et al. (2011) and Chaochang et al. (2007)
Epileptic	Chaovalitwongse et al. (2011)			Bourien et al. (2004)	
Sleep disorders	Al-Angari and Sahakian (2012)			Laxminarayan et al. (2006)	Kwiatkowska et al. (2007) and Alonso et al. (2002),
Other disease	Su et al. (2012), Shusaku (2000), Chang and Chen (2009), Samuel et al. (2013), Froelich et al. (2012), Mookiah et al. (2012), Yeh, Wu, et al. (2011), Riganello et al. (2010), Tsumoto (2004), Kantardzic et al. (2002) and Ferreira et al. (2012)		Yildirim, Çeken, Hassanpour, and Tolun (2012)	Imberman et al. (2002)	Ho et al. (2007), Chen et al. (2013), Subasi (2012), Hsu and Ho (2004), Wongseree et al. (2007), Roychowdhury et al. (2004) and Magoulas et al. (2004)

and preventing disease progression. Most of works in these tasks can be summarized as identification of effective treatment factors, determination of appropriate methods and medications based on drug side effects.

4.3.1. Classification

Shah, Kusiak, and O'Donnell (2006) proposed a model to predict successful bladder cancer treatments and identify important parameters in therapy by SVM, decision tree and decision rules. The advantage of this model is being able to provide effective treatment guidelines for each patient. Also Farion, Michalowski, Wilk, O'Sullivan, and Matwin (2010), Suner, Çelikoğlu, Dicle, and Sökmen (2012) and Ting, Wu, Chan, Lin, and Chen (2010) used decision tree to help treatment of Appendicitis, Asthmain children and Rectal Cancer respectively.

Toussi, Lamy, Le Toumelin, and Venot (2009) proposed a model that can fill the knowledge gap between medical guidelines and physicians' therapeutic decisions for the management of type 2 diabetes by C.5 algorithm. This gap occurs in the situation where clinical guidelines cannot cover or provide recommendations for special clinical conditions. KNN and ANN can be used for treatment Brain injury (Guenther et al., 2009) and Diabetes type2 (Lee, Lee, & Liew, 2013) respectively.

Karaolis, Moutiris, Hadjipanayi, and Pattichis (2010) assessed heart event-related risk factors to reduce coronary heart disease events based on decision tree. The risk factors have been investigated in three situations: before the event, during the event and after the event. In this approach each of the risk factors are classified into non-modifiable and modifiable groups. Then, for each group, importance of risk factors is determined.

Acquired brain injury is a dangerous event in the trauma. Brain injury leads to death and disability in a higher rate and imposes high health care costs. Patients often require acute treatment and long term rehabilitation. Marcano-Cedeño et al. (2013) proposed models to predict outcomes of cognitive rehabilitation with acquired brain injury. This was done based on decision tree, multi-

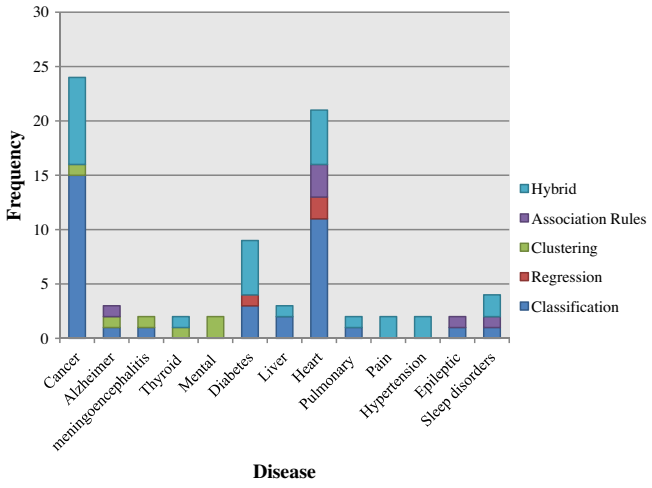

Fig. 14. Frequency of published papers in diagnosis.

layer perceptron and general regression neural network. Naïve bayes can be used for predict the need for CT scanning in children with minor head injuries (Klement et al., 2012). Crash scene data can be used to evaluate severe and vehicular injuries by CART algorithm. This helps in transporting patients into appropriate trauma-care (Scheetz et al., 2009).

4.3.2. Regression

Pre-hospital trauma criteria are used to facilitate the treatment of severely injured patients. Cox et al. (2012), has extracted a model based on multivariate logistic regression in the first stage to determine pre-hospital rules. In the second stage, determined rules have been reviewed and an improved model has been replaced. Almazyad, Ahamad, Siddiqui, and Almazyad (2010) present a system for treatment Hypertension disease by regression.

4.3.3. Clustering

Liao and Tsai's work, identifies important DNA viruses in breast cancer progression by clustering methods with the aim of improving treatment process (Liao & Tsai, 2007). In mentioned paper, combinations of DNA viruses associated with risk factors such as inheritance are extracted by artificial neural network. Their common characteristics are also determined by agglomerative hierarchical clustering technique. There has also been some works focusing on other types of cancer such as: prediction of prostate cancer by FCM (Froelich, Papageorgiou, Samarinas, & Skriapas, 2012).

Tang et al. (2010), provided a useful system to analyze H1N1 influenza and developed a treatment system by hierarchical clustering.

4.3.4. Association rule analysis

Association rules can be used for detect Adverse drug reaction (Erxin, Liang, Xinsheng, Yuping, & Jinao, 2010; Huidong et al., 2008) and treatment Bladder, Breast, Cervical, Lung, Prostate and Skin Cancer (Nahar, Tickle, Ali, & Chen, 2011). In Yanqing et al. (2011), fuzzy association rule was performed to predict unknown adverse drug reactions.

4.3.5. Hybrid approaches

Grosan et al. proposed a novel multiple criteria procedure for evolutionary algorithms referring to the ranking of Trigeminal Neuralgia treatments (Grosan, Abraham, & Tigan, 2008). The result of this approach is similar to a standard mathematical approach for multi objective optimization. They differ in the fact that proposed approach does not require any additional information.

In Shah and Kusiak (2004), breast cancer was predicted using rule-based algorithms based on genetic algorithm. Instance-based learning, ANN and Decision table can be used for treatment process of breast cancer (Çakir & Demirel, 2011).

It is important to predict hospitalization of hemodialysis patients and perform immediate treatment. Yeh, Wu, et al. (2011) proposed a hybrid system for rule production in order to avoid hospitalization. Temporal abstraction algorithm with C4.5 and apriori were used to generate helpful rules.

4.3.6. Summarization and discussion

In Table 9, review papers in this task are summarized due to the data mining algorithm and disease. Other disease referring to cases that only have one paper and contains: H1N1 influenza, Appendicitis, antidepressant and ICU. Stacked bar graph of papers associated with treatment based on data mining algorithms and disease is depicted in Fig. 15. The disease that seen in this figure are elicit from Table 9. We can show that classification and association are the most widely used than others. Also with the respect to Table 9, cancer, drug application and trauma have more attention. In drug application researches are focused on discovering drug interactions. Association rule mining is more conform to the nature of this domain. In trauma only predictive approach is used. This is because, in trauma most of works focus on predict severity of injury and determine appropriate bed.

4.4. Prognosis

Prognosis is predicting current disease outcomes in terms of morbidity and mortality and patient's chances of recovery. The most important applications of data mining in this task are predicting survivability and recurrence of diseases. This method has been compared with traditional statistical techniques. Saving time and cost are the benefits of accurate prognosis.

4.4.1. Classification

De Falco (2013) extracted knowledge by differential evolution in the form of *if*...*then*... rules to predict outcome of diseases. In the this approach, results of data mining was compared with an expert oncologist's work and a less experienced one. Data mining method proved to perform better than the less experienced person. Mortality of a patient admitted to ICU can be predicted by multilayer artificial neural network (Silva, Cortez, Santos, Gomes, & Neves, 2006). Kusiak, Dixon, and Shah (2005) made a model to predict outcome of kidney dialysis with the aim of improving patient outcomes and reduce the cost of dialysis by ensemble decision tree. A method based on combination of rough set and decision tree was presented for Arrhythmia study in the case of children born with a malformation of the heart in Kusiak, Law, and MacDonald (2001). Wang et al. (2013) discovered a method for reducing recurrence of Endometriosis by decision tree. Better diagnosis and treatment, are results of this system. Delen, Oztekin, and Tomak (2012) provides a solid study on coronary surgeries with the aim of achieving better understanding and management of surgeries. The authors have considered artificial neural network, SVM, CART and C5 and concluded that SVM performed better. Rough set-based multiple criteria linear programming (Zhang, Shi, & Gao, 2009) and swarm optimization (Yeh, 2012) can be used to improve performance of classification. Survivability of HIV (Ramirez, Cook, Peterson, & Peterson, 2000) and burn patients (Patil, Joshi, Toshniwal, & Biradar, 2011) can be predicted by Classification algorithms.

Age is the most influential parameter on outcome of surgery according to this algorithm. In Hsieh, Hung, Shih, Keh, and Chan

Table 9
Medical data mining papers associated with treatment.

Disease	Classification	Regression	Clustering	Association rules	Hybrid
Cancer	Shah, Kusiak, and O'Donnell (2006), Suner et al. (2012), Bertolazzi et al. (2008), Aussem, de Morais, and Corbex (2012) and Lee, Lushington, and Visvanathan (2011)		Liao and Tsai (2007) and Froelich et al. (2012)	Nahar et al. (2011))	Shah and Kusiak (2004) and Çakır and Demirel (2011)
Asthma	Farion, Michalowski, Wilk, O'Sullivan, and Matwin (2010)				Grosan et al. (2008)
Trauma	Scheetz et al. (2009) and Klement et al. (2012)	Cox et al. (2012)			
Drug application			Altiparmak et al. (2006)	Erxin et al. (2010), Huidong et al. (2008) and Yanqing et al. (2011)	
Diabetes	Lee et al. (2013) and Toussi et al. (2009)				
Stress	Chen et al. (2010)			Panagiotakopoulos et al. (2010)	
Brain injury	Marcano-Cedeño et al. (2013) and Guenther et al. (2009)				
Other disease	Taft et al. (2009) and Ting et al. (2010)	Almazyad et al. (2010) and Huang, Wulsin, Li, and Guo (2009)	Tang et al. (2010)		Gülçin Yıldırım, Karahoca, and Uçar (2011)

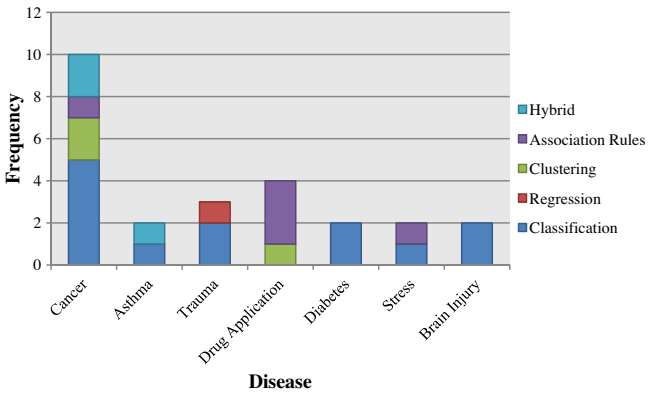

Fig. 15. Frequency of published papers in treatment.

(2012) a model was also made to predict post-operative morbidity after Endovascular Aneurysm Repair by ensemble approach.

Jonsdottir, Hvannberg, Sigurdsson, and Sigurdsson (2008) tries to address breast cancer survivability. In this paper, feature selection was done by three ways: manual, semi-automatic (combined manual and data mining) and automatic. The authors have considered expert knowledge as an individual field to improve performance. Several predictive models were built separately based on seventeen algorithms. Between all these models, decision tree and naïve bayes proved to perform the best.

It is important to investigate indicators that can discover dependencies between first observations of patients in hospital and early death. For this reason, rule induction algorithm was performed for prognosis of diabetes (Richards, Rayward-Smith, Sönksen, Carey, & Weng, 2001). Trauma surgeons have a decision making problem when facing severe traumatic injury patients. For this reason, Demšar et al. (2001), has proposed prognostic models to predict first surgery outcome of severe trauma patients.

4.4.2. Regression

In some chronic failure of the major organs, transplantation is the only treatment. The main limitations of organ transplantation are shortage of donor and large number of requests. However, there are several rejections and failures due to mismatch of donor and patient. For this reason, Oztekin, Delen, and Kong (2009) proposed a model to predict the outcome of graft surgery and survivability. Initially, the model was built based on regression. Then, the designed model was repaired according to expert knowledge and literature research. The advantages of this study are overcoming the problems of traditional and statistical approaches and its ability to extract a model from large amount of variables.

Regression can be used to estimate outcome of brain trauma (Niki Kunene & Roland Weistroffer, 2008), Success of weight reduction after bariatric surgery (Lee et al., 2007) and survivability of breast cancer (Delen, Walker, & Kadam, 2005; Lee, Mangasarian, & Wolberg, 2003).

Making a decision support system for diagnosis of dementia is a complicated task due to its complexity and lack of comprehensive tools. Mazzocco and Hussain (2012) improved efficiency of bayesian network by logistic regression. Sut and Simsek (2011) provided a comparison study on six regression trees for mortality prediction of head injury. Results show that boosted tree classifiers and regression have a higher performance. In Chen et al. (2007) a prognostic system for lung cancer was proposed based on microarray data. In this study, impact of trophinin on metastasis was explored by regression.

4.4.3. Clustering

Survivability of breast cancer can be predicted by *K*-means (Delen et al., 2005; Lee et al., 2003). Results show that, grade, stage of cancer, radiation and number of primaries are the most prognostic factors.

4.4.4. Association rule analysis

Literature-based discovery by association rules was used to explore candidate genes for diseases (Hristovski, Peterlin, Mitchell, & Humphrey, 2005).

4.4.5. Hybrid approaches

A prediction model has been made to explore metastasis or recurrence of breast cancer by decision tree and genetic algorithm

(Razavi, Gill, Åhlfeldt, & Shahsavar, 2007; Šprogar, Lenič, & Alayon, 2002); also incidence of cardiovascular as a result of renal disease in hemodialysis patients can be predicted by logistic regression and random forest (Ion Titapiccolo et al., 2013).

4.4.6. Summarization and discussion

In Table 10, review papers in this task are summarized due to the data mining algorithm and disease. Other disease referring to cases that only have one paper and contains: dialysis, Dementia, Cardiovascular, HIV, Burn Patient, and Liver. Stacked bar graph of papers associated with prognosis based on data mining algorithms and disease is depicted in Fig. 16. The disease that seen in this Figure are elicit from Table 10. We can show that classification and regression are the most widely used than others. This is because, regression known as the oldest statistical techniques that used for predict survivability. Also with the respect to Table 10, cancer, heart and trauma have more attention.

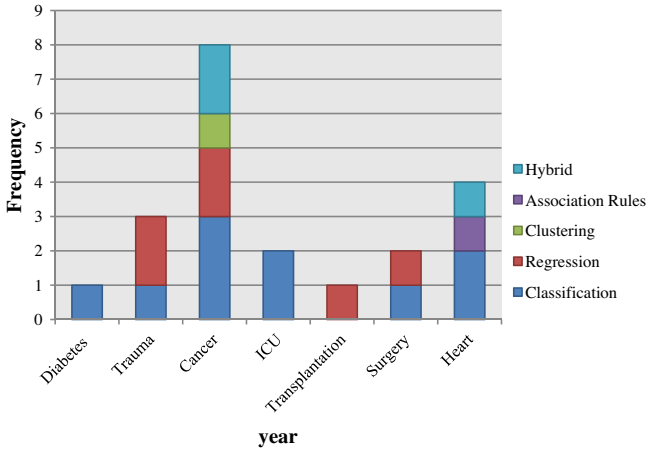

Fig. 16. Frequency of published papers in prognosis.

4.5. Monitoring

Monitoring is identified in medical domain as observation of disease and patient conditions over the time. Three important jobs in this task are continuously measuring certain parameters (e.g. body temperature), performing medical tests (e.g. blood test) and immediate detection of abnormalities (e.g. Cardiac arrest).

4.5.1. Classification

Abdominal pains in childhood such as appendicitis are highly prevalent and delayed diagnoses or treatments of them may be painful. Michalowski, Rubin, Slowinski, and Wilk (2003) provides a good mobile clinical support system by rough-set for monitoring children that admission for abdominal pains in the emergency room (ER).

Patients are admitted to the ICU after major surgery (e.g. cardiac surgery) or unplanned event (e.g. trauma) that requires full support and monitoring. Güiza, Eyck, and Meyfroidt (2012) has analyzed time series data collected from various electrical equipment in ICU, developed an early-warning system and examined long term outcome prediction model. Research shows that, with the increase of staying time in ICU, survivability rate decreases. Therefore, Ramon et al. (2007) proposed a prediction model for evolution of patients by classification algorithms.

Exploration of factors associated with glycemic control and determination of impact of educational interventions in type 2 diabetes by decision tree, are given by Sigurdardottir, Jonsdottir, and Benediktsson (2007). Automatic seizure detection can be made feasible with the help of epilepsy long-term monitoring. For this purpose, wavelet and SVM were used to predict long-term intracranial EEG (Yinxia, Weidong, Qi, & Shuangshuang, 2012). Pecchia, Melillo, and Bracale (2011) proposed monitoring remote systems for early detection of any deterioration inpatient's heart rate variability by feature extraction and CART algorithm.

4.5.2. Regression

Hemodialysis treatment needs nephrologist to plan it. It guarantees quality of Hemodialysis sessions. Bellazzi et al. (2012) describes a computerized system to plan the Hemodialysis scheduling. The main idea behind this work is to detect and report failure-to-adhere treatment by regression analysis of six clinical parameters. As a consequence, their results show decrease in failure-to-adhere planned treatment.

Gregori et al. (2011) made a comparative study on monitoring of diabetic patients. It has been shown that artificial neural network performs worse than logistic regression, generalized additive model, projection pursuit regression, linear discriminant analysis and quadratic discriminant analysis.

4.5.3. Clustering

Huang, Zhang, Cao, Steyn, and Taraporewalla (2013), proposed three novel algorithms based on clustering to reduce the data size of medical data. An effective data mining algorithm was developed with the aim of detecting and ranking abnormalities.

In Lin et al. (2010) has used two-staged clustering and decision tree to analyze abnormal patients admitted in emergency depart-

Table 10
Medical data mining papers associated with prognosis.

Disease	Classification	Regression	Clustering	Association rules	Hybrid
Diabetes	Richards et al. (2001)				
Trauma	Demšar et al. (2001)	Niki Kunene and Roland Weistroffer (2008)			
Cancer	Jonsdottir et al. (2008), Zhang, Shi, et al. (2009) and Thongkam et al. (2009)	Delen et al. (2005) and Chen et al. (2007)	Lee et al. (2003)		Šprogar et al. (2002) and Razavi et al. (2007)
(ICU)	Silva et al. (2006) and Frize et al. (2001)				
Transplantation		Oztekin et al. (2009)			Delen et al. (2010)
Surgery	Delen et al. (2012)	Lee et al. (2007)			
Heart	De Falco (2013) and Hsieh et al. (2012)			Kusiak et al. (2001)	Huang et al. (2007)
Others	Ramirez et al. (2000), Patil et al. (2011), Nakayama et al. (2012), Wang et al. (2013) and Kusiak et al. (2005)	Mazzocco and Hussain (2012)		Hristovski et al. (2005)	Ion Titapiccolo et al. (2013)

ment. This approach can be used in triage to classify patients with the aim of achieving effective utilization of distributed resource. Query-based self organization map (Chang, Chu, Wu, & Chen, 2010) can be used as well to overcome drawbacks such as time complexity and curse of dimensionality.

4.5.4. Association rule analysis

Produced alarm in ICU presented in Hu et al. (2012), has managed to predict cardiac arrest early using association rule mining. Klema, Novakova, Karel, Stepankova, and Zelezny (2008) tried to make a comparative study on windowing, episode rules, and inductive logic programming. This was followed by presenting a case study on atherosclerosis risk factors. More efficient rules can be achieved by rule pruning approach. One of these approaches is hybrid pruning method based on objective and subjective analysis and ontology. Performance of this method was evaluated by medical data sets (Mansingh et al., 2011).

4.5.5. Hybrid approaches

Chen, Larbani, Hsieh, and Chen (2009) tried to ensure patient safety and reduce error risk by affinity set. In this work, affinity set was compared with rough set, artificial neural network, SVM and logistic regression in the context of prediction accuracy and explanation power.

Works of Chittaro, Combi, and Trapasso (2003) and Gueguin et al. (2008), presented an optimized system to follow-up heart and renal failure disease by mining time series data; Czabanski, Jezewski, Matonia, and Jezewski (2012) used fuzzy- SVM and artificial neural network to monitor heart rate signal and predict neonatal acidemia. Cardiac arrhythmia and ischemic episode can be automatically detected by a combination of *K*-means + SVM.

4.5.6. Summarization and discussion

In Table 11, review papers in this task are summarized due to the data mining algorithm and disease. Other disease referring to cases that only have one paper and contains: Abdominal pain in childhood and Triage. Stacked bar graph of papers associated with prognosis based on data mining algorithms and disease is depicted in Fig. 17. The disease that seen in this Figure are elicit from Table 11. We can show that clustering, association and hybrid have similar frequency. Also with the respect to Table 11, heart, ICU and diabetes have more attention. This is because, monitoring in these disease known as a part of regaining health process.

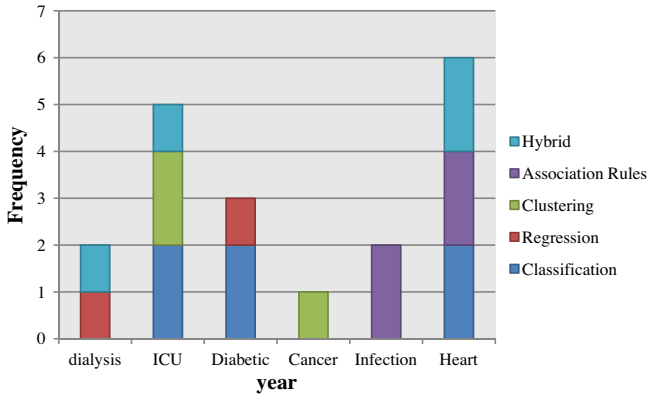

Fig. 17. Frequency of published papers in monitoring.

4.6. Management

The goals of management are health promotion and medical services. The most important data mining Approaches regarding this task are hospital resource management and staff scheduling.

4.6.1. Classification

Some authors have attempted to make systems for resource management and inpatient beds allocation by decision tree and artificial neural network (Delen, Fuller, McCann, & Ray, 2009; Isken & Rajagopalan, 2002). Goodwin, VanDyne, Lin, and Talbert (2003) proposed an assist system to obtain knowledge from clinical data, nursing interventions and outcomes by CART, artificial neural network and SVM. Performance of this approach has been illustrated in preterm risk prediction case study. A comparative study between artificial neural network and decision tree with the aim of predicting hospital charges on gastric cancer has been presented in Wang, Li, Hu, and Zhu (2009).

In Chae, Kim, Tark, Park, and Ho (2003) a decision support system was proposed to monitor trends of quality by decision tree. The result of this system is a guideline for quality improvement activities. Authors claim that, this approach can be integrated with hospital Order Communication System (OCS). Chi, Street, and Ward (2008) has built a hospital-selection system based on SVM. Breault, Goodall, and Fos (2002) proposed an approach to extract new rules from diabetes data warehouse that can be useful for clinicians and administrators. In this approach, time-series data was pre-processed initially and then CART algorithm was applied.

Wang, Lin, Wu, and Chaovalitwongse (2011) described an early detection system to identify numerical typing errors made by

Table 11
Medical data mining papers associated with monitoring.

Disease	Classification	Regression	Clustering	Association rules	Hybrid
Dialysis		Bellazzi et al. (2012)			Chittaro et al. (2003)
ICU	Ramon et al. (2007) and Güiza et al. (2012)		Apiletti et al. (2009) and Huang et al. (2013)		Chen et al. (2009)
Diabetic	Sigurdardottir et al. (2007) and Jannin and Morandi (2007)	Gregori et al. (2011)			
Cancer			Chang, Chu, et al. (2010)		
Infection				Mansingh et al. (2011) and Lamma et al. (2006)	
Heart	Shen, Kao, et al. (2012) and Pecchia, Melillo, and Bracale (2011)			Hu et al. (2012) and Klema et al. (2008)	Czabanski et al. (2012) and Gueguin et al. (2008)
Other disease	Michalowski et al. (2003)		Lin et al. (2010)		

human operators of EEG recordings using linear discriminant analysis and SVM. In Madigan and Curet (2006), a home health care system based on CART was proposed for older persons. Some attempts have been made to make scalable models for large scale data (Mullins et al., 2006; Semenova, 2004). The result of these models can be used to achieve a better health care system. Phillips-Wren, Sharkey, and Dy (2008) proposed a decision support system to assess the utilization of resources for lung cancer patients by combination of decision tree and artificial neural network. Triage is known as the key clinical component in trauma care. Its main task is to rapidly match patients to appropriate resources. On the other hand, mistriage leads to delayed allocation and patient's access to resource and increases mortality and morbidity. SVM, artificial neural network and decision tree with relevance feedback can be used to improve triage. This approach is more adaptive and dynamic than similar approaches (Talbert, Honeycutt, & Talbert, 2011).

Hypoplastic Left Heart Syndrome (HLHS) is a rare heart disease that affects infants. The only way to treat HLHS is surgery. Fetal rate is high in case of choosing non-surgical options. The most important challenge of surgical success is existence of numerous variables and requires post-operative management. Kusiak et al. (2006) discovered relations among these variables by rough set and extracted a real-time prediction model. This model predicted intervention of different variables for patients admitted to the ICU.

4.6.2. Regression

A simple estimation of surgery time has been addressed by Devi, Rao, and Sangeetha (2012). This approach was developed by Multiple Linear Regression Analysis, Adaptive Neuro Fuzzy Inference Systems and artificial neural network.

4.6.3. Clustering

Some authors have attempted to make systems for public health modeling by clustering (Lavrač et al., 2007). Shen, Jigjidsuren, et al. (2012) improved efficiency of medical resources utilization by clustering methods. As a case study on diabetic care, Antonelli et al. (2013) provides a system based on DBSCAN to identify examination pathways used by patients.

4.6.4. Association rule analysis

Koyuncugil and Ozgulbas have also proposed a system based on association rule mining to identify the appropriate donor in organ transplantations with respect to time constraints (Koyuncugil & Ozgulbas, 2010).

4.6.5. Hybrid approaches

The hybrid approach proposed by Shun Ngan, Leung Wong, Lam, Leung, and Cheng (1999) can be used to extract parameters associated with length of stay at hospital for limb fracture and classification of scoliosis. This approach contains bayesian network, Rule induction and Evolutionary computing. Bayesian network can be used to determine overall structure of the relationship between attributes. In López-Vallverdú, Riaño, and Bohada (2012) decision tree was combined with background knowledge for decreasing human error and increasing accuracy of extracted knowledge.

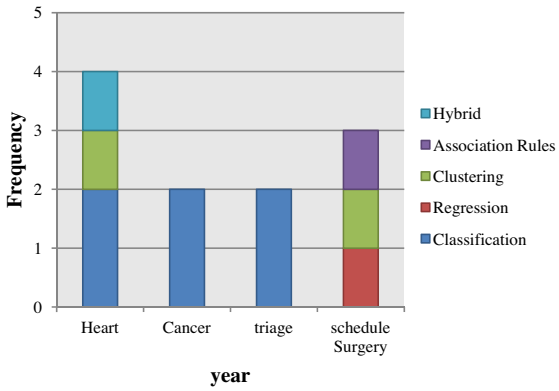

Fig. 18. Frequency of published papers in management.

Clustering combined with rough set and Hidden Markov Models with the aim of increasing quality of medical services and decreasing operative costs (Lin, Wu, Zheng, & Chen, 2011), and identifying common sequence of events based on insurance data (Tsoi, Zhang, & Hagenbuchner, 2005). A nice combination of clustering and SVM seems to have solved the problem of processing large data in Zhong, Chow, and He (2012).

4.6.6. Summarization and discussion

In Table 12, review papers in this task are summarized due to the data mining algorithm and disease. Other disease referring to cases that only have one paper and contains: total hip arthroplasty, thyroid, Mental, preterm birth prevention, Diabetes and Seizure. In this task some issue are important such as: resource management, quality of services, select best hospital, and schedule surgery. Stacked bar graph of papers associated with prognosis based on data mining algorithms and disease is depicted in Fig. 18. The disease that seen in this Figure are elicit from Table 12. We can show that classification and clustering are most used. Also with the respect to Table 12, heart and schedule surgery have more attention.

5. Discussion: current issue and future trend

In this Section, as a brief summarization, twelve important issues in medical data mining are considered according to the liter-

Table 12
Medical data mining papers associated with management.

Disease	Classification	Regression	Clustering	Association rules	Hybrid
Heart	Kusiak et al. (2006) and Madigan and Curet (2006)		Antonelli et al. (2013)		Zhong et al. (2012)
Cancer	Phillips-Wren et al. (2008) and Wang et al. (2009)				
Triage	Lin et al. (2011) and Yang et al. (2009)				
Schedule surgery		Devi et al. (2012)	Isken and Rajagopalan (2002)	Koyuncugil and Ozgulbas (2010)	
Others disease	Semenova (2004), Mullins et al. (2006), Chae et al. (2003), Chi et al. (2008), Delen et al. (2009), Cheng (2012), Yan et al. (2008), Tsoi et al. (2005), Yinxia et al. (2012), Breault et al. (2002), Goodwin et al. (2003) and Wang et al. (2011)		Lavrač et al. (2007), Shen, Jigjidsuren, et al. (2012) and Bentham and Hand (2012)		López-Vallverdú et al. (2012) and Chang, Yen, et al. (2010)

atures. In each item, more and less attention problems are analyzed. In the following, current issue and future trend are investigated. This can draw some guidelines to help researchers.

5.1. Brief summarization

5.1.1. Medical data mining goals

Among four medical data mining goals, little work has been carried out on decreasing time and cost, based on Fig. 6. This is because accuracy is more important in medicine and a few numbers of diseases have costly tests. Extracting hidden knowledge is also known as a main objective of data mining.

5.1.2. Types of data

Based on Fig. 7, non-numeric data such as image and signal should be converted and adapted to be used in common data mining algorithms. The main purpose of this task is feature selection. For example, features in image data are manually extracted by expert and in signal data, wavelet transformation is used. General weakness of using non-numeric data is that it requires extra step. Therefore, little effort has been devoted to them.

5.1.3. Integration

In medical domain, several methods such as clinical examination, tests and imaging are used for accurate diagnosis. Data integration should be used to make an automatic diagnosis system with the ability of using several methods with different data types. The amount of works published on this idea is negligible and most of them have considered a single data type.

5.1.4. Scalability

Data mining algorithms should be scalable in case of huge amount of records or variables. Scalability means retaining accuracy and efficiency of knowledge when dealing with huge data. Popular algorithms such as decision tree cannot handle huge data because of memory restriction. With respect to Figs. 8 and 9, most of papers focus on small data. Additionally, some of papers that focus on scalability have tried to solve the problem of high dimensional data rather than large amount of records.

5.1.5. Medical data mining challenges

Based on Section 3.2.3, there are three data challenges. The general challenge can be solved by data pre-processing. However, collection and medical domain challenge have received less attention. It is suggested that comprehensible standards can overcome some challenges.

5.1.6. Users

Types of users should be taken into account in knowledge extraction. Some papers have tried to address this issue. For i.e. Systems have been made for nurses (Goodwin et al., 2003) or patients (Chi et al., 2008). This issue is considered in the adaptive standard framework in Section 3.2.4.

5.1.7. Medical tasks

Among six medical tasks mentioned in Section 4, researchers have been more attention on diagnosis. Cancer has the biggest share of research in diagnosis, screening and prognosis. Also, heart disease and ICU in monitoring, and heart and scheduling in management, have more works carried out on. Treatment research has a uniform frequency among diseases, except on cancer. Predict outcome of some surgeries such as heart, brain and transplant according to high risk and cost, is essential. Therefore, prognosis should be considered more.

5.1.8. Diseases

Based on WHO fact and figures, there some diseases that have a more mortality rate and incidence known as risky disease. This study shown that three of risky diseases have more attention than other disease: cancer, heart and diabetes. Percentage of papers focusing on these diseases has been depicted in Fig. 19. In heart disease, prevention and early diagnosis should be taken into account. This is because of high mortality and morbidity in patients.

5.1.9. Data pre-processing

The most important step of knowledge extraction is data pre-processing. This step takes most of the project time. Percentage of papers using data pre-processing in literature can be observed in Fig. 20. According to this figure, about half of papers have explicitly mentioned data pre-processing methods. We have described the amount of papers using different data pre-processing approaches in literature in Fig. 21. Feature selection is more common than other data pre-processing techniques. Noise, outlier and missing value are common topics in data pre-processing. Imbalanced data should also be considered more. In the one hand the nature of medical data is imbalanced and, on the other hand, efficiency of algorithms has been degraded when facing this data. Select appropriate data pre-processing techniques with respect to characteristics of data is summarized as a shortcoming of previous researches.

5.1.10. Data modeling

Frequency of published papers in six medical tasks can be shown in Fig. 22. In each task, frequency of data mining approach is given. Classification algorithms appear as the most used in all tasks. Regression based algorithms are known as basic statistical techniques used to do valuable traditional studies on prognosis. Therefore, most works in regression have focused on prognosis.

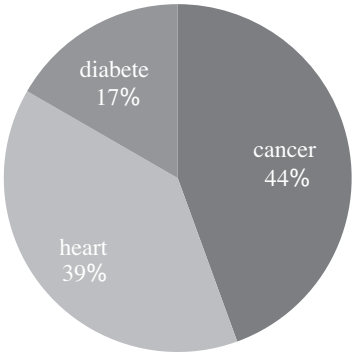

Fig. 19. Percentage of published papers in three risky diseases.

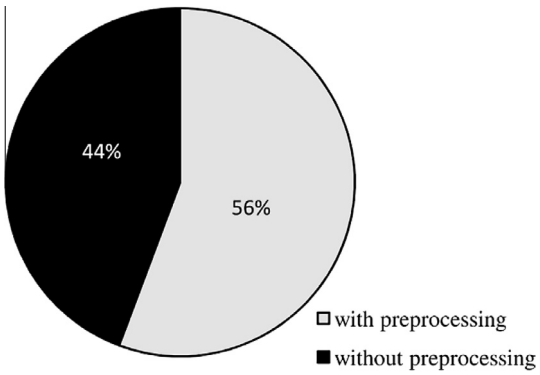

Fig. 20. Percentage of papers on literature that using explicitly data pre-processing.

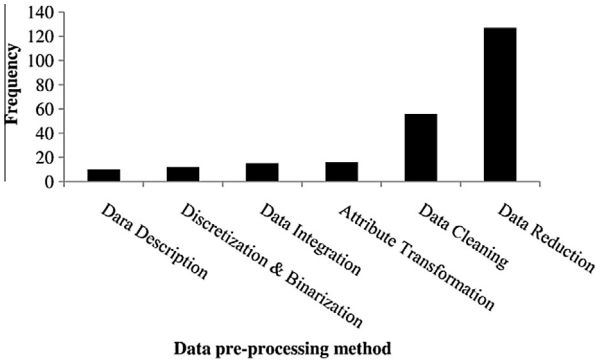

Fig. 21. Amount of usage of different data pre-processing in literature.

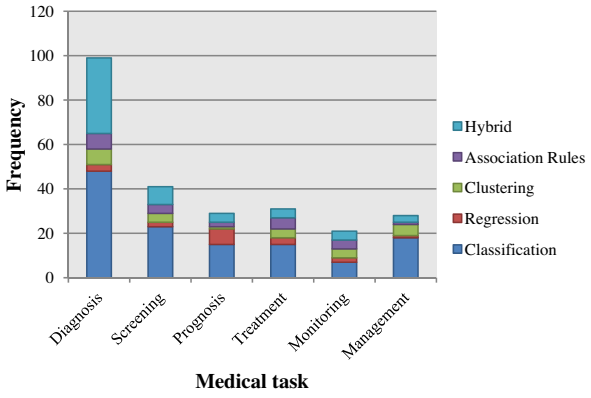

Fig. 22. Frequency of medical tasks and data mining approaches in each of them.

Association rule in treatment is more focused. Data clustering algorithm has been used more in management but it can also be performed in data pre-processing and visualization. Recently, the amount of work published on hybrid approach is substantial for making it possible to combine advantages of several algorithms. The main goal of this approach is increase performance. Combination of fuzzy and evolutionary is a more suitable and popular instance. Fuzzy solves the uncertainly problem and evolutionary can be used as an optimization tool. In Fig. 23, frequency of data mining algorithms in medical domain has been depicted.

5.1.11. Classification

Among the five data mining approaches, classification known as most important due to the high frequency in literature. Interpretability of model is the key factor to select the best algorithm for extracting knowledge. It is important for the expert to understand extracted knowledge. Therefore, decision tree is the most popular method in medical data mining. SVM and artificial neural network are proved efficient but less popular compared with decision tree, due to the incomprehensibility. SVM can achieve more popularity if provide a better way to represent and visualize models. Both major limitations of decision tree in medical data are imbalancement and cost-sensitivity problem. In the first problem, class values are divided into majority and minority groups that tree is biased by majority value. The leaves associated with minority value may also be removed in pruning step. In second problem, different class values have the different cost. These issues should be taken into account in the growing phase and performance evaluation (Kotsiantis, 2013). In order to solve cost-sensitivity problem, we can refer to (Lomax & Vadera, 2013).

5.1.12. Data post-processing

In Fig. 24, frequencies of usage of mentioned methods in performance evaluation are presented. The simplest and most widely used method is accuracy, which has appeared in about 90% of papers. Main limitation of this method is skew data. Weighting approach has been recommended to solve this problem. In Fig. 24, accuracy is not mentioned. Typically, in medical data sets, distribution of each class is not balanced. Also, cost of each class is different. For example, in binary classes, positive value is stated as incidence of disease. In these classes, correct prediction of positive values is more important than negative values. On the other hand, amount of positive values may be less than negatives. If a healthy person is incorrectly diagnosed as a patient, this result can be corrected by re-trial or by experience of corresponding physician. But, if a patient is diagnosed as healthy incorrectly, his disease may progress before correct results are obtained. Sensitivity can evaluate models for negative values and specificity is used to evaluate positive class values. ROC and AUC are also popular. This popularity is because of their interpretability and ability to compare several models.

5.2. Current issue and future trend

We showed that application of data mining in medicine is growing in recent years. Based on Fig. 12, trend of medical data

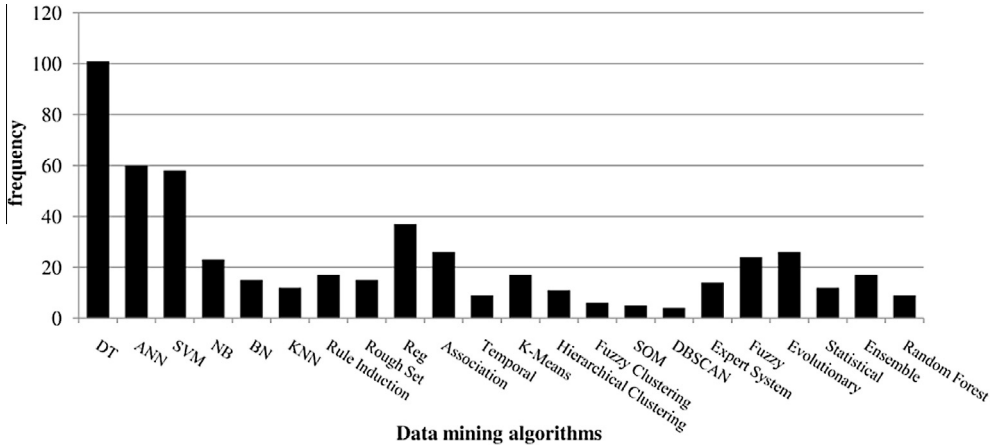

Fig. 23. Frequency of using data mining algorithms in medicine.

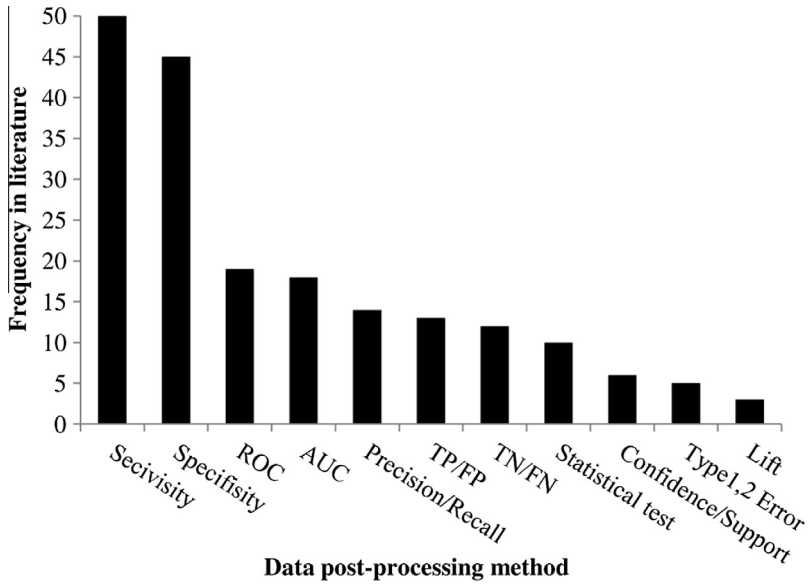

Fig. 24. Frequency of usage of performance evaluation methods in literature.

mining is same as data mining trend. In 2005 and 2010, significant growth in medical data mining publication is observed. Several factors affected this tendency such as success of data mining in other areas, need of new technology in medicine for automatic analyzing, and transformation of data mining from academic into industrial.

Percentage of studies based on medical tasks during the time has been illustrated in Fig. 25. Diagnosis has higher expected percentage in 2002 and 2004 than others. This task is also uniformly distributed over the time. In contrast, monitoring has received little effort. This may be because of overlapping with various fields such as signal processing or being considered in other fields instead of data mining. Statistical methods such as regression are known as the oldest algorithms in medicine for prediction of survivability and occurrence of disease. This has led to more focus on prognosis in 1999–2004. In the middle of time, treatment and screening are growing and prognosis and management are reduced. Treatment is known as the most difficult task for automation because it is associated directly with human life. According to mentioned notes, treatment application has started to increase after usefulness of application of data mining in others tasks has been proved. The reason of increase in screening from 2005 is development of microarray technology to be used as screening tools. Recently, microarray data is playing an important role in early diagnosis and screening of cancer. As a consequence, treatment, diagnosis and screening can be considered as popular tasks in the future.

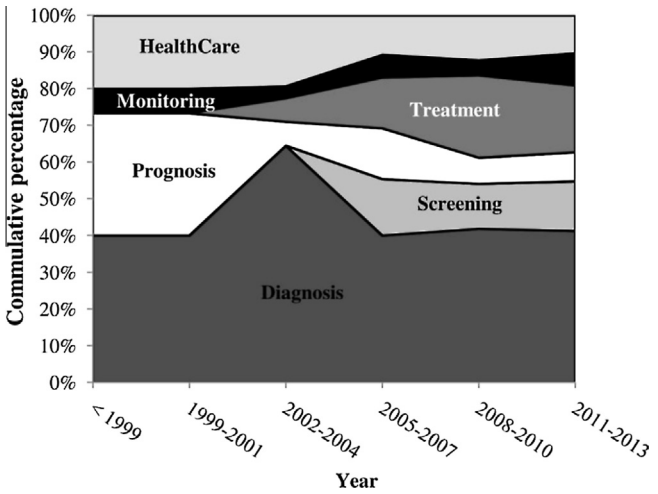

Fig. 25. Area chart of application of data mining in different medical tasks during the time.

Trend of cause of death is investigated in Organization (2007). Although heart disease was the most important cause of death in early years of twentieth century, nowadays cancer has replaced it. Given that the increase in cancer mortality in coming years and popularity of application of data mining in this disease, cancer seems to be an important future focus.

It is better that there is correlation between cancer world statistics and amount of research on various types of cancer. if there is correlation, then we can anticipate future and tendency. Amount of papers published in cancer in literature, and incident and mortality rates in the world (http://globocan.iarc.fr) is mentioned in Fig. 26. By comparing two graphs we realized that there is a weak correlation between two graphs. Breast is more incident and lung has a more mortality in a real world. Exactly, these two cancers also have more attention in literature. But some of them do not follow this rule. For i.e. prostate cancer that has high incident but the frequency in literature is low.

At now, we decide to draw a future path that helps researchers to determine which cancer potentially should be selected for research. Some types of cancer such as prostate, pancreatic and liver cancer can be considered more than past. Prostate cancer has high incident in men and pancreatic and liver cancer has a high mortality per diagnosed patient rate. Patients Follow-up and thereby decreasing occurrence of cancer can also be stated an important subject of research mentioned in a small number of works.

Although the decision tree is known as the most popular algorithm in medical data mining, its usage has decreased in recent years. In contrast, SVM has been more used in recent works. If SVM combined with other algorithms to add ability of knowledge visualization such as (Arun Kumar & Gopal, 2010), usage of SVM in medical data mining may be increased. Artificial neural network has uniform frequency during the time. The most advantage of artificial neural network is its ability to extract complex relations. But artificial neural network is known as a black box algorithm that cannot visualize knowledge. Regression algorithm has been accepted as a reliable method to extract knowledge due to the histor-

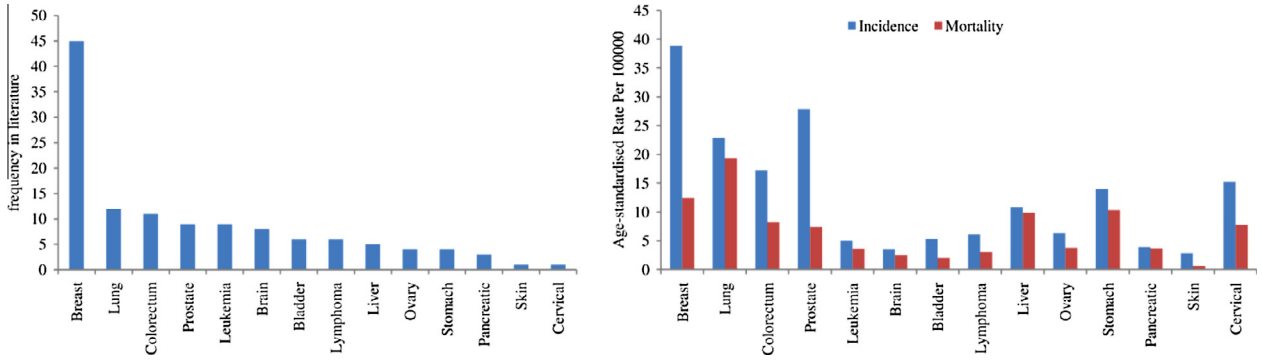

Fig. 26. Amount of published paper for various cancers (left) – incident and mortality rates in the world standardized based on age per 100,000 (right).

ical background in medical domain. In the beginning, regression algorithm was used in prognosis as a statistical tool but it is now used in all medical tasks.

Hybrid approaches try to combine two or more algorithms with the aim of enhancing performance. In literature three approaches have appeared to be combined with data mining algorithms: fuzzy, expert system and evolutionary algorithms. Fuzzy approach has more attention to construct hybrid models. Hence, SVM and fuzzy approach are expected to be popular methods in the future.

6. Concluding remarks

Inadequacy of traditional data when facing challenges such as diversity of data types and large amount of records and variables has led to creation of data mining. Application of data mining is rising in recent years. Application of data mining in medicine also rising due to the two issues: importance of medicine and its effects on the human life, and success of application of data mining in the others fields. There is wide variety of applications that covered most of the medicine field. Recently, data mining can be used in the most of the disease from prevalent to rare. The author's contribution show that different disciplines and views are interacting with each other.

It is important to provide a comprehensive survey. Therefore in this paper, related works concerning knowledge extraction from structural medical data are reviewed. For this purpose 291 papers from wide variety of journals such as data mining and medicine were selected. The main goal of this review is to clarify main issues such as definition, challenges and framework, and make a review based on medical tasks.

After describing the framework that used to provide this review in Section 2, similar papers in the context of review and survey were explained. The contribution of these papers in the medicine and data mining was different. Therefore, writing a review paper that update researches and focus on the knowledge extraction, is necessary.

In Section 3, data mining concepts and three phase process model for data mining is introduced. In the following, definition and goals for medical data mining is elicited based on literature. The main characteristics of medical data and various medical data types are described. In this subsection, we noted that numeric and microarray data is the most frequently used data type in this field. Also with the study on the size of these data, we found that most of the papers used small size of data. An adaptive standard framework was also proposed to cover all data mining activities from problem definition to deployment. The main part of Section 3 is forming by drawing research trend. By this trend we can noticed two tips: leaping the researches in recent years and, similarity of data mining trend and medical data mining trend.

In Section 4, each paper reviews based on six medical tasks: screening, diagnosis, treatment, prognosis, monitoring and management. In each task, five data mining approach are examined: classification, regression, clustering, association rule and hybrid. In Section 5, firstly, a brief summarization of study in the form of twelve issues was explained. Then, current issue and future trend were investigated.

Finally, we can summarize literature around three point:

(1) *Based on the six medical tasks*: diagnosis is the most data mining usage. Frequency of each task was draw during the time. In contrast, monitoring has received little effort. As a consequence, treatment, diagnosis and screening can be considered as popular tasks in the future.
(2) *Based on disease*: three risky diseases according to WHO facts and figures are heart disease, cancer and diabetes that known as high risk health problems. These risky diseases have most focus. Mortality rate of cancer is rising in recent years to make it a global concern. Therefore, more attention has been devoted to cancer.
(3) *Based on data mining*: Data mining process includes three steps. Most works on data pre-processing step have focused on data reduction. Among various algorithms in data modeling step, decision tree is known as the most popular due to its simplicity and interpretability. Recently, more efficient algorithms such as SVM have also become popular. Application of SVM, Fuzzy systems and regression has increased during recent years. Performance evaluation methods in data post-processing step should be selected based on type of extracted knowledge and data.

As for weakness of application of data mining in medicine, lack of standard in the overall of knowledge extraction process from data gathering to knowledge evaluation, can be cited. For example a standard can be imposed for data preparation that unified gathering and integration. The main goal of this field is extracted knowledge that be used in medicine. Then we need a way to put the knowledge into the medicine process. It is not useful to produce knowledge and remain inutile. On the other hand, a system that will be used in medicine should have especial characteristics such as high accuracy and reliability. Although the efficiency of data mining in medicine has been proved but there is not yet confidence to used in medicine. Lack of proven package or tools to use in a real world except in microarray analysis, can be noted as a major reason for this weakness.

The strength of this study can be pointed in three cases: (1) present a comprehensive study to cover most of the medicine and data mining area. (2) Integration of medicine and data mining view to cover all users. (3) Provide definition, goals and challenges to clarify this domain.

In this study, we focused only on a structural data. Therefore as for future works, provide a review on application of text mining in medicine is recommended. Also, Introducing the tools, package and successful implemented programs in this field is proposed.

References

Abbass, H. A. (2002). An evolutionary artificial neural networks approach for breast cancer diagnosis. *Artificial Intelligence in Medicine, 25*, 265–281.

Abbod, M. F., Linkens, D. A., Mahfouf, M., & Dounias, G. (2002). Survey on the use of smart and adaptive engineering systems in medicine. *Artificial Intelligence in Medicine, 26*, 179–209.

Abbod, M. F., von Keyserlingk, D. G., Linkens, D. A., & Mahfouf, M. (2001). Survey of utilisation of fuzzy technology in medicine and healthcare. *Fuzzy Sets and Systems, 120*, 331–349.

Al-Angari, H. M., & Sahakian, A. V. (2012). Automated recognition of obstructive sleep apnea syndrome using support vector machine classifier. *IEEE Transactions on Information Technology in Biomedicine, 16*, 463–468.

Almazyad, A., Ahamad, M., Siddiqui, M., & Almazyad, A. (2010). Effective hypertensive treatment using data mining in Saudi Arabia. *Journal of Clinical Monitoring and Computing, 24*, 391–401.

Alonso, F., Caraça-Valente, J. P., González, A. L., & Montes, C. (2002). Combining expert knowledge and data mining in a medical diagnosis domain. *Expert Systems with Applications, 23*, 367–375.

Alonso, F., Caraça-Valente, J., Martínez, L., & Montes, C. (2003). Discovering similar patterns for characterizing time series in a medical domain. *Knowledge and Information Systems, 5*, 183–200.

Altiparmak, F. F. H., Erdal, S., & Trost, D. C. (2006). Information mining over heterogeneous and high-dimensional time-series data in clinical trials databases. *IEEE Transactions on Information Technology in Biomedicine, 10*(254), 263.

Anagnostou, T., Remzi, M., Lykourinas, M., & Djavan, B. (2003). Artificial neural networks for decision-making in urologic oncology. *European Urology, 43*, 596–603.

Antonelli, D., Baralis, E., Bruno, G., Cerquitelli, T., Chiusano, S., & Mahoto, N. (2013). Analysis of diabetic patients through their examination history. *Expert Systems with Applications, 40*, 4672–4678.

Antonelli, D., Baralis, E., Bruno, G., Chiusano, S., Mahoto, N., & Petrigni, C. (2012). Analysis of diagnostic pathways for colon cancer. *Flexible Services and Manufacturing Journal, 24*, 379–399.

Apiletti, D., Baralis, E., Bruno, G., & Cerquitelli, T. (2009). Real-time analysis of physiological data to support medical applications. *IEEE Transactions on Information Technology in Biomedicine, 13*, 313–321.

Arun Kumar, M., & Gopal, M. (2010). A hybrid SVM based decision tree. *Pattern Recognition, 43*, 3977–3987.

Au, W.-H., Chan, K. C. C., Wong, A. K. C., & Wang, Y. (2005). Attribute clustering for grouping, selection, and classification of gene expression data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2*, 83–101.

Aussem, A., de Morais, S. R., & Corbex, M. (2012). Analysis of nasopharyngeal carcinoma risk factors with Bayesian networks. *Artificial Intelligence in Medicine, 54*, 53–62.

Azar, A., & El-Metwally, S. (2012). Decision tree classifiers for automated medical diagnosis. *Neural Computing and Applications*, 1–17.

Barakat, N. H., Bradley, A. P., & Barakat, M. N. H. (2010). Intelligible support vector machines for diagnosis of diabetes mellitus. *Transactions on Information Technology in Biomedicine, 14*, 1114–1120.

Becker, H. (2001). Computing with words and machine learning in medical diagnostics. *Information Sciences, 134*, 53–69.

Bellazzi, R., Sacchi, L., Caffi, E., de Vincenzi, A., Nai, M., Manicone, F., et al. (2012). Implementation of an automated system for monitoring adherence to hemodialysis treatment: A report of seven years of experience. *International Journal of Medical Informatics, 81*, 320–331.

Bellazzi, R., & Zupan, B. (2008). Predictive data mining in clinical medicine: Current issues and guidelines. *International Journal of Medical Informatics, 77*, 81–97.

Bentham, J., & Hand, D. (2012). Data mining from a patient safety database: the lessons learned. *Data Mining and Knowledge Discovery, 24*, 195–217.

Benton, A., Ungar, L., Hill, S., Hennessy, S., Mao, J., Chung, A., et al. (2011). Identifying potential adverse effects using the web: A new approach to medical hypothesis generation. *Journal of Biomedical Informatics, 44*, 989–996.

Berman, J. J. (2002). Confidentiality issues for medical data miners. *Artificial Intelligence in Medicine, 26*, 25–36.

Bertolazzi, P., Felici, G., Festa, P., & Lancia, G. (2008). Logic classification and feature selection for biomedical data. *Computers & Mathematics with Applications, 55*, 889–899.

Bilge, U., Bozkurt, S., & Durmaz, S. (2013). Application of data mining techniques for detecting asymptomatic carotid artery stenosis. *Computers & Electrical Engineering, 39*, 1499–1505.

Bojarczuk, C. C., Lopes, H. S., & Freitas, A. A. (2000). Genetic programming for knowledge discovery in chest-pain diagnosis. *IEEE Engineering in Medicine and Biology Magazine, 19*, 38–44.

Bontempi, G. (2007). A blocking strategy to improve gene selection for classification of gene expression data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 4*, 293–300.

Bourien, J., Bellanger, J. J., Bartolomei, F., Chauvel, P., & Wendling, F. (2004). Mining reproducible activation patterns in epileptic intracerebral EEG signals: Application to interictal activity. *IEEE Transactions on Biomedical Engineering, 51*, 304–315.

Breault, J. L., Goodall, C. R., & Fos, P. J. (2002). Data mining a diabetic data warehouse. *Artificial Intelligence in Medicine, 26*, 37–54.

Briones, N., & Dinu, V. (2012). Data mining of high density genomic variant data for prediction of Alzheimer's disease risk. *BMC Medical Genetics, 13*, 1–12.

Çakır, A., & Demirel, B. (2011). A software tool for determination of breast cancer treatment methods using data mining approach. *Journal of Medical Systems, 35*, 1503–1511.

Cao, X., Maloney, K., & Brusic, V. (2008). Data mining of cancer vaccine trials: a bird's-eye view. *Immunome Research, 4*, 1–11.

Chae, Y. M., Kim, H. S., Tark, K. C., Park, H. J., & Ho, S. H. (2003). Analysis of healthcare quality indicator using data mining and decision support system. *Expert Systems with Applications, 24*, 167–172.

Chang, C.-L., & Chen, C.-H. (2009). Applying decision tree and neural network to increase quality of dermatologic diagnosis. *Expert Systems with Applications, 36*, 4035–4041.

Chang, R.-I., Chu, C.-C., Wu, Y.-Y., & Chen, Y.-L. (2010). Gene clustering by using query-based self-organizing maps. *Expert Systems with Applications, 37*, 6689–6694.

Chang, C.-D., Wang, C.-C., & Jiang, B. C. (2011). Using data mining techniques for multi-diseases prediction modeling of hypertension and hyperlipidemia by common risk factors. *Expert Systems with Applications, 38*, 5507–5513.

Chang, W.-W., Yeh, W.-C., & Huang, P.-C. (2010). A hybrid immune-estimation distribution of algorithm for mining thyroid gland data. *Expert Systems with Applications, 37*, 2066–2071.

Chaochang, C., Kuang-Hung, H., Hsu, P. L., Chi, I. H., Po-Chi, L., Wen-Ko, C., et al. (2007). Mining three-dimensional anthropometric body surface scanning data for hypertension detection. *IEEE Transactions on Information Technology in Biomedicine, 11*, 264–273.

Chaovalitwongse, W. A., Pottenger, R. S., Shouyi, W., Ya-Ju, F., & Iasemidis, L. D. (2011). Pattern- and network-based classification techniques for multichannel medical data signals to improve brain diagnosis. *IEEE Transactions on Systems, Man, and Cybernetics Part A: Systems and Humans, 41*, 977–988.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0 step-by-step data mining guide. In *The CRISP-DM consortium*.

Chaves, R., Ramírez, J., Górriz, J. M., & Puntonet, C. G. (2012). Association rule-based feature selection method for Alzheimer's disease diagnosis. *Expert Systems with Applications, 39*, 11766–11774.

Chen, H.-Y., Chuang, C.-H., Yang, Y.-J., & Wu, T.-P. (2011). Exploring the risk factors of preterm birth using data mining. *Expert Systems with Applications, 38*, 5384–5387.

Chen, H.-Y., Hou, T.-W., & Chuang, C.-H. (2010). Applying data mining to explore the risk factors of parenting stress. *Expert Systems with Applications, 37*, 598–601.

Chen, T.-C., & Hsu, T.-C. (2006). A GAs based approach for mining breast cancer pattern. *Expert Systems with Applications, 30*, 674–681.

Chen, H.-L., Huang, C.-C., Yu, X.-G., Xu, X., Sun, X., Wang, G., et al. (2013). An efficient diagnosis system for detection of Parkinson's disease using fuzzy *k*-nearest neighbor approach. *Expert Systems with Applications, 40*, 263–271.

Chen, Y.-W., Larbani, M., Hsieh, C.-Y., & Chen, C.-W. (2009). Introduction of affinity set and its application in data-mining example of delayed diagnosis. *Expert Systems with Applications, 36*, 10883–10889.

Chen, K.-Y., Lee, Y.-C. G., Lai, J.-M., Chang, Y.-L., Lee, Y.-C., Yu, C.-J., et al. (2007). Identification of trophinin as an enhancer for cell invasion and a prognostic factor for early stage lung cancer. *European Journal of Cancer, 43*, 782–790.

Chen, L., McKenna, T. M., Reisner, A. T., Gribok, A., & Reifman, J. (2008). Decision tool for the early diagnosis of trauma patient hypovolemia. *Journal of Biomedical Informatics, 41*, 469–478.

Cheng, C.-H. (2012). Discovering knowledge of medical quality in total hip arthroplasty (THA). *Archives of Gerontology and Geriatrics, 55*, 323–330.

Chi, C.-L., Street, W. N., & Ward, M. M. (2008). Building a hospital referral expert system with a prediction and optimization-based decision support system algorithm. *Journal of Biomedical Informatics, 41*, 371–386.

Chittaro, L., Combi, C., & Trapasso, G. (2003). Data mining on temporal data: a visual approach and its clinical application to hemodialysis. *Journal of Visual Languages & Computing, 14*, 591–620.

Choi, W.-J., & Choi, T.-S. (2012). Genetic programming-based feature transform and classification for the automatic detection of pulmonary nodules on computed tomography images. *Information Sciences, 212*, 57–78.

Chou, S.-M., Lee, T.-S., Shao, Y. E., & Chen, I. F. (2004). Mining the breast cancer pattern using artificial neural networks and multivariate adaptive regression splines. *Expert Systems with Applications, 27*, 133–142.

Chuang, C.-L. (2011). Case-based reasoning support for liver disease diagnosis. *Artificial Intelligence in Medicine, 53*, 15–23.

Cilla, M., Martinez, J., Pena, E., Marti, x., & nez, M. A. (2012). Machine learning techniques as a helpful tool toward determination of plaque vulnerability. *IEEE Transactions on Biomedical Engineering, 59*, 1155–1161.

Çınar, M., Engin, M., Engin, E. Z., & Ziya Ateşçi, Y. (2009). Early prostate cancer diagnosis by using artificial neural networks and support vector machines. *Expert Systems with Applications, 36*, 6357–6361.

Cios, K. J., & William Moore, G. (2002). Uniqueness of medical data mining. *Artificial Intelligence in Medicine, 26*, 1–24.

Combes, C., Meskens, N., Rivat, C., & Vandamme, J. P. (2008). Using a KDD process to forecast the duration of surgery. *International Journal of Production Economics, 112*, 279–293.

Cox, S., Currell, A., Harriss, L., Barger, B., Cameron, P., & Smith, K. (2012). Evaluation of the Victorian state adult pre-hospital trauma triage criteria. *Injury, 43*, 573–581.

Cruz-Ramírez, N., Acosta-Mesa, H.-G., Carrillo-Calvet, H., & Barrientos-Martínez, R.-E. (2009). Discovering interobserver variability in the cytodiagnosis of breast cancer using decision trees and Bayesian networks. *Applied Soft Computing, 9*, 1331–1342.

Czabanski, R., Jezewski, J., Matonia, A., & Jezewski, M. (2012). Computerized analysis of fetal heart rate signals as the predictor of neonatal acidemia. *Expert Systems with Applications, 39*, 11846–11860.

Dai, J., & Xu, Q. (2013). Attribute selection based on information gain ratio in fuzzy rough set theory with application to tumor classification. *Applied Soft Computing, 13*, 211–221.

De Falco, I. (2013). Differential evolution for automatic rule extraction from medical databases. *Applied Soft Computing, 13*, 1265–1283.

Delen, D., Fuller, C., McCann, C., & Ray, D. (2009). Analysis of healthcare coverage: A data mining approach. *Expert Systems with Applications, 36*, 995–1003.

Delen, D., Oztekin, A., & Kong, Z. (2010). A machine learning-based approach to prognostic analysis of thoracic transplantations. *Artificial Intelligence in Medicine, 49*, 33–42.

Delen, D., Oztekin, A., & Tomak, L. (2012). An analytic approach to better understanding and management of coronary surgeries. *Decision Support Systems, 52*, 698–705.

Delen, D., Walker, G., & Kadam, A. (2005). Predicting breast cancer survivability: a comparison of three data mining methods. *Artificial Intelligence in Medicine, 34*, 113–127.

Demšar, J., Zupan, B., Aoki, N., Wall, M. J., Granchi, T. H., & Robert Beck, J. (2001). Feature mining and predictive model construction from severe trauma patient's data. *International Journal of Medical Informatics, 63*, 41–50.

Devi, S. P., Rao, K. S., & Sangeetha, S. S. (2012). Prediction of surgery times and scheduling of operation theaters in optholmology department. *Journal of Medical Systems, 36*, 415–430.

Diederich, J., Al-Ajmi, A., & Yellowlees, P. (2007). Ex-ray: Data mining and mental health. *Applied Soft Computing, 7*, 923–928.

Engreitz, J. M., Daigle, B. J., Jr, Marshall, J. J., & Altman, R. B. (2010). Independent component analysis: Mining microarray data for fundamental human gene expression modules. *Journal of Biomedical Informatics, 43*, 932–944.

Erxin, S., Liang, Y., Xinsheng, F., Yuping, T., & Jinao, D. (2010). Discovery of association rules between TCM properties in drug pairs by association mining between datasets and probability tests. *World Science and Technology, 12*, 377–382.

Exarchos, T. P., Papaloukas, C., Fotiadis, D. I., & Michalis, L. K. (2006). An association rule mining-based methodology for automated detection of ischemic ECG beats. *IEEE Transactions on Biomedical Engineering, 53*, 1531–1540.

Fakih, S. J., & Das, T. K. (2006). LEAD: A methodology for learning efficient approaches to medical diagnosis. *Transactions on Information Technology in Biomedicine, 10*, 220–228.

Fan, C.-Y., Chang, P.-C., Lin, J.-J., & Hsieh, J. C. (2011). A hybrid model combining case-based reasoning and fuzzy decision tree for medical data classification. *Applied Soft Computing, 11*, 632–644.

Farion, K., Michalowski, W., Wilk, S., O'Sullivan, D., & Matwin, S. (2010). A tree-based decision model to support prediction of the severity of asthma exacerbations in children. *Journal of Medical Systems, 34*, 551–562.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine, 17*, 37–54.

Ferreira, D., Oliveira, A., & Freitas, A. (2012). Applying data mining techniques to improve diagnosis in neonatal jaundice. *BMC Medical Informatics and Decision Making, 12*, 1–6.

Filipovic, N., Ivanovic, M., Krstajic, D., & Kojic, M. (2011). Hemodynamic flow modeling through an abdominal aorta aneurysm using data mining tools. *IEEE Transactions on Information Technology in Biomedicine, 15*, 189–194.

Frize, M., Ennett, C. M., Stevenson, M., & Trigg, H. C. E. (2001). Clinical decision support systems for intensive care units: using artificial neural networks. *Medical Engineering & Physics, 23*, 217–225.

Froelich, W., Papageorgiou, E. I., Samarinas, M., & Skriapas, K. (2012). Application of evolutionary fuzzy cognitive maps to the long-term prediction of prostate cancer. *Applied Soft Computing, 12*, 3810–3817.

Ge, G., & Wong, G. (2008). Classification of premalignant pancreatic cancer mass-spectrometry data using decision tree ensembles. *BMC Bioinformatics, 9*, 1–12.

Ghazavi, S. N., & Liao, T. W. (2008). Medical data mining by fuzzy modeling with selected features. *Artificial Intelligence in Medicine, 43*, 195–206.

Goodwin, L., VanDyne, M., Lin, S., & Talbert, S. (2003). Data mining issues and opportunities for building nursing knowledge. *Journal of Biomedical Informatics, 36*, 379–388.

Gregori, D., Petrinco, M., Bo, S., Rosato, R., Pagano, E., Berchialla, P., et al. (2011). Using data mining techniques in monitoring diabetes care. The simpler the better? *Journal of Medical Systems, 35*, 277–281.

Grosan, C., Abraham, A., & Tigan, S. (2008). Multicriteria programming in medical diagnosis and treatments. *Applied Soft Computing, 8*, 1407–1417.

Gueguin, M., Roux, E., Hernandez, A. I., Porce, F., Mabo, P., Graindorge, L., et al. (2008). Exploring time series retrieved from cardiac implantable devices for optimizing patient follow-up. *IEEE Transactions on Biomedical Engineering, 55*, 2343–2352.

Guenther, T., Mueller, I., Preuss, M., Kruse, R., & Sabel, B. A. (2009). A treatment outcome prediction model of visual field recovery using self-organizing maps. *IEEE Transactions on Biomedical Engineering, 56*, 572–581.

Güiza, F., Eyck, J., & Meyfroidt, G. (2012). Predictive data mining on monitoring data from the intensive care unit. *Journal of Clinical Monitoring and Computing*, 1–5.

Gülçin Yıldırım, E., Karahoca, A., & Uçar, T. (2011). Dosage planning for diabetes patients using data mining methods. *Procedia Computer Science, 3*, 1374–1380.

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

Hassanien, A. E., & Kim, T.-H. (2012). Breast cancer MRI diagnosis approach using support vector machine and pulse coupled neural networks. *Journal of Applied Logic, 10*, 277–284.

Hautemanière, A., Florentin, A., Hartemann, P., & Hunter, P. R. (2011). Identifying possible deaths associated with nosocomial infection in a hospital by data mining. *American Journal of Infection Control, 39*, 118–122.

Hijazi, M. H. A., Coenen, F., & Zheng, Y. (2012). Data mining techniques for the screening of age-related macular degeneration. *Knowledge-Based Systems, 29*, 83–92.

Hirano, S., Sun, X., & Tsumoto, S. (2004). Comparison of clustering methods for clinical databases. *Information Sciences, 159*, 155–165.

Ho, T.-B., Nguyen, C.-H., Kawasaki, S., Le, S.-Q., & Takabayashi, K. (2007). Exploiting temporal relations in mining hepatitis data. *New Generation Computing, 25*, 247–262.

Hristovski, D., Peterlin, B., Mitchell, J. A., & Humphrey, S. M. (2005). Using literature-based discovery to identify disease candidate genes. *International Journal of Medical Informatics, 74*, 289–298.

Hsieh, N.-C., Hung, L.-P., Shih, C.-C., Keh, H.-C., & Chan, C.-H. (2012). Intelligent postoperative morbidity prediction of heart disease using artificial intelligence techniques. *Journal of Medical Systems, 36*, 1809–1820.

Hsu, K.-H., Chiu, C., Chiu, N.-H., Lee, P.-C., Chiu, W.-K., Liu, T.-H., et al. (2011). A case-based classifier for hypertension detection. *Knowledge-Based Systems, 24*, 33–39.

Hsu, C.-C., & Ho, C.-S. (2004). A new hybrid case-based architecture for medical diagnosis. *Information Sciences, 166*, 231–247.

Hu, X., Sapo, M., Nenov, V., Barry, T., Kim, S., Do, D. H., et al. (2012). Predictive combinations of monitor alarms preceding in-hospital code blue events. *Journal of Biomedical Informatics, 45*, 913–921.

Huang, M.-J., Chen, M.-Y., & Lee, S.-C. (2007). Integrating data mining with case-based reasoning for chronic diseases prognosis and diagnosis. *Expert Systems with Applications, 32*, 856–867.

Huang, Y., McCullagh, P., Black, N., & Harper, R. (2007). Feature selection and classification model construction on type 2 diabetic patients' data. *Artificial Intelligence in Medicine, 41*, 251–262.

Huang, S. H., Wulsin, L. R., Li, H., & Guo, J. (2009). Dimensionality reduction for knowledge discovery in medical claims database: Application to antidepressant medication utilization study. *Computer Methods and Programs in Biomedicine, 93*, 115–123.

Huang, G., Zhang, Y., Cao, J., Steyn, M., & Taraporewalla, K. (2013). Online mining abnormal period patterns from multiple medical sensor data streams. *World Wide Web*, 1–19.

Huidong, J., Jie, C., Hongxing, H., Williams, G. J., Kelman, C., & O''Keefe, C. M. (2008). Mining unexpected temporal associations: Applications in detecting adverse drug reactions. *IEEE Transactions on Information Technology in Biomedicine, 12*, 488–500.

Illán, I. A., Górriz, J. M., López, M. M., Ramírez, J., Salas-Gonzalez, D., Segovia, F., et al. (2011). Computer aided diagnosis of Alzheimer's disease using component based SVM. *Applied Soft Computing, 11*, 2376–2382.

Imamura, T., Matsumoto, S., Kanagawa, Y., Tajima, B., Matsuya, S., Furue, M., et al. (2007). A technique for identifying three diagnostic findings using association analysis. *Medical & Biological Engineering & Computing, 45*, 51–59.

Imberman, S. P., Domanski, B., & Thompson, H. W. (2002). Using dependency/association rules to find indications for computed tomography in a head trauma dataset. *Artificial Intelligence in Medicine, 26*, 55–68.

Ion Titapiccolo, J., Ferrario, M., Cerutti, S., Barbieri, C., Mari, F., Gatti, E., et al. (2013). Artificial intelligence models to stratify cardiovascular risk in incident hemodialysis patients. *Expert Systems with Applications, 40*, 4679–4686.

Isken, M., & Rajagopalan, B. (2002). Data mining to support simulation modeling of patient flow in hospitals. *Journal of Medical Systems, 26*, 179–197.

Itchhaporia, D., Snow, P. B., Almassy, R. J., & Oetgen, W. J. (1996). Artificial neural networks: Current status in cardiovascular medicine. *Journal of the American College of Cardiology, 28*, 515–521.

Jannin, P., & Morandi, X. (2007). Surgical models for computer-assisted neurosurgery. *NeuroImage, 37*, 783–791.

Jen, C.-H., Wang, C.-C., Jiang, B. C., Chu, Y.-H., & Chen, M.-S. (2012). Application of classification techniques on development an early-warning system for chronic illnesses. *Expert Systems with Applications, 39*, 8852–8858.

Jiang, D., Tang, C., & Zhang, A. (2004). Cluster analysis for gene expression data: A survey. *IEEE Transactions on Knowledge and Data Engineering, 16*, 1370–1386.

Jin, B., Tang, Y. C., & Zhang, Y.-Q. (2007). Support vector machines with genetic fuzzy feature transformation for biomedical data classification. *Information Sciences, 177*, 476–489.

Jinyan, L., & Qiang, Y. (2007). Strong compound-risk factors: efficient discovery through emerging patterns and contrast sets. *IEEE Transactions on Information Technology in Biomedicine, 11*, 544–552.

Jonsdottir, T., Hvannberg, E. T., Sigurdsson, H., & Sigurdsson, S. (2008). The feasibility of constructing a predictive outcome model for breast cancer using the tools of data mining. *Expert Systems with Applications, 34*, 108–118.

Jourdan, L., Dhaenens, C., Talbi, E. G., & Gallina, S. (2002). A data mining approach to discover genetic and environmental factors involved in multifactorial diseases. *Knowledge-Based Systems, 15*, 235–242.

Kahramanli, H., & Allahverdi, N. (2008). Design of a hybrid system for the diabetes and heart diseases. *Expert Systems with Applications, 35*, 82–89.

Kantardzic, M., Djulbegovic, B., & Hamdan, H. (2002). A data-mining approach to improving polycythemia vera diagnosis. *Computers & Industrial Engineering, 43*, 765–773.

Karabulut, E., & İbrikçi, T. (2012). Effective diagnosis of coronary artery disease using the rotation forest ensemble method. *Journal of Medical Systems, 36*, 3011–3018.

Karaolis, M. A., Moutiris, J. A., Hadjipanayi, D., & Pattichis, C. S. (2010). Assessment of the risk factors of coronary heart events based on data mining with decision trees. *IEEE Transactions on Information Technology in Biomedicine, 14*, 559–566.

Kirkos, E., Spathis, C., & Manolopoulos, Y. (2007). Data mining techniques for the detection of fraudulent financial statements. *Expert Systems with Applications, 32*, 995–1003.

Klema, J., Novakova, L., Karel, F., Stepankova, O., & Zelezny, F. (2008). Sequential data mining: A comparative case study in development of atherosclerosis risk factors. *Transactions on Systems, Man, and Cybernetics Part C, 38*, 3–15.

Klement, W., Wilk, S., Michalowski, W., Farion, K. J., Osmond, M. H., & Verter, V. (2012). Predicting the need for CT imaging in children with minor head injury using an ensemble of naive bayes classifiers. *Artificial Intelligence in Medicine, 54*, 163–170.

Kohli, R., Krishnamurti, R., & Jedidi, K. (2006). Subset-conjunctive rules for breast cancer diagnosis. *Discrete Applied Mathematics, 154*, 1100–1112.

Kotsiantis, S. B. (2013). Decision trees: A recent overview. *Artificial Intelligence Review, 39*, 261–283.

Koyuncugil, A., & Ozgulbas, N. (2010). Donor research and matching system based on data mining in organ transplantation. *Journal of Medical Systems, 34*, 251–259.

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

Kusiak, A., Caldarone, C. A., Kelleher, M. D., Lamb, F. S., Persoon, T. J., & Burns, A. (2006). Hypoplastic left heart syndrome: Knowledge discovery with a data mining approach. *Computers in Biology and Medicine, 36*, 21–40.

Kusiak, A., Dixon, B., & Shah, S. (2005). Predicting survival time for kidney dialysis patients: A data mining approach. *Computers in Biology and Medicine, 35*, 311–327.

Kusiak, A., Law, I. H., & MacDonald, D. I. (2001). The G-algorithm for extraction of robust decision rules – Children's postoperative intra-atrial arrhythmia case study. *Transactions on Information Technology in Biomedicine, 5*, 225–235.

Kwiatkowska, M., Atkins, M. S., Ayas, N. T., & Ryan, C. F. (2007). Knowledge-based data analysis: First step toward the creation of clinical prediction rules using a new typicality measure. *IEEE Transactions on Information Technology in Biomedicine, 11*, 651–660.

Kwong-Sak, L., Kin Hong, L., Jin-Feng, W., Ng, E. Y. T., Chan, H. L. Y., et al. (2011). Data mining on DNA sequences of hepatitis B virus. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 8*, 428–440.

Lahsasna, A., Ainon, R., Zainuddin, R., & Bulgiba, A. (2012). Design of a fuzzy-based decision support system for coronary heart disease diagnosis. *Journal of Medical Systems, 36*, 3293–3306.

Lamma, E., Mello, P., Nanetti, A., Riguzzi, F., Storari, S., & Valastro, G. (2006). Artificial intelligence techniques for monitoring dangerous infections. *Transactions on Information Technology in Biomedicine, 10*, 143–155.

Sacha, J. P., Cios, K. J., & Goodenday, L. S. (2000). Issues in automating cardiac SPECT diagnosis. *IEEE Engineering in Medicine and Biology, 19*, 78–88.

Lavrač, N. (1999). Selected techniques for data mining in medicine. *Artificial Intelligence in Medicine, 16*, 3–23.

Lavrač, N., Bohanec, M., Pur, A., Cestnik, B., Debeljak, M., & Kobler, A. (2007). Data mining and visualization for decision support and modeling of public health-care resources. *Journal of Biomedical Informatics, 40*, 438–447.

Laxminarayan, P. A. S. A., Ruiz, C., & Moonis, M. (2006). Mining statistically significant associations for exploratory analysis of human sleep data. *IEEE Transactions on Information Technology in Biomedicine, 10*, 440–450.

Lee, C.-P., & Leu, Y. (2011). A novel hybrid feature selection method for microarray data analysis. *Applied Soft Computing, 11*, 208–213.

Lee, D., Ryu, K., Bashir, M., Bae, J.-W., & Ryu, K. (2013). Discovering medical knowledge using association rule mining in young adults with acute myocardial infarction. *Journal of Medical Systems, 37*, 1–10.

Lee, G., Rodriguez, C., & Madabhushi, A. (2008). Investigating the efficacy of nonlinear dimensionality reduction schemes in classifying gene and protein expression studies. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 5*, 368–384.

Lee, I.-H., Lushington, G., & Visvanathan, M. (2011). A filter-based feature selection approach for identifying potential biomarkers for lung cancer. *Journal of Clinical Bioinformatics, 1*, 1–8.

Lee, Y.-C., Lee, W.-J., Lee, T.-S., Lin, Y.-C., Wang, W., Liew, P.-L., et al. (2007). Prediction of successful weight reduction after bariatric surgery by data mining technologies. *Obesity Surgery, 17*, 1235–1241.

Lee, Y.-C., Lee, W.-J., & Liew, P.-L. (2013). Predictors of remission of type 2 diabetes mellitus in obese patients after gastrointestinal surgery. *Obesity Research & Clinical Practice*.

Lee, Y. J., Mangasarian, O. L., & Wolberg, W. H. (2003). Survival-time classification of breast cancer patients. *Computational Optimization and Applications, 25*, 151–166.

Li, D.-C., Fang, Y.-H., Lai, Y.-Y., & Hu, S. C. (2009). Utilization of virtual samples to facilitate cancer identification for DNA microarray data in the early stages of an investigation. *Information Sciences, 179*, 2740–2753.

Li, J., Fu, A. W.-C., & Fahey, P. (2009). Efficient discovery of risk patterns in medical data. *Artificial Intelligence in Medicine, 45*, 77–89.

Li, K., Yang, M., Sablok, G., Fan, J., & Zhou, F. (2013). Screening features to improve the class prediction of acute myeloid leukemia and myelodysplastic syndrome. *Gene, 512*, 348–354.

Li, L., Tang, H., Wu, Z., Gong, J., Gruidl, M., Zou, J., et al. (2004). Data mining techniques for cancer detection using serum proteomic profiling. *Artificial Intelligence in Medicine, 32*, 71–83.

Liao, H.-C., & Tsai, J.-H. (2007). Data mining for DNA viruses with breast cancer, fibroadenoma, and normal mammary tissue. *Applied Mathematics and Computation, 188*, 989–1000.

Liao, S.-H., Chu, P.-H., & Hsiao, P.-Y. (2012). Data mining techniques and applications – A decade review from 2000 to 2011. *Expert Systems with Applications, 39*, 11303–11311.

Lin, R.-H. (2009). An intelligent model for liver disease diagnosis. *Artificial Intelligence in Medicine, 47*, 53–62.

Lin, W.-T., Wang, S.-T., Chiang, T.-C., Shi, Y.-X., Chen, W.-Y., & Chen, H.-M. (2010). Abnormal diagnosis of emergency department triage explored with data mining technology: An emergency department at a medical center in Taiwan taken as an example. *Expert Systems with Applications, 37*, 2733–2741.

Lin, W. T., Wu, Y. C., Zheng, J. S., & Chen, M. Y. (2011). Analysis by data mining in the emergency medicine triage database at a Taiwanese regional hospital. *Expert Systems with Applications, 38*, 11078–11084.

Lisboa, P. J. G. (2002). A review of evidence of health benefit from artificial neural networks in medical intervention. *Neural networks, 15*, 11–39.

Liu, G.-P., Li, G.-Z., Wang, Y.-L., & Wang, Y.-Q. (2010). Modelling of inquiry diagnosis for coronary heart disease in traditional Chinese medicine by using multi-label learning. *BMC Complementary and Alternative Medicine, 10*, 1–12.

Lomax, S., & Vadera, S. (2013). A survey of cost-sensitive decision tree induction algorithms. *ACM Computing Surveys, 45*, 1–35.

López-Vallverdú, J. A., Riaño, D., & Bohada, J. A. (2012). Improving medical decision trees by combining relevant health-care criteria. *Expert Systems with Applications, 39*, 11782–11791.

Luk, J. M., Lam, B. Y., Lee, N. P. Y., Ho, D. W., Sham, P. C., Chen, L., et al. (2007). Artificial neural networks and decision tree model analysis of liver cancer proteomes. *Biochemical and Biophysical Research Communications, 361*, 68–73.

Luo, S.-T., & Cheng, B.-W. (2012). Diagnosing breast masses in digital mammography using feature selection and ensemble methods. *Journal of Medical Systems, 36*, 569–577.

Madigan, E., & Curet, O. (2006). A data mining approach in home healthcare: Outcomes and service use. *BMC Health Services Research, 6*, 1–10.

Magoulas, G. D., Plagianakos, V. P., & Vrahatis, M. N. (2004). Neural network-based colonoscopic diagnosis using on-line learning and differential evolution. *Applied Soft Computing, 4*, 369–379.

Mahfouf, M., Abbod, M. F., & Linkens, D. A. (2001). A survey of fuzzy logic monitoring and control utilisation in medicine. *Artificial Intelligence in Medicine, 21*, 27–42.

Mandal, I., & Sairam, N. (2012). Accurate prediction of coronary artery disease using reliable diagnosis system. *Journal of Medical Systems, 36*, 3353–3373.

Mansingh, G., Osei-Bryson, K.-M., & Reichgelt, H. (2011). Using ontologies to facilitate post-processing of association rules by domain experts. *Information Sciences, 181*, 419–434.

Marcano-Cedeño, A., Chausa, P., García, A., Cáceres, C., Tormos, J. M., & Gómez, E. J. (2013). Data mining applied to the cognitive rehabilitation of patients with acquired brain injury. *Expert Systems with Applications, 40*, 1054–1060.

Martis, R. J., Acharya, U. R., Mandana, K. M., Ray, A. K., & Chakraborty, C. (2012). Application of principal component analysis to ECG signals for automated diagnosis of cardiac health. *Expert Systems with Applications, 39*, 11792–11800.

Mazzocco, T., & Hussain, A. (2012). Novel logistic regression models to aid the diagnosis of dementia. *Expert Systems with Applications, 39*, 3356–3361.

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*, 93–99.

Michalowski, W., Rubin, S., Slowinski, R., & Wilk, S. (2003). Mobile clinical support system for pediatric emergencies. *Decision Support Systems, 36*, 161–176.

Mohanty, A., Senapati, M., & Lenka, S. (2012). An improved data mining technique for classification and detection of breast cancer from mammograms. *Neural Computing and Applications*, 1–8.

Mookiah, M. R. K., Rajendra Acharya, U., Lim, C. M., Petznick, A., & Suri, J. S. (2012). Data mining technique for automated diagnosis of glaucoma using higher order spectra and wavelet energy features. *Knowledge-Based Systems, 33*, 73–82.

Mueller, J., von Eggeling, F., Driesch, D., Schubert, J., Melle, C., & Junker, K. (2005). ProteinChip technology reveals distinctive protein expression profiles in the urine of bladder cancer patients. *European Urology, 47*, 885–894.

Mullins, I. M., Siadaty, M. S., Lyman, J., Scully, K., Garrett, C. T., Greg Miller, W., et al. (2006). Data mining and clinical data repositories: Insights from a 667,000 patient data set. *Computers in Biology and Medicine, 36*, 1351–1377.

Muthukaruppan, S., & Er, M. J. (2012). A hybrid particle swarm optimization based fuzzy expert system for the diagnosis of coronary artery disease. *Expert Systems with Applications, 39*, 11657–11665.
Nahar, J., Imam, T., Tickle, K. S., & Chen, Y.-P. P. (2013a). Association rule mining to detect factors which contribute to heart disease in males and females. *Expert Systems with Applications, 40*, 1086–1093.
Nahar, J., Imam, T., Tickle, K. S., & Chen, Y.-P. P. (2013b). Computational intelligence for heart disease diagnosis: A medical knowledge driven approach. *Expert Systems with Applications, 40*, 96–104.
Nahar, J., Imam, T., Tickle, K. S., Shawkat Ali, A. B. M., & Chen, Y.-P. P. (2012). Computational intelligence for microarray data and biomedical image analysis for the early diagnosis of breast cancer. *Expert Systems with Applications, 39*, 12371–12377.
Nahar, J., Tickle, K., Ali, A. B. M. S., & Chen, Y.-P. (2011). Significant cancer prevention factor extraction: An association rule discovery approach. *Journal of Medical Systems, 35*, 353–367.
Nakayama, N., Oketani, M., Kawamura, Y., Inao, M., Nagoshi, S., Fujiwara, K., et al. (2012). Algorithm to determine the outcome of patients with acute liver failure: A data-mining analysis using decision trees. *Journal of Gastroenterology, 47*, 664–677.
Ngai, E. W. T., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems, 50*, 559–569.
Ngai, E. W. T., Xiu, L., & Chau, D. C. K. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications, 36*, 2592–2602.
Niki Kunene, K., & Roland Weistroffer, H. (2008). An approach for predicting and describing patient outcome using multicriteria decision analysis and decision rules. *European Journal of Operational Research, 185*, 984–997.
Organization, W. H. (2007). World health statistics 2007: WORLD HEALTH ORGN.
Orozova-Bekkevold, I., Jensen, H., Stensballe, L., & Olsen, J. (2007). Maternal vaccination and preterm birth: Using data mining as a screening tool. *Pharmacy World & Science, 29*, 205–212.
Ozcift, A. (2012). Enhanced cancer recognition system based on random forests feature elimination algorithm. *Journal of Medical Systems, 36*, 2577–2585.
Oztekin, A., Delen, D., & Kong, Z. (2009). Predicting the graft survival for heart–lung transplantation patients: An integrated data mining methodology. *International Journal of Medical Informatics, 78*, e84–e96.
Padilla, P., Lopez, M., Gorriz, J. M., Ramirez, J., Salas-Gonzalez, D., & Alvarez, I. (2012). NMF-SVM based CAD tool applied to functional brain images for the diagnosis of Alzheimer's disease. *IEEE Transactions on Medical Imaging, 31*, 207–216.
Panagiotakopoulos, T. C., Lyras, D. P., Livaditis, M., Sgarbas, K. N., Anastassopoulos, G. C., & Lymberopoulos, D. K. (2010). A contextual data mining approach toward assisting the treatment of anxiety disorders. *IEEE Transactions on Information Technology in Biomedicine, 14*, 567–581.
Pandey, B., & Mishra, R. B. (2009). Knowledge and intelligent computing system in medicine. *Computers in Biology and Medicine, 39*, 215–230.
Papachristoudis, G., Diplaris, S., & Mitkas, P. A. (2010). SoFoCles: Feature filtering for microarray classification based on gene ontology. *Journal of Biomedical Informatics, 43*, 1–14.
Patel, V. L., Shortliffe, E. H., Stefanelli, M., Szolovits, P., Berthold, M. R., Bellazzi, R., et al. (2009). The coming of age of artificial intelligence in medicine. *Artificial Intelligence in Medicine, 46*, 5–17.
Patil, B., Joshi, R., Toshniwal, D., & Biradar, S. (2011). A new approach: Role of data mining in prediction of survival of burn patients. *Journal of Medical Systems, 35*, 1531–1542.
Patil, B. M., Joshi, R. C., & Toshniwal, D. (2010). Hybrid prediction model for type-2 diabetic patients. *Expert Systems with Applications, 37*, 8102–8108.
Paul, T. K., & Iba, H. (2009). Prediction of cancer class with majority voting genetic programming classifier using gene expression data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 6*, 353–367.
Pecchia, L., Melillo, P., & Bracale, M. (2011). Remote health monitoring of heart failure with data mining via CART method on HRV features. *IEEE Transactions on Biomedical Engineering, 58*, 800–804.
Pecchia, L., Melillo, P., Sansone, M., & Bracale, M. (2011). Discrimination power of short-term heart rate variability measures for CHF assessment. *IEEE Transactions on Information Technology in Biomedicine, 15*, 40–46.
Pechenizkiy, M., Tsymbal, A., & Puuronen, S. (2006). Local dimensionality reduction and supervised learning within natural clusters for biomedical data analysis. *Transactions on Information Technology in Biomedicine, 10*, 533–539.
Peña-Reyes, C. A., & Sipper, M. (2000). Evolutionary computation in medicine: an overview. *Artificial Intelligence in Medicine, 19*, 1–23.
Perner, P. (2002). Image mining: Issues, framework, a generic tool and its application to medical-image diagnosis. *Engineering Applications of Artificial Intelligence, 15*, 205–216.
Peterson, C., & Ringnér, M. (2003). Analyzing tumor gene expression profiles. *Artificial Intelligence in Medicine, 28*, 59–74.
Pham, H. N. A., & Triantaphyllou, E. (2009). An application of a new meta-heuristic for optimizing the classification accuracy when analyzing some medical datasets. *Expert Systems with Applications, 36*, 9240–9249.
Phillips-Wren, G., Sharkey, P., & Dy, S. M. (2008). Mining lung cancer patient data to assess healthcare resource utilization. *Expert Systems with Applications, 35*, 1611–1619.
Piatetsky-Shapiro, G. (2007). Data mining and knowledge discovery 1996 to 2005: Overcoming the hype and moving from "university" to "business" and "analytics". *Data Mining and Knowledge Discovery, 15*, 99–105.
Piatetsky-Shapiro, G., & Frawley, W. (1991). *Knowledge discovery in databases*. California: AAAI/MIT Press.
Pietraszek, T., & Tanner, A. (2005). Data mining and machine learning—Towards reducing false positives in intrusion detection. *Information Security Technical Report, 10*, 169–183.
Plant, C., Teipel, S. J., Oswald, A., Böhm, C., Meindl, T., Mourao-Miranda, J., et al. (2010). Automated detection of brain atrophy patterns based on MRI for the prediction of Alzheimer's disease. *NeuroImage, 50*, 162–174.
Podgorelec, V., Kokol, P., Stiglic, M. M., Heričko, M., & Rozman, I. (2005). Knowledge discovery with classification rules in a cardiovascular dataset. *Computer Methods and Programs in Biomedicine, 80*(Suppl. 1), S39–S49.
Polat, K. (2012). Application of attribute weighting method based on clustering centers to discrimination of linearly non-separable medical datasets. *Journal of Medical Systems, 36*, 2657–2673.
Polat, K., & Güneş, S. (2007). Breast cancer diagnosis using least square support vector machine. *Digital Signal Processing, 17*, 694–701.
Pospisil, P., Iyer, L., Adelstein, S., & Kassis, A. (2006). A combined approach to data mining of textual and structured data to identify cancer-related targets. *BMC Bioinformatics, 7*, 1–11.
Prati, R. C., Batista, G. E. A. P. A., & Monard, M. C. (2011). A survey on graphical methods for classification predictive performance evaluation. *IEEE Transactions on Knowledge and Data Engineering, 23*, 1601–1618.
Qiang, Y., Guo, Y., Li, X., Wang, Q., Chen, H., & Cuic, D. (2007). The diagnostic rules of peripheral lung cancer preliminary study based on data mining technique. *Journal of Nanjing Medical University, 21*, 190–195.
Ramírez, J., Górriz, J. M., Salas-Gonzalez, D., Romero, A., López, M., Álvarez, I., et al. (2013). Computer-aided diagnosis of Alzheimer's type dementia combining support vector machines and discriminant set of features. *Information Sciences, 237*, 59–72.
Ramirez, J. C. G., Cook, D. J., Peterson, L. L., & Peterson, D. M. (2000). Temporal pattern discovery in course-of-disease data. *IEEE Engineering in Medicine and Biology Magazine, 19*, 63–71.
Ramon, J., Fierens, D., Güiza, F., Meyfroidt, G., Blockeel, H., Bruynooghe, M., et al. (2007). Mining data from intensive care patients. *Advanced Engineering Informatics, 21*, 243–256.
Ramos-Pollán, R., Guevara-López, M., Suárez-Ortega, C., Díaz-Herrero, G., Franco-Valiente, J., Rubio-del-Solar, M., et al. (2012). Discovering mammography-based machine learning classifiers for breast cancer diagnosis. *Journal of Medical Systems, 36*, 2259–2269.
Raymer, M. L., Doom, T. E., Kuhn, L. A., & Punch, W. F. (2003). Knowledge discovery in medical and biological datasets using a hybrid Bayes classifier/evolutionary algorithm. *Transactions on Systems, Man, and Cybernetics Part B, 33*, 802–813.
Razavi, A., Gill, H., Åhlfeldt, H., & Shahsavar, N. (2007). Predicting metastasis in breast cancer: comparing a decision tree with domain experts. *Journal of Medical Systems, 31*, 263–273.
Razavi, A., Gill, H., Stål, O., Sundquist, M., Thorstenson, S., Åhlfeldt, H., et al. (2005). Exploring cancer register data to find risk factors for recurrence of breast cancer – application of canonical correlation analysis. *BMC Medical Informatics and Decision Making, 5*, 1–7.
Richards, G., Rayward-Smith, V. J., Sönksen, P. H., Carey, S., & Weng, C. (2001). Data mining for indicators of early mortality in a database of clinical records. *Artificial Intelligence in Medicine, 22*, 215–231.
Riganello, F., Candelieri, A., Quintieri, M., Conforti, D., & Dolce, G. (2010). Heart rate variability: An index of brain processing in vegetative state? An artificial intelligence, data mining study. *Clinical Neurophysiology, 121*, 2024–2034.
Roman, M., Jitaru, P., Agostini, M., Cozzi, G., Pucciarelli, S., Nitti, D., et al. (2012). Serum seleno-proteins status for colorectal cancer screening explored by data mining techniques – A multidisciplinary pilot study. *Microchemical Journal, 105*, 124–132.
Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. *Transactions on Systems, Man, and Cybernetics Part C, 40*, 601–618.
Romero, C., Ventura, S., & García, E. (2008). Data mining in course management systems: Moodle case study and tutorial. *Computers & Education, 51*, 368–384.
Roy Walker, P., Smith, B., Liu, Q. Y., Fazel Famili, A., Valdés, J. J., Liu, Z., et al. (2004). Data mining of gene expression changes in Alzheimer brain. *Artificial Intelligence in Medicine, 31*, 137–154.
Roychowdhury, A., Pratihar, D. K., Bose, N., Sankaranarayanan, K. P., & Sudhahar, N. (2004). Diagnosis of the diseases––Using a GA-fuzzy approach. *Information Sciences, 162*, 105–120.
Ryu, Y. U., Chandrasekaran, R., & Jacob, V. S. (2007). Breast cancer prediction using the isotonic separation technique. *European Journal of Operational Research, 181*, 842–854.
Šajn, L., & Kukar, M. (2011). Image processing and machine learning for fully automated probabilistic evaluation of medical images. *Computer Methods and Programs in Biomedicine, 104*, e75–e86.
Sallaberry, A., Pecheur, N., Bringay, S., Roche, M., & Teisseire, M. (2011). Sequential patterns mining and gene sequence visualization to discover novelty from microarray data. *Journal of Biomedical Informatics, 44*, 760–774.
Samanta, B., Bird, G. L., Kuijpers, M., Zimmerman, R. A., Jarvik, G. P., Wernovsky, G., et al. (2009). Prediction of periventricular leukomalacia. Part I: Selection of hemodynamic features using logistic regression and decision tree algorithms. *Artificial Intelligence in Medicine, 46*, 201–215.

Samuel, O. W., Omisore, M. O., & Ojokoh, B. A. (2013). A web based decision support system driven by fuzzy logic for the diagnosis of typhoid fever. *Expert Systems with Applications, 40*, 4164–4171.

Santos, R. S., Malheiros, S. M. F., Cavalheiro, S., & de Oliveira, J. M. P. (2013). A data mining system for providing analytical information on brain tumors to public health decision makers. *Computer Methods and Programs in Biomedicine, 109*, 269–282.

Schaefer, G., & Nakashima, T. (2010). Data mining of gene expression data by fuzzy and hybrid fuzzy methods. *IEEE Transactions on Information Technology in Biomedicine, 14*, 23–29.

Scheetz, L. J., Zhang, J., & Kolassa, J. (2009). Classification tree modeling to identify severe and moderate vehicular injuries in young and middle-aged adults. *Artificial Intelligence in Medicine, 45*, 1–10.

Semenova, T. (2004). Discovering patterns of medical practice in large administrative health databases. *Data & Knowledge Engineering, 51*, 149–160.

Shah, S., & Kusiak, A. (2007). Cancer gene search with data-mining and genetic algorithms. *Computers in Biology and Medicine, 37*, 251–261.

Shah, S. C., & Kusiak, A. (2004). Data mining and genetic algorithm based gene/SNP selection. *Artificial Intelligence in Medicine, 31*, 183–196.

Shah, S. C., Kusiak, A., & O'Donnell, M. A. (2006). Patient-recognition data-mining model for BCG-plus interferon immunotherapy bladder cancer treatment. *Computers in Biology and Medicine, 36*, 634–655.

Shaik, J., & Yeasin, M. (2009). Fuzzy-adaptive-subspace-iteration-based two-way clustering of microarray data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 6*, 244–259.

Shen, C.-P., Jigjidsuren, C., Dorjgochoo, S., Chen, C.-H., Chen, W.-H., Hsu, C.-K., et al. (2012). A data-mining framework for transnational healthcare system. *Journal of Medical Systems, 36*, 2565–2575.

Shen, C.-P., Kao, W.-C., Yang, Y.-Y., Hsu, M.-C., Wu, Y.-T., & Lai, F. (2012). Detection of cardiac arrhythmia in electrocardiograms using adaptive feature extraction and modified support vector machines. *Expert Systems with Applications, 39*, 7845–7852.

Shenghuo, Z., Dingding, W., Kai, Y., Tao, L., & Yihong, G. (2010). Feature selection for gene expression using model-based entropy. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 7*, 25–36.

Shilaskar, S., & Ghatol, A. (2013). Feature selection for medical diagnosis: Evaluation for cardiovascular diseases. *Expert Systems with Applications, 40*, 4146–4153.

Shun Ngan, P., Leung Wong, M., Lam, W., Leung, K. S., & Cheng, J. C. Y. (1999). Medical data mining using evolutionary computation. *Artificial Intelligence in Medicine, 16*, 73–96.

Shusaku (2000). Knowledge discovery in clinical databases and evaluation of discovered knowledge in outpatient clinic. *Information Sciences, 124*, 125–137.

Sigurdardottir, A. K., Jonsdottir, H., & Benediktsson, R. (2007). Outcomes of educational interventions in type 2 diabetes: WEKA data-mining analysis. *Patient Education and Counseling, 67*, 21–31.

Silva, Á., Cortez, P., Santos, M. F., Gomes, L., & Neves, J. (2006). Mortality assessment in intensive care units via adverse events using artificial neural networks. *Artificial Intelligence in Medicine, 36*, 223–234.

Silver, H., & Shmoish, M. (2008). Analysis of cognitive performance in schizophrenia patients and healthy individuals with unsupervised clustering models. *Psychiatry Research, 159*, 167–179.

Simek, K., Fujarewicz, K., Świerniak, A., Kimmel, M., Jarząb, B., Wiench, M., et al. (2004). Using SVD and SVM methods for selection, classification, clustering and modeling of DNA microarray data. *Engineering Applications of Artificial Intelligence, 17*, 417–427.

Šprogar, M., Lenič, M., & Alayon, S. (2002). Evolution in medical decision making. *Journal of Medical Systems, 26*, 479–489.

Straszecka, E. (2006). Combining uncertainty and imprecision in models of medical diagnosis. *Information Sciences, 176*, 3026–3059.

Su, C.-T., Wang, P.-C., Chen, Y.-C., & Chen, L.-F. (2012). Data mining techniques for assisting the diagnosis of pressure ulcer development in surgical patients. *Journal of Medical Systems, 36*, 2387–2399.

Su, C.-T., Yang, C.-H., Hsu, K.-H., & Chiu, W.-K. (2006). Data mining for the diagnosis of type II diabetes from three-dimensional body surface anthropometrical scanning data. *Computers & Mathematics with Applications, 51*, 1075–1092.

Subasi, A. (2012). Medical decision support system for diagnosis of neuromuscular disorders using DWT and fuzzy support vector machines. *Computers in Biology and Medicine, 42*, 806–815.

Suner, A., Çelikoğlu, C. C., Dicle, O., & Sökmen, S. (2012). Sequential decision tree using the analytic hierarchy process for decision support in rectal cancer. *Artificial Intelligence in Medicine, 56*, 59–68.

Sunghan, K., Scalzo, F., Bergsneider, M., Vespa, P., Martin, N., & Xiao, H. (2012). Noninvasive intracranial pressure assessment based on a data-mining approach using a nonlinear mapping function. *IEEE Transactions on Biomedical Engineering, 59*, 619–626.

Sut, N., & Simsek, O. (2011). Comparison of regression tree data mining methods for prediction of mortality in head injury. *Expert Systems with Applications, 38*, 15534–15539.

Taft, L. M., Evans, R. S., Shyu, C. R., Egger, M. J., Chawla, N., Mitchell, J. A., et al. (2009). Countering imbalanced datasets to improve adverse drug event predictive models in labor and delivery. *Journal of Biomedical Informatics, 42*, 356–364.

Talbert, D. A., Honeycutt, M., & Talbert, S. (2011). A machine learning and data mining framework to enable evolutionary improvement in trauma triage. In *Proceedings of the 7th international conference on machine learning and data mining in pattern recognition* (pp. 348–361). New York, NY: Springer-Verlag.

Tan, P.-N., Steinbach, M., & Kumar, V. (2006). *Introduction to data mining*. Boston: Pearson Education Inc.

Tang, S.-H., Chen, J.-X., Li, G., Wu, H.-W., Chen, C., Zhang, N., et al. (2010). Research on component law of Chinese patent medicine for anti-influenza and development of new recipes for anti-influenza by unsupervised data mining methods. *Journal of Traditional Chinese Medicine, 30*, 288–293.

Temurtas, H., Yumusak, N., & Temurtas, F. (2009). A comparative study on diabetes disease diagnosis using neural networks. *Expert Systems with Applications, 36*, 8610–8615.

Thongkam, J., Xu, G., Zhang, Y., & Huang, F. (2009). Toward breast cancer survivability prediction models through improving training space. *Expert Systems with Applications, 36*, 12200–12209.

Ting, H.-W., Wu, J.-T., Chan, C.-L., Lin, S.-L., & Chen, M.-H. (2010). Decision model for acute appendicitis treatment with decision tree technology—A modification of the alvarado scoring system. *Journal of the Chinese Medical Association, 73*, 401–406.

Toussi, M., Lamy, J.-B., Le Toumelin, P., & Venot, A. (2009). Using data mining techniques to explore physicians' therapeutic decisions when clinical guidelines do not provide recommendations: methods and example for type 2 diabetes. *BMC Medical Informatics and Decision Making, 9*, 1–12.

Tsipouras, M. G., Exarchos, T. P., Fotiadis, D. I., Kotsia, A. P., Vakalis, K. V., Naka, K. K., et al. (2008). Automated diagnosis of coronary artery disease based on data mining and fuzzy modeling. *IEEE Transactions on Information Technology in Biomedicine, 12*, 447–458.

Tsoi, A. C., Zhang, S., & Hagenbuchner, M. (2005). Pattern discovery on Australian medical claims data – A systematic approach. *IEEE Transactions on Knowledge and Data Engineering, 17*, 1420–1435.

Tsumoto, S. (1998). Automated extraction of medical expert system rules from clinical databases based on rough set theory. *Information Sciences, 112*, 67–84.

Tsumoto, S. (2004). Mining diagnostic rules from clinical databases using rough sets and medical diagnostic model. *Information Sciences, 162*, 65–80.

Uçar, T., & Karahoca, A. (2011). Predicting existence of mycobacterium tuberculosis on patients using data mining approaches. *Procedia Computer Science, 3*, 1404–1411.

Vatankhah, M., Asadpour, V., & Fazel-Rezai, R. (2012). Perceptual pain classification using ANFIS adapted RBF kernel support vector machine for therapeutic usage. *Applied Soft Computing*.

Voznuka, N., Granfeldt, H., Babic, A., Storm, M., Lönn, U., & Ahn, H. (2004). Report generation and data mining in the domain of thoracic surgery. *Journal of Medical Systems, 28*, 497–509.

Wagholikar, K. B., Sundararajan, V., & Deshpande, A. W. (2012). Modeling paradigms for medical diagnostic decision support: A survey and future directions. *Journal of Medical Systems, 36*, 3029–3049.

Wang, J., Li, M., Hu, Y.-T., & Zhu, Y. (2009). Comparison of hospital charge prediction models for gastric cancer patients: Neural network vs. decision tree models. *BMC Health Services Research, 9*, 1–6.

Wang, S.-L., Li, X., Zhang, S., Gui, J., & Huang, D.-S. (2010). Tumor classification by combining PNN classifier ensemble with neighborhood rough set based gene reduction. *Computers in Biology and Medicine, 40*, 179–189.

Wang, S., Lin, C.-J., Wu, C., & Chaovalitwongse, W. A. (2011). Early detection of numerical typing errors using data mining techniques. *Transactions on Systems, Man, and Cybernetics Part A, 41*, 1199–1212.

Wang, Y.-F., Chang, M.-Y., Chiang, R.-D., Hwang, L.-J., Lee, C.-M., & Wang, Y.-H. (2013). Mining medical data: A case study of endometriosis. *Journal of Medical Systems, 37*, 1–7.

Wang, Y., Ma, L., & Liu, P. (2009). Feature selection and syndrome prediction for liver cirrhosis in traditional Chinese medicine. *Computer Methods and Programs in Biomedicine, 95*, 249–257.

Warren Liao, T. (2011). Diagnosis of bladder cancers with small sample size via feature selection. *Expert Systems with Applications, 38*, 4649–4654.

WHO. (2013). The top 10 causes of death.

Wiggins, M., Saad, A., Litt, B., & Vachtsevanos, G. (2008). Evolving a Bayesian classifier for ECG-based age classification in medical applications. *Applied Soft Computing, 8*, 599–608.

Wongseree, W., Chaiyaratana, N., Vichittumaros, K., Winichagoon, P., & Fucharoen, S. (2007). Thalassaemia classification by neural networks and genetic programming. *Information Sciences, 177*, 771–786.

Yamaguchi, M., Kaseda, C., Yamazaki, K., & Kobayashi, M. (2006). Prediction of blood glucose level of type 1 diabetics using response surface methodology and data mining. *Medical and Biological Engineering and Computing, 44*, 451–457.

Yan, H., Zheng, J., Jiang, Y., Peng, C., & Xiao, S. (2008). Selecting critical clinical features for heart diseases diagnosis with a real-coded genetic algorithm. *Applied Soft Computing, 8*, 1105–1111.

Yang, C. C., Lin, W. T., Chen, H. M., & Shi, Y. H. (2009). Improving scheduling of emergency physicians using data mining analysis. *Expert Systems with Applications, 36*, 3378–3387.

Yanqing, J., Hao, Y., Dews, P., Mansour, A., Tran, J., Miller, R. E., et al. (2011). A potential causal association mining algorithm for screening adverse drug reactions in postmarketing surveillance. *IEEE Transactions on Information Technology in Biomedicine, 15*, 428–437.

Yardimci, A. (2009). Soft computing in medicine. *Applied Soft Computing, 9*, 1029–1043.

Yeh, D.-Y., Cheng, C.-H., & Chen, Y.-W. (2011). A predictive model for cerebrovascular disease using data mining. *Expert Systems with Applications, 38*, 8970–8977.

Yeh, J.-Y., Wu, T.-H., & Tsao, C.-W. (2011). Using data mining techniques to predict hospitalization of hemodialysis patients. *Decision Support Systems, 50*, 439–448.
Yeh, W.-C. (2012). Novel swarm optimization for mining classification rules on thyroid gland data. *Information Sciences, 197*, 65–76.
Yildirim, P., Çeken, Ç., Hassanpour, R., & Tolun, M. (2012). Prediction of Similarities among Rheumatic Diseases. *Journal of Medical Systems, 36*, 1485–1490.
Yinxia, L., Weidong, Z., Qi, Y., & Shuangshuang, C. (2012). Automatic seizure detection using wavelet transform and SVM in long-term intracranial EEG. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 20*, 749–755.
Yoo, I., Alafaireet, P., Marinov, M., Pena-Hernandez, K., Gopidi, R., Chang, J.-F., et al. (2012). Data mining in healthcare and biomedicine: A survey of the literature. *Journal of Medical Systems, 36*, 2431–2448.
Yoon, Y., Lee, J., Park, S., Bien, S., Chung, H. C., & Rha, S. Y. (2008). Direct integration of microarrays for selecting informative genes and phenotype classification. *Information Sciences, 178*, 88–105.
Zhang, S., Tjortjis, C., Zeng, X., Qiao, H., Buchan, I., & Keane, J. (2009). Comparing data mining methods with logistic regression in childhood obesity prediction. *Information Systems Frontiers, 11*, 449–460.
Zhang, Z., Shi, Y., & Gao, G. (2009). A rough set-based multiple criteria linear programming approach for the medical diagnosis and prognosis. *Expert Systems with Applications, 36*, 8932–8937.
Zhong, W., Chow, R., & He, J. (2012). Clinical charge profiles prediction for patients diagnosed with chronic diseases using multi-level support vector machine. *Expert Systems with Applications, 39*, 1474–1483.
Zhou, X., Chen, S., Liu, B., Zhang, R., Wang, Y., Li, P., et al. (2010). Development of traditional Chinese medicine clinical data warehouse for medical knowledge discovery and decision support. *Artificial Intelligence in Medicine, 48*, 139–152.