EASR

Engineering and Applied Science Research

https://www.tci-thaijo.org/index.php/easr/index

Published by the Faculty of Engineering, Khon Kaen University, Thailand

Applications of data mining in healthcare area: A survey

Ehsan Shirzad*[1)], Ghazal Ataei[2)] and Hamid Saadatfar[1)]

[1)]Department of Computer Engineering, Faculty of Electrical and Computer Engineering, University of Birjand, Birjand, Iran
[2)]Department of Biology, Faculty of Sciences, Payame Noor University, Tehran, Iran

Received 28 June 2020
Revised 7 August 2020
Accepted 25 August 2020

Abstract

Data mining is the modern way of discovering knowledge among databases that leads to statistical analysis, pattern recognition, and information prediction. Today, one of the most important applications of data mining is in the healthcare field which leads to many advances in this area in order to increase the effectiveness of treatments, reduce the risks, decrease the costs, better patient relationships, early disease diagnosis, and etc. This article attempts to provide a comprehensive overview with a new classification of services that data mining has created or facilitated in the healthcare field. It includes disease diagnosis, early detection of diseases, managing pandemic diseases, dimension reduction, health monitoring, treatment effectiveness, system biology, management of hospital resources, hospital ranking, customer relationship management, public health policy planning, fraud and abuse detection, and control data overload. Furthermore, the strengths and weaknesses of data mining in the healthcare field are discussed and future directions in this area are mentioned. Finally, it can be concluded that although data mining has abundant applications in the healthcare area, especially in the diagnosis and prediction of diseases and healthcare business, medical data mining is still young and needs more attention.

Keywords: Data mining, Machine learning, Healthcare, Medical informatics, Review

1. Introduction

Nowadays, everything generates data and this huge amount of data existing around the world (which leads to the emergence of big data phenomenon) needs to be refined like oil in order to discover useful knowledge [1]. Due to the high volume of data, the traditional methods of data processing are no longer responsive. So, a new method is needed that is data mining. Data mining is the science of extracting logical, useful, and usually understandable knowledge, including patterns, models, instructions, statistics, and so on by using computer systems.

Data mining has been used extensively for many purposes and one of which is healthcare. Data mining can greatly benefit healthcare applications and solve many problems. For example, data mining can help patient caretakers to detect diseases or other problems from the medical records, healthcare organizations to improve customer relationship management, patients to receive better and more affordable healthcare services, and physicians to find effective treatments and early prediction of diseases.

The data generated by the health organizations (which is used for data mining) is very enormous and complex. This data contains details about hospitals, patients, physicians, medical claims, treatment costs, and etc. Therefore, using data mining on such raw data can extract knowledge and generate important information about the various factors that are responsible for diseases (shown in Figure 1 [2]). So, using data mining in healthcare is becoming increasingly popular and essential.

In this paper, we try to review the works done in the area of using data mining in healthcare by the view of what services are provided by data mining for healthcare. We have provided new classifying with new categories of services that data mining has created or facilitated in healthcare. The categories include disease diagnosis (by using image processing and voice pathology), early detection of diseases, managing pandemic diseases, dimension reduction (reduce the number of symptoms needed to diagnose a disease), health monitoring, treatment effectiveness, system biology, management of hospital resources, hospital ranking, customer relationship management, public health policy planning, fraud and abuse detection, and control data overload.

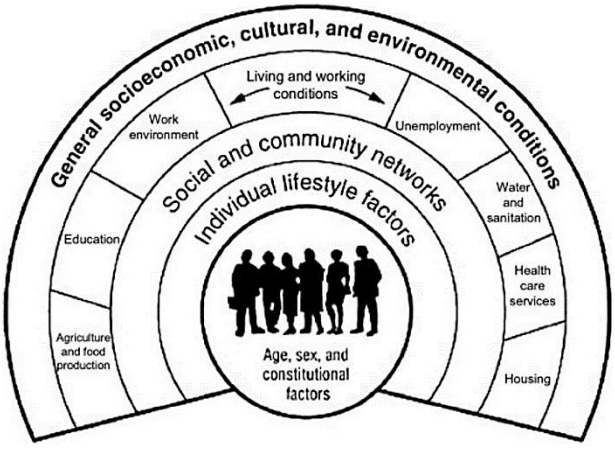

Figure 1 The factors responsible for diseases around the world [2].

*Corresponding author. Tel.: +9893 0835 0785
Email address: e71.shirzad@gmail.com
doi: 10.14456/easr.2021.34

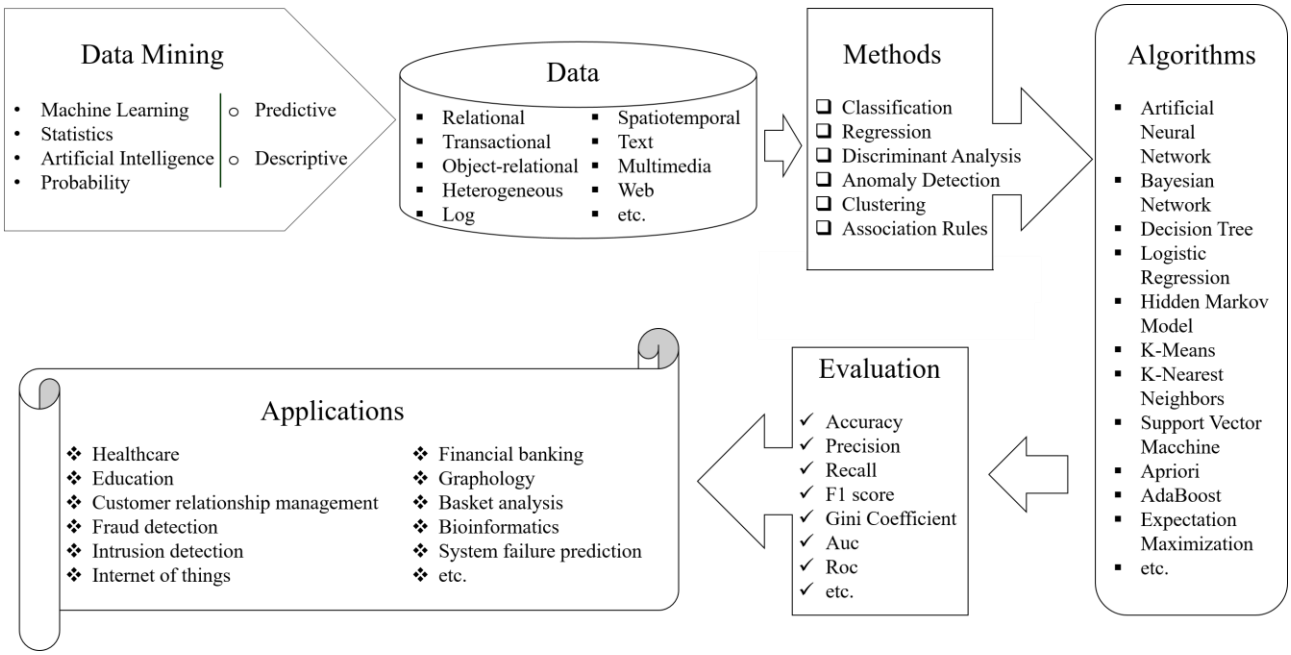

Figure 2 A general data mining scheme.

For each category, some works are surveyed and discussed with investigating the data mining method and its results.

The main contribution of our work can be listed as follows:

- Highlighting the role of data mining in the field of healthcare services.
- Categorizing the services that data mining provides in the healthcare area.
- Realizing the services by surveying some examples of scientific works done in each category.
- Noticing the challenges and future directions in the medical data mining area.

The remaining parts of the paper are organized as follows: section 2 provides an overview of data mining and its popular algorithms and section 3 discusses the applications and services of data mining in healthcare. While, we speak about the results and challenges of this area in section 4. Eventually, section 5 discusses future directions and concludes the paper.

2. Data mining

Before introducing and reviewing applications of data mining in healthcare, some basic data mining concepts and methods must be explained.

Data mining has been called by a variety of names. First of all, statisticians have used terms like “data fishing” or “data grubbing” to refer to what they considered a practice of analyzing data without a prior hypothesis [3]. But, the term “data mining” appeared around the 1990s in the database community. Other terms which have been used are Data Archeology, Information Harvesting, Information Discovery, and Knowledge Extraction [4].

Data mining is the extract of useful knowledge from large data repositories and emerged as a beneficial interdisciplinary field in computer science. Data mining techniques have been widely used in engineering, science, industry, and governments; and it has an impact on our society [5]. In fact, data mining is based on the concepts that include machine learning [6], artificial intelligence [7], probability [8], and statistics [9]. The scheme of data mining is visible in Figure 2.

Generally, there are two kinds of data mining models: predictive model and descriptive model. The descriptive model applies unsupervised learning functions to discover hidden patterns in a dataset [10]. On the other hand, the predictive model applies supervised learning functions to derive an operation from a labeled training dataset in order to predict future situations or values [11]. Most of the works reviewed in this paper are related to the predictive model (supervised learning).

The famous data mining algorithms that have the most utilization in healthcare works are Artificial Neural Networks [12], Decision Trees [13], Bayesian Networks [14], Support Vector Machines [15], Regression (Linear [16] and Logistic [17]), and K-Nearest Neighbor [18] which are the supervised models and Clustering (K-means [19] and Hierarchical [20]) and Association Rules [21] models which are unsupervised ones. In order to prevent an inessential increase in the volume of the article, it has been avoided to explain these algorithms one by one; so, the necessary information can be obtained by referring to the mentioned sources [12-21].

3. Data mining applications and services in healthcare

Today, the healthcare industry generates huge amounts of complex data about patients, hospitals, disease diagnosis, treatment costs, medical devices, and etc. These huge amounts of data (called healthcare big data) must be processed and analyzed to extract useful knowledge by using data mining in order to improve healthcare services. In this section, the data mining applications in healthcare are grouped and reviewed. The surveyed works are summarized in Table 1.

3.1 Disease diagnosis

Disease diagnosis is one of the important applications of medical data mining. This field of data mining has more to do with image processing and voice recognition, and is meant to provide a smart system for image and sound analysis in order to discover the things that a human being cannot normally find. Pandit and Shah [22] used data mining to improve the medical palmistry. In their proposed system, the human palm images enter the system. Then, the system applies digital image processing and analysis techniques on input images to identify certain features (island, spot, square, star, and grille) in the image and predict probable diseases (e.g. heart disease, urinal diseases, and problems with the reproductive system). Jia et al. [23] employed a system named DCNN to detect bleeding in gastrointestinal by processing 10,000 wireless capsule endoscopy

Table 1 A summary of the works surveyed in this section, categorized based on the healthcare services.

Healthcare aspect	Authors	Year	Data mining technique	A short description
Disease diagnosis	Pandit and Shah [22]	2011	Neural networks	Identify certain diseases by processing human palm images
	Jia et al. [23]	2016	Neural networks	Detect bleeding in Gastrointestinal by processing Endoscopy images
	Wimmer et al. [24]	2016	Neural networks	Detection of celiac disease using duodenum endoscopy images
	Pratumgul and Sa-ngiamwibool [25]	2016	Neural networks	Diagnosing diabetic retinopathy using image processing
	Sarraf et al. [26]	2016	Neural networks	Recognize Alzheimer disease in adults by using MRI and functional MRI images
	Karimi Rouzbahani and Daliri [27]	2011	K-nearest neighbor and support vector machine	Diagnosing Parkinson disease by processing voice signals
	Hashim et al. [28]	2012	Linear and quadratic classifiers (statistical-based)	Detection of depression in male patients by analyzing timing patterns of speech
	Rosa et al. [29]	1998	Neural networks	Identification of laryngeal pathologies by classifying acoustic measures
Early detection of diseases	Krishnaiah et al. [30]	2013	Decision tree, naive Bayes, and neural networks	Predict Lung cancer disease in patients by using their certain symptoms
	Wongtrairat et al. [31]	2016	linear regression	Analyzing the relationship between the brain response time and arm movement characteristics to early detection of Parkinson's disease
	Mokhtar and Elsayad [32]	2013	Decision tree, neural networks, and support vector machine	Predict the severity of breast masses
	Abdar et al. [33]	2017	C5.0 and CHAID decision trees	Predict the liver disease
Managing pandemic diseases	Wong et al. [34]	2003	Association rules and Bayesian networks	Detect outbreaks in their early stages
	Caduff [35]	2014	Traditional statistical analysis	Using crowd-sourcing data for the surveillance of the expansion of epidemic infectious diseases
	Gu et al. [36]	2015	Linear regression	Forecasting the rate of erythromelalgia disease outbreak
Dimension reduction	Joloudari et al. [37]	2020	Decision trees (C5.0, CHAID, and random trees) and support vector machine	Reduce the attributes needed to discover coronary artery disease
	Huda et al. [38]	2016	Decision trees and neural networks	Combine feature selection with classification algorithms to reduce the attributes of Oligodendroglioma tumor disease and diagnosis it faster
	Tayefi et al. [39]	2017	CART decision tree	Apply a decision tree to select the appropriate features to detect coronary heart disease
Health monitoring	Sareen et al [40]	2018	Naïve Bayes and decision trees	Monitoring and detecting symptoms of Ebola virus in patients
	Papamatthaiakis et al [41]	2010	Association rules mining	Monitoring old people's indoor activities
	Pandey [42]	2017	Logistic regression, support vector machine, and naïve Bayes	Monitoring a person's heartbeat rate and alarm the risk of myocardial infarction
	Verma et al [43]	2018	support vector machine, Decision trees, naive Bayes and K-nearest neighbor	Monitoring and detecting the waterborne diseases (e.g. cholera and hepatitis) in students
Treatment effectiveness	Aljumah et al [44]	2013	Regression	Analyzing the effectiveness of the treatment types for diabetes mellitus patients in different age groups
	Wilson et al. [45]	2004	MGPS (based on Bayesian analysis)	Discover the drug side effects in the US Drug Administration database

Table 2 (continued) A summary of the works surveyed in this section, categorized based on the healthcare services.

Healthcare aspect	Authors	Year	Data mining technique	A short description
System biology	Wang et al [46]	2000	Bayesian and neural networks	Classifying protein sequences
	Arango-Lopez et al [47]	2017	K-means clustering, Decision trees, and Bayesian networks	Classifying repeated genome sequences
	Manda [48]	2020	Association rule mining, clustering, and text mining	Investigating data mining techniques on gene ontology (GO) (gene expression) datasets
Management of hospital resources	J. Alapont et al [49]	2005	Linear regression, decision trees, and neural networks	Managing physical and human resources of hospitals
	Belciug [50]	2009	Hierarchical clustering	Estimating the length of stay of the patients in the hospital
	Ng et al. [51]	2006	Statistical-based methods	Predict the patients who require longer hospital care
	Ceglowski et al. [52]	2016	CART tree and neural networks	Classify patients for a better workflow scheduling in an emergency ward
	Testik et al. [53]	2012	Hierarchical / k-means clustering and CART tree	Recognize the patterns of arrival rates for blood donors
Hospital ranking	Cerrito et al [54]	2002	Decision trees and clustering	Examine hospitals data to rank them in case of cardiac care services
Customer relationship management	Rafalski [55]	2002	Traditional statistical analysis	Identify marketing opportunities in a hospital
	Song [56]	2018	Decision trees	Improve business management and decision making in pharmaceutical companies
Public health policy planning	Goodall [57]	1999	Multiple logistic regression (in COREPLUS) and Traditional statistic (in SAFS)	Investigating two famous systems in analyzing the outcomes of hospital cares and modeling the resources between clinical elements
	Lavrac et al. [58]	2007	Decision trees	Discovering the patterns among health centers to provide appropriate policy recommendations for them
	Kniesner and Leeth [59]	2004	Regression	Proposing a framework to make a healthcare policy for Mine Safety and Health Administration
Fraud and abuse detection	Ortega et al. [60]	2006	Neural networks	Proposing a fraud detection system in a Chilean private health insurance company
	Liou et al. [61]	2008	Neural networks, logistic regression, and decision trees	Detecting the fraud and abuses in diabetic outpatient services
	He et al. [62]	1998	K-nearest neighbor and Bayesian rules	Recognizing the weights of the features and classifying the general practitioner profiles
	Yang et al. [63]	2006	Association rules	Detecting frauds performed by medical service providers
Control data overload	Chandola et al. [64]	2013	Logistic regression	Studying the medical big data analytics and then applying a data mining method on a healthcare large dataset

(a non-invasive image video method for examination small bowel disease) images using a deep convolutional neural network. Wimmer et al. [24] proposed a system which learned features from the ImageNet dataset and then the learned feature vector fed to a convolutional neural network for classification and detection of celiac disease using duodenum endoscopic images. Pratumgul and Sa-ngiamwibool [25] proposed a classification method for diagnosing diabetic retinopathy using image processing and neural networks. They considered the basic morphology-related features such as blood vessels, exudates, micro-aneurysms, and texture-identification (entropy and homogeneity) to diagnosis retinopathy among diabetic patients. Their method eventually achieved a good performance, with an accuracy of 98.89%, a sensitivity of 99.26%, and a specificity of 97.77%. Sarraf et al. [26] used data mining on MRI and functional MRI images to recognize Alzheimer's disease in adults (above 75 years old). The authors used a convolutional neural network model to detect healthy or Alzheimer's brains and reported an accuracy of 99.9% for fMRI data and 98.84% for MRI data, respectively.

Voice recognition or voice pathology is another way to diagnosis a person's disease by processing his/her voice signals. This method works on the diseases that affect the vocal features of the patient. For example, Parkinson's disease causes changes in pressure at lips and shaking vocal cords. This made Karimi Rouzbahani and Daliri [27] to record and process voice signals in order to let a computer decide whether a person is suffering

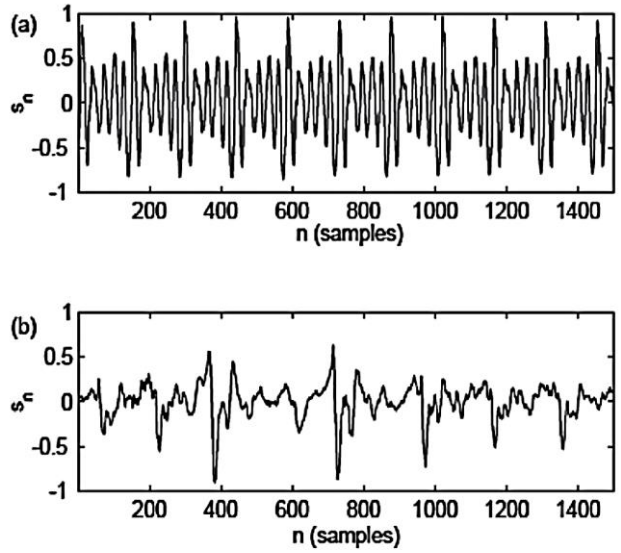

Figure 3 Two selected examples of speech signals: (a) healthy, (b) subject with PD. The horizontal axis is time in seconds, the vertical axis is signal amplitude [27].

from Parkinson's disease or not (Figure 3). They used KNN and SVM after feature selection and achieved about 93% accuracy with KNN in diagnosing a person's Parkinson's disease. Hashim et al. [28] found that the information in the speech signals contains characteristics changeable associated with depression in male patients. The analysis was based on the features related to the timing patterns of speech (voiced, unvoiced, and silence), specifically the transition parameters and interval probability density functions. They believe their method is a reliable way to detect and prevent of suicide attempts. Rosa et al. [29] also used ANN for the identification of laryngeal pathologies.

3.2 Early detection of diseases

Using classification algorithms to aid in the early detection of diseases is an important application of data mining in healthcare. Data mining can be used as a tool to assist in trends monitoring in the clinical experiments and treatment processes. Also, by using data mining and visualization, medical experts can recognize patterns and anomalies better than just looking at a set of normal tabulated data. For the pointed purposes, Krishnaiah et al. [30] used classification-based data mining techniques such as decision tree, naive Bayes, and artificial neural network on some healthcare data repositories in order to predict Lung cancer disease in patients by using their general symptoms related to lung cancer such as age, sex, wheezing, shortness of breath, and pain in shoulder, chest, and arms. Wongtrairat et al. [31] proposed a method to early detect Parkinson's disease by using linear regression to find the relationship between the brain response time and arm movement characteristics. They considered 120 persons in different age groups and calculated the brain signals for each person while moving arms. The authors analyzed the data using ANOVA and computed the linear correlation using Pearson, Kendall, and Spearman methods. Eventually, they showed that their proposed method has a high accuracy of 99.58% in identifying healthy people without Parkinson's disease. Mokhtar and Elsayad [32] have analyzed three different classification models including decision tree, artificial neural network, and support vector machine to predict the severity of breast masses (benign or malignant) using a mammographic dataset. Abdar et al. [33] also used C5.0 and CHAID decision trees to extract common risk factors of liver disease and predict it with the attributes including age, gender, total Bilirubin, direct Bilirubin, Alkaline Phosphatase, Sgpt Alanine Aminotransferase, Sgot Aspartate Aminotransferase, total proteins, Albumin, Albumin / Globulin ratio. They showed that a boosted C5.0 has better results on the used dataset.

3.3 Managing pandemic diseases

Another aspect of healthcare data mining is applying data mining for early detection and management of pandemic diseases. For example, Wong et al. [34] introduced WSARE, an algorithm based on association rules and Bayesian networks to detect outbreaks in their early stages. As they claimed, applying WSARE on simulation models has relatively accurate prediction results on simulated disease outbreaks. Caduff [35] conducted a research on using data mining and crowd-sourcing as a tool to use citizens in contributing to the surveillance of the expansion of epidemic infectious diseases. Also, Gu et al. [36] used the linear regression model in order to forecast the rate of an outbreak and spatial progression of erythromelalgia disease in China.

3.4 Dimension reduction

Data mining can be used to reduce the features needed to diagnose a disease. This application that has more to do with feature selection and decision trees, is to select a subset of features from whole features of a disease in order to speed up the diagnosis of it. In another word, this should remove the useless features. For instance, Joloudari et al. [37] ranked the important features needed to predict coronary artery disease based on decision trees-based models. Huda et al. [38] with highlighting the negative impact of imbalanced data on disease prediction, tried to use a feature selection combining an ensemble-based classification to achieve a fast and inexpensive diagnosis of Oligodendroglioma tumor (this tumor is one kind of brain tumor that has a good response to treatment if the tumor subtype is distinguished accurately). Eventually, their experiment results showed the advantage of their proposed approach in the problem of overcoming the imbalanced characteristics of medical data in the brain tumor classification problems. Tayefi et al. [39] also studied the factors affecting coronary heart disease. They provided a dataset of 2346 individuals including 1159 healthy participants and 1187 participants who had undergone coronary angiography (782 participants with positive angiography and 405 participants with negative angiography). They then entered 10 variables and applied a CART decision tree to predict the probability of the existence of coronary heart disease and select the appropriate features. They found that hs-CRP (highly sensitive C-reactive protein) with FBG (fasting blood glucose), gender, and age were more than 95% of the determinants for the presence of coronary heart disease. Their findings show that using decision trees is a good way to reduce the dimensions of an early disease prediction system.

3.5 Health monitoring

Traditional healthcare monitoring systems require patients to apply to the healthcare provider (center) such as a hospital for a scheduled visit or in case of an emergency situation. Such clinical visits might either lose the symptoms or be too late for any intervention. In addition, long-term healthcare costs increase year-by-year. Therefore, the healthcare systems are transforming from clinical settings to the patient or home-centered settings with the help of internet of things (IoT). In healthcare IoT or The Internet of Medical Things, the medical devices communicate with each other to share sensitive data [65] by using wireless sensor networks [66]. IoT facilities health monitoring by collecting health data using sensors in any situation (house, mobile, and wearable systems) of a person [67]. For instance, Figure 4 illustrates an application of IoT in the healthcare area [68].

On-demand services of healthcare IoT have undoubtedly dependent on data mining to analyze the collected raw data

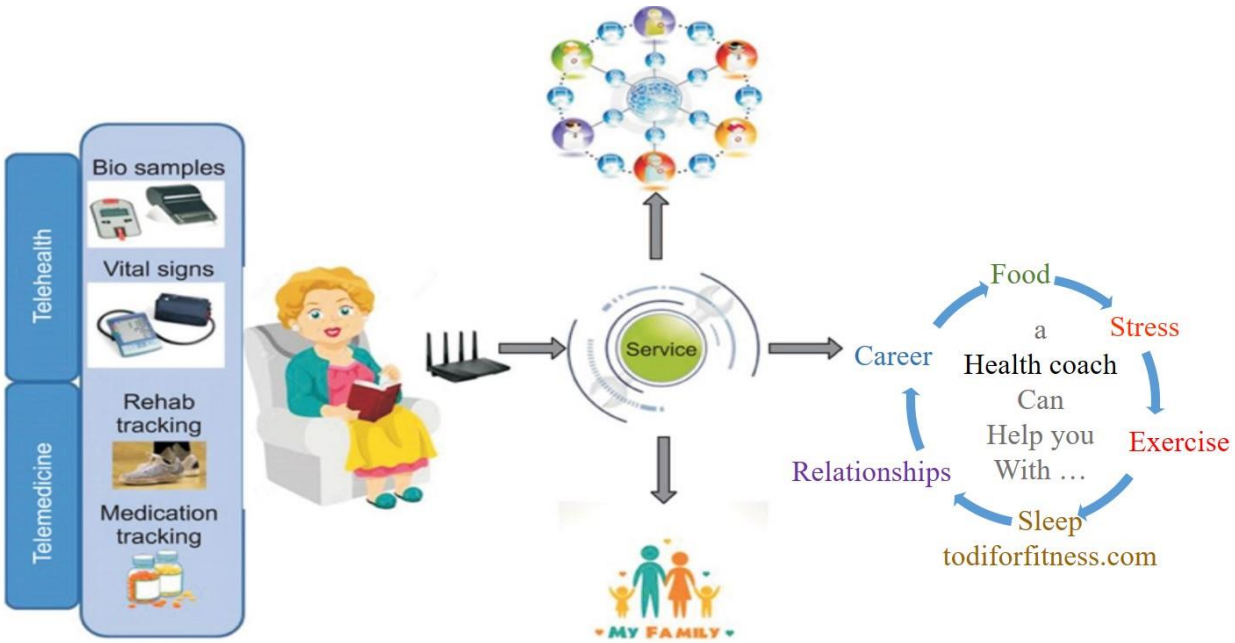

Figure 4 Elderly care is an important example of using IoT in the healthcare area [66].

and discover useful information. Therefore, this application of data mining in healthcare is growing fast. For this purpose, Sareen et al [40] proposed an architecture based on Internet of Things, cloud computing, and RFID for monitoring and detecting symptoms of Ebola virus in patients. They applied naïve Bayes and some decision trees algorithms on the data stored in the cloud and predict the disease with 94% accuracy. Papamatthaiakis et al [41] presented a method of indoor activity recognition based on the association rules mining algorithm to monitor old people's activities. Pandey [42] used IoT with data mining in order to monitor a person's heartbeat rate, detect that s/he is under stress or not, and alarm the situation when s/he is in real risk of myocardial infarction. The author tested logistic regression, SVM, and naïve Bayes classification algorithms and achieved near 100% accuracy. Verma et al [43] also designed a smart student health monitoring system to analyze the healthcare of the students. They focused on detecting waterborne diseases such as cholera, typhoid, hepatitis, and etc. among the students. In their proposed system, the health situation of the students first monitored by using IoT devices and then a data mining algorithm is applied to predict the waterborne diseases. The researchers tested SVM, DT, NB, and KNN algorithms and found that the DT classifier is the best model with 85.19% accuracy.

3.6 Treatment effectiveness

By using data mining, physicians and patients can evaluate the effectiveness of medical therapies, analyze existing therapies, and determine which method is best. Data mining also can assist them to compare the causes, symptoms, and courses of treatments, in order to identify specific side effects of treatments; so that, they can make appropriate decisions and reduce the risks of treatments. For instance, in [44], datasets of Non-Communicable Diseases (NCD) risk factors in Saudi Arabia were obtained from the World Health Organization (WHO) and studied and analyzed to recognize the effectiveness of various treatment types for diabetes mellitus patients in different age groups. The authors applied a regression technique due to the numeric nature of the data and used the Oracle Data Miner as a mining tool. They concluded that in diabetic patients, drug treatments can be delayed to avoid side effects; but, an immediate drug treatment is needed. They also suggested some preferential orders according to the data mining results for better treatment of diabetes. Data mining also helps to discover adverse drug events (ADE). Some drugs are harmful to humans in a special situation (for example in long-term consumption or drug interaction) that may not be normally detected by a physician. Wilson et al. [45] showed that the US Food and Drug Administration is using data mining to discover knowledge about drug side effects in their database. They indicated that the algorithm which is called MGPS (Multi-item Gamma Poisson Shrinker) was able to find 67% of ADEs five years before they could be detected using traditional ways.

3.7 System biology

Systems biology is the mathematical and computational analysis and modeling of complex biological systems. Data mining is a basic tool for system biology. A biological database contains a great diversity of data types, sometimes with rich relational structure and it needs multi-relational data mining techniques in order to explore databases and extract information from DNA, RNA, and proteins. In this way, Wang et al [46] proposed a Bayesian neural network approach in order to classify protein sequences. They tried to develop a new algorithm to extract the global/local similarities from the sequences that are used as input attributes of the Bayesian neural network and develop new measurements for evaluating the significance of 2-gram patterns and frequently occurring motifs used in classifying the sequences and eventually compare the proposed algorithm to similar ones. Arango-Lopez et al [47] also used classification algorithms to analyze and classify repeated genome sequences of the Robusta coffee tree. A good review on the applications of data mining techniques such as association rule mining, clustering, and text mining on gene ontology (GO) (gene expression) datasets can be found in the work of Manda [48] too.

3.8 Management of hospital resources

Data mining helps to manage the hospital resources which is an important application in healthcare. Using data mining, it is possible to prioritize the patients based on the complexity of the patient disease, predict the waiting/queuing time of patients, hospital/clinic overcrowded times, and etc. So that, the patients will get more effective treatment in a timely and better manner. J. Alapont et al [49] proposed an automated tool using data mining for managing the physical and human resources of

hospitals. Belciug [50] used a hierarchical clustering approach to divide the patients into different clusters according to their length of stay in the hospital in order to enhance the managing of hospital resources. Ng et al. [51] used data mining to predict the patients who require longer hospital care. This helps to construct a short to medium term hospital resources planning for improving inpatient/outpatient care. Ceglowski et al. [52] also used data mining techniques to classify patients by finding relationships between the patient urgency, treatment and disposal, and the occurrence of queues for treatment. These features were used as the inputs for the discrete event simulation model for workflow scheduling in an emergency ward. Testik et al. [53] used data mining to recognize the patterns of differences between daily and weekly arrival rates for blood donors. They used this information in order to provide an adaptive work scheduling for a facilitated healthcare system in hospitals.

3.9 Hospital ranking

Data mining techniques can be helpful in analyzing the hospital details in order to determine their ranks. Hospital ranking is based on the hospitals' capability to handle high-risk patients. A hospital with a higher rank can handle the high-risk patients on its top priority while a lower rank one does not consider the risk factors. The determined ranks are based on information reported by healthcare providers which may be not truly honest. Data mining is a way to test reporting ways by using classification, clustering, and association analyses on attributes of diseases and codes (risk factors) [54]. It is a fair way for meaningful comparisons across hospitals.

3.10 Customer relationship management (CRM)

Data Mining assists the healthcare institutes to understand their customers' priories, requirements, behaviors, and usage patterns in order to make better relations with them or predict the health products that they may like to buy. For these purposes, Rafalski [55] studied the use of data mining for CRM and healthcare marketing in the Sinai Health System. Furthermore, data mining and healthcare CRM can make a profit for pharmaceutical companies. For example, Song [56] with highlighting the benefits of using data mining in the CRM of the pharmaceutical industry, proposed a framework based on decision trees to achieve a better knowledge about customer needs and thus improve the business management and decision making in pharmaceutical companies.

3.11 Public health policy planning

Data mining plays an important role to make effective policies about healthcare in order to improve the health services quality and the costs. COREPLUS and SAFS are two famous models that were developed by HDS (Healthcare Design Systems) using data mining techniques to analyze the results and costs of medical services supplied by clinics and hospitals [57]. COREPLUS (Clinical Outcomes Resource Evaluation Plus) is a system to analyze outcomes of hospital cares. SAFS (Severity Adjustment Factor computations) is a system to model resources between clinical experts, healthcare staff, statisticians, managers, and marketers. Lavrac et al. [58] also combined GIS (Geographic Information System) with a data mining tool (a decision tree method in Weka) to analyze similarities between health centers in Slovenia. They could discover patterns among health centers by using data mining in order to provide appropriate policy recommendations for their Institute of Public Health. They concluded that data mining and decision support methods can lead to better performance in health policy decision making. Kniesner and Leeth [59] tried to propose a data mining framework to make a healthcare policy for MSHA (Mine Safety and Health Administration) in order to use regression methods on mines' safety datasets for more heart disease screening or defibrillators at worksites.

3.12 Fraud and abuse detection

Fraud/abuse is a big problem in the financial subject of healthcare services such as insurance. By using data mining, experts can appoint norms and then recognize the abnormal patterns in the claims of physicians, laboratories, clinics, or other healthcare participants. Data mining can highlight inappropriate prescriptions or referrals and fraudulent insurance and medical claims. For example, Ortega et al. [60] proposed a fraud detection system in a Chilean private health insurance company based on multilayer perceptron neural networks each one of healthcare entities (e.g. medical claims, affiliates, medical professionals, and employers) involved in a fraud or abuse problem. Liou et al. [61] used neural networks, logistic regression, and decision trees to detect fraud and abuses in diabetic outpatient services, and determined the variables which were most significant for the classification (e.g. drug cost, consultation/treatment fees, diagnosis/dispensing fees, and medical expenditure). The authors showed that their method works with only an error rate of 9%. He et al. [62] combined the KNN and genetic algorithms to recognize the weights of the features in profile data of the general practitioners. They then used Bayesian rules to classify the general practitioner profiles. Along with classification methods, association rules algorithms have been widely used in studies such as [63] to detect frauds performed by medical service providers in healthcare systems.

3.13 Control data overload

The last and simplest application of data mining in the healthcare area is to encounter with very large datasets that have led to the emergence of the health Big Data phenomenon. Nowadays, there is so much medical data produced that it cannot be dealt with and stored traditionally. Data mining helps to analyze the generated medical data to diagnosis outlier and noise data records and store only useful data [64].

4. Issues and challenges

While data mining has a variety of useful applications in the healthcare area and its usage is growing day by day, it is facing with some limitations and challenges as follows:

1. Raw medical data (which is the inputs for data mining processes) is very heterogeneous and complex in nature, because, it is collected from different sources such as laboratory reports, patient conversations, physician examinations, and etc. Therefore, the complication in healthcare data is one of the significant obstacles for analyzing it. So, it is essential to collect, integrate, and maintain the quality of raw medical datasets for a data mining process to reach effective results.
2. Another issue in healthcare data mining is the scarcity of useful real medical data. Healthcare organizations do not like to share their data due to monopolization and ethical and privacy concerns. For data mining to be more accurate, needing a significant amount of real records is undeniable.
3. While it seems that a data warehouse should be built before data mining is efforted, starting up a data warehouse is very costly and time-consuming. Furthermore, the design of a healthcare data warehouse should be faultless due to the sensitivity of the medical results.
4. Healthcare data mining processes are complicated. Unlike standard data mining practices that simply start with a dataset without a defined hypothesis, in medical researches, data mining starts with a hypothesis and then the results should be adjusted to fit the hypothesis. This makes a healthcare data mining process more complex.

5. Traditional data mining mentions comprehensive patterns and trends, but, data mining in medicine is more interested in the minority samples that do not match with the patterns and trends. This diversity is due to the sensitivity that a slight difference could change the balance between life and death. For example, both COVID-19 and the flu have many similar symptoms, but, they need a very accurate diagnosis. This is another reason for the complexity of healthcare data mining processes.

6. In many of the surveyed papers, the results are usually ambiguous and cautious. They usually report encouraging results but recommend further study. This lack of conviction shows that data mining is still unreliable in some particular parts of the healthcare area.

7. Even with reliable results of data mining, physicians usually resist to change their acts and habits. It is a bigger problem, because, data mining results need to be put into practice and it depends on the collaboration of health practitioners with specialists in this area.

8. Another challenge in healthcare data mining is the law and ethics issues. As mentioned earlier, the law sometimes prohibits the dissemination of medical data and it limits the extension of data mining researches. For example, Federal Privacy Rule that implements the HIPPA (Health Insurance Portability and Accountability Act) prohibits any share or disclosure of health information for marketing purposes (although it has a marketing title, it makes many limitations on research conductions). The ethical issue comes due to the profitability aspect of data mining. Many data mining firms might extract datasets and sell it for their own profits. Healthcare service providers such as hospitals and clinics have commercial aspects too and may violate ethics for more financial gain. A hospital can sell the collected data under the title of research purposes for its own profit. This business might be legal but it is totally unethical. Some international law should be formed to cover transnational and ethical aspects of healthcare data mining.

9. Another problem in healthcare data mining is the lack of expertise and the need for more training. As can be seen in the works done in this area, authors usually interested in using simple data mining models such as SVM and MLP, while more complex novel algorithms like Random Forest and Boltzmann Machine have been reported to give better results. Furthermore, some articles with the keyword “data mining” in their titles, just provided a simple use of statistical graphs. It shows that a new definition of data mining in the healthcare sector is needed and more experts must be trained in this area.

10. The last challenge that needs to be mentioned is the security and inference speed issues in on-demand healthcare monitoring services. With the advent of 5G technology that is related to IoT on-demand services, the health IoT services which usually used for medical monitoring need data mining approaches with more learning (modeling) speed to provide a high-speed inference system in order to convert raw collected data into useful information. The security of the data acquired at any moment by sensors and transmitted to a cloud through the Internet for data mining purposes must also be ensured to maintain the confidentiality and integrity of the data. Of course, data mining has also come to the aid of communication security and many studies such as [69-71] have been done in the field of information security in communication networks, especially in the Internet of Things; but, due to the sensitivity of medical information, more researches should be done on the security and facility of healthcare data transfer.

The existence of these challenges and issues in healthcare data mining indicates that this field is still young and needs more attention. It requires more management and financial supports, cleared user expectations, more expertise, a sufficient volume of validated data, and trust to implement the result. In other words, all parties involved in this field have to collaborate to provide an intensive planning, good project supports, and technological preparation work. Furthermore, physicians and health practitioners must be convinced about the usefulness and trustworthiness of data mining and be willing to change their work processes.

5. Conclusions

Data mining can play an important role in the medical industry and this role is getting more colorful day by day. This paper provided an overview of the works done in the area of using data mining approaches in the healthcare field. We first surveyed the popular and important data mining models and algorithms that were supervised (including Artificial Neural Networks, Bayesian Networks, Decision Trees, Support Vector Machines, and Regression models) and unsupervised (including Clustering models and Association Rules). We then classified data mining applications in the healthcare field (including disease diagnosis, early detection of diseases, managing pandemic diseases, dimension reduction, health monitoring, treatment effectiveness, system biology, management of hospital resources, hospital ranking, customer relationship management, public health policy planning, fraud and abuse detection, and control data overload) and reviewed the works in this area separated by this classification. Eventually, we discussed the challenges in the healthcare data mining field.

Many applications of medical data mining were reviewed, and for each, several examples were provided from recent researches. Some applications are more popular among researchers and have been developed more. For example, more researches have been done in diagnosing a disease by image processing or predicting the probability of a disease occurrence by using some features of individuals. Also, healthcare IoT and healthcare business are hot topics that need data mining to grow. Among the data mining techniques, neural network-based algorithms (deep learning) that show stronger results on medical data, especially in image processing, and decision tree-based algorithms that provide understandable knowledge along with predicting and reducing the dimensions are more popular. Of course, attention to other similar successful algorithms such as Recurrent Neural Networks, Long Short-Term Memory Networks, and Random Forests was lacking in the reviewed articles.

The provided survey can help healthcare organizations and experts to find ideas about the goals of extract knowledge from their own database systems. Developing an efficient data mining tool for an application can reduce the cost and time constraint in terms of human resources and expertise. Data mining also helps to explore knowledge from the medical datasets that are noisy, irrelevant, and massive. Eventually, this survey can be used as a platform for other researchers in this area as a starting point for their own study.

Lastly, we must denote that despite we tried to provide a structured approach, we cannot claim that we have covered all works in this area. Nevertheless, we hope this paper can make a contribution to all parties involved in the healthcare area to learn about the benefits and practices of data mining in their directions.

Because the success of a healthcare data mining project strongly depends on the availability of clean healthcare data, for the future direction of the healthcare industry experts, it is critical to consider how medical data can be better captured, stored, prepared, and shared. Meanwhile, our future direction for the data mining experts is to enhance the data mining approaches by using hybrid models in order to increase the accuracy of the results and eliminate the limitations in medical data mining applications.

6. References

[1] Shirzad E, Saadatfar H. Job failure prediction in Hadoop based on log file analysis. Int J Comput Appl. 2020;29:1732081.

[2] Kandwal R, Garg P, Garg R. Health GIS and HIV/AIDS studies: perspective and retrospective. J Biomed Informat. 2009;42(4):748-55.
[3] Lovell MC. Data mining. The Rev Econ Stat. 1983;65(1): 1-12.
[4] Piatetsky-Shapiro G. The journey of knowledge discovery. In: Gaber MM, editor. Journeys to data mining. Berlin: Springer; 2012. p. 173-96.
[5] Chakrabarti S, Ester M, Fayyad U, Gehrke J, Han J, Morishita S, et al. Data mining curriculum: a proposal (Version 1.0). Intensive working group of ACM SIGKDD curriculum committee. 2006:1-10.
[6] Kubat M. An introduction to machine learning. 2nd ed. Switzerland: Springer; 2017.
[7] Tecuci G. Artificial intelligence. Wiley Interdiscip Rev Comput Stat. 2012;4(2):168-80.
[8] Karegar M, Isazadeh A, Fartash F, Saderi T, Navin AH. Data-mining by probability-based patterns. International Conference on Information Technology Interfaces; 2008 June 23-26; Cavtat, Croatia. USA: IEEE; 2014. p. 353-60.
[9] Hill T, Lewicki P. Statistics: methods and applications: a comprehensive reference for science, industry, and data mining. USA: StatSoft; 2006.
[10] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. In: Hastie T, Tibshirani R, Friedman J, editors. Unsupervised learning. 2nd ed. New York: Springer; 2009. p. 485-585.
[11] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. In: Hastie T, Tibshirani R, Friedman J, editors. Overview of supervised learning. 2nd ed. New York: Springer; 2009. p. 9-41.
[12] Graupe D. Principles of artificial neural networks. 3rd ed. Singapore: World Scientific; 2013.
[13] Premchaisawatt S, Ruangchaijatupon N. Enhancing indoor positioning based on filter partitioning cascade machine learning models. Eng Appl Sci Res. 2016;43(3):146-52.
[14] Saadatfar H, Fadishei H, Deldari H. Predicting job failures in AuverGrid based on workload log analysis. New Generat Comput. 2012;30(1):73-94.
[15] Noyunsan C, Katanyukul T, Saikaew K. Performance evaluation of supervised learning algorithms with various training data sizes and missing attributes. Eng Appl Sci Res. 2018;45(3):221-9.
[16] Thongkam J, Sukmak V, Mayusiri W. A comparison of regression analysis for predicting the daily number of anxiety-related outpatient visits with different time series data mining. Eng Appl Sci Res. 2015;42(3):243-9.
[17] Peng CY, Lee KL, Ingersoll GM. An introduction to logistic regression analysis and reporting. J Educ Res. 2002;96(1):3-14.
[18] Ayyad SM, Saleh AI, Labib LM. Gene expression cancer classification using modified K-Nearest Neighbors technique. Bio Syste. 2019;176:41-51.
[19] Jain AK. Data clustering: 50 years beyond K-means. Pattern Recogn Lett. 2010;31(8):651-66.
[20] Kimes PK, Liu Y, Neil Hayes D, Marron JS. Statistical significance for hierarchical clustering. Biometrics. 2017;73(3):811-21.
[21] Thongkam J, Sukmak V. Enhancing the performance of association rule models by filtering instances in colorectal cancer patients. Eng Appl Sci Res. 2017;44(2):76-83.
[22] Pandit H, Shah DM. Application of digital image processing and analysis in healthcare based on medical palmistry. IJCA. 2011:56-9.
[23] Jia X, Meng MQ-H. A deep convolutional neural network for bleeding detection in wireless capsule endoscopy images. 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); 2016 Aug 16-20; Orlando, USA. USA: IEEE; 2016. p. 639-42.
[24] Wimmer G, Hegenbart S, Vecsei A, Uhl A. Convolutional neural network architectures for the automated diagnosis of celiac disease. In: Peter T, editor. Computer-Assisted and Robotic Endoscopy. Cham: Springer; 2016. p. 104-13.
[25] Pratumgul W, Sa-ngiamwibool W. Classification of diabetic retinopathy using artificial neural network. Eng Appl Sci Res. 2016;43(1):74-7.
[26] Sarraf S, DeSouza DD, Anderson JAE, Tofighi G. DeepAD: Alzheimer's disease classification via deep convolutional neural networks using MRI and fMRI. Bio Rxiv. 2016:1-6.
[27] Karimi-Rouzbahani H, Daliri M. Diagnosis of Parkinson's disease in human using voice signals. Basic Clin Neurosci. 2011;2(3):12-20.
[28] Hashim NW, Wilkes M, Salomon R, Meggs J. Analysis of timing pattern of speech as possible indicator for near-term suicidal risk and depression in male patients. 2012 International Conference on Signal Processing Systems (ICSPS 2012); 2012 Dec 21-22; Kuala Lumpur, Malaysia. Singapore: IACSIT Press; 2012. p. 6-13.
[29] Rosa MdO, Pereira JC, Carvalho A. Evaluation of neural classifiers using statistic methods for identification of laryngeal pathologies. Proceedings 5th Brazilian Symposium on Neural Networks (Cat No 98EX209); 1998 Dec 9-11; Belo Horizonte, Brazil. USA: IEEE; 2002.
[30] Krishnaiah V, Narsimha G, Chandra DNS. Diagnosis of lung cancer prediction system using data mining classification techniques. Int J Comput Sci Inform Tech. 2013;4(1):39-45.
[31] Wongtrairat W, Namwet P, Pornnimitra S. Early detection of Parkinson's diseases by using the relationship between time response and movement characteristics of human arms. Eng Appl Sci Res. 2016;43(3):127-34.
[32] Mokhtar SA, Elsayad A. Predicting the severity of breast masses with data mining methods. arXiv:1305.7057. 2013:1-9.
[33] Abdar M, Zomorodi-Moghadam M, Das R, Ting IH. Performance analysis of classification algorithms on early detection of liver disease. Expert Syst Appl. 2017;67: 239-51.
[34] Wong WK, Moore A, Cooper G, Wagner M. WSARE: What's strange about recent events?. J Urban Health. 2003;80(1):66-75.
[35] Caduff C. Sick weather ahead: on data-mining, crowd-sourcing and white noise. Camb Anthropol. 2014;32(1): 32-46.
[36] Gu Y, Chen F, Liu T, Lv X, Shao Z, Lin H, et al. Early detection of an epidemic erythromelalgia outbreak using Baidu search data. Sci Rep. 2015;5(1):12649.
[37] Joloudari JH, Hassannataj Joloudari E, Saadatfar H, GhasemiGol M, Razavi SM, Mosavi A, et al. Coronary artery disease diagnosis; ranking the significant features using a random trees model. Int J Environ Res Publ Health. 2020;17(3):731.
[38] Huda S, Yearwood J, Jelinek HF, Hassan MM, Fortino G, Buckland M. A hybrid feature selection with ensemble classification for imbalanced healthcare data: A case study for brain tumor diagnosis. IEEE Access. 2016;4:9145-54.
[39] Tayefi M, Tajfard M, Saffar S, Hanachi P, Amirabadizadeh AR, Esmaeily H, et al. hs-CRP is strongly associated with coronary heart disease (CHD): a data mining approach using decision tree algorithm. Comput Meth Programs Biomed. 2017;141:105-9.
[40] Sareen S, Sood SK, Gupta SK. IoT-based cloud framework to control Ebola virus outbreak. J Ambient Intell Humaniz Comput. 2018;9(3):459-76.
[41] Papamatthaiakis G, Polyzos GC, Xylomenos G. Monitoring and modeling simple everyday activities of the elderly at home. 2010 7th IEEE Consumer

Communications and Networking Conference; 2010 Jan 9-12; Las Vegas, USA. USA: IEEE; 2010. p. 1-5.
[42] Pandey PS. Machine learning and IoT for prediction and detection of stress. 2017 17th International Conference on Computational Science and Its Applications (ICCSA); 2017 July 3-6; Trieste, Italy. USA: IEEE; 2017.
[43] Verma P, Sood SK, Kalra S. Cloud-centric IoT based student healthcare monitoring framework. J Ambient Intell Humaniz Comput. 2018;9(5):1293-309.
[44] Aljumah AA, Ahamad MG, Siddiqui MK. Application of data mining: Diabetes health care in young and old patients. J King Saud Univ Comp Info Sci. 2013;25(2):127-36.
[45] Wilson AM, Thabane L, Holbrook A. Application of data mining techniques in pharmacovigilance. Br J Clin Psychol. 2004;57(2):127-34.
[46] Wang JT, Ma Q, Shasha D, Wu CH. Application of neural networks to biological data mining: a case study in protein sequence classification. Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining; 2000 Aug; Massachusetts, USA. USA: Association for Computing Machinery; 2000. p. 305-9.
[47] Arango-Lopez J, Orozco-Arias S, Salazar JA, Guyot R. Application of data mining algorithms to classify biological data: the coffea canephora genome case. In: Solano A, Ordoñez H, editors. Advances in Computing. Cham: Springer; 2017. p. 156-70.
[48] Manda P. Data mining powered by the gene ontology. J Data Min Knowl Discov. 2020;10(3):e1359.
[49] Alapont J, Bella-Sanjuan A, Ferri C, Hernandez-Orallo J, Llopis-Llopis J, Ramirez-Quintana M. Specialised tools for automating data mining for hospital management. Proc First East European Conference on Health Care Modelling and Computation; 2005 Aug 31 – Sep 2; Craiova, Romania.
[50] Belciug S. Patients length of stay grouping using the hierarchical clustering algorithm. Math Comp Sci Ser. 2009;36(2):79-84.
[51] Ng SK, McLachlan GJ, Lee AH. An incremental EM-based learning approach for on-line prediction of hospital resource utilization. Artif Intell Med. 2006;36(3):257-67.
[52] Ceglowski R, Churilov L, Wasserthiel J. Combining Data Mining and Discrete Event Simulation for a Value-Added View of a Hospital Emergency Department. J Oper Res Soc. 2016;1(1):119-38.
[53] Testik MC, Ozkaya BY, Aksu S, Ozcebe OI. Discovering blood donor arrival patterns using data mining: a method to investigate service quality at blood centers. J Med Syst. 2012;36(2):579-94.
[54] Cerrito PB, Cox JA, Mayes M, Thompson W. Using text analysis to examine ICD-9 codes to determine uniformity in the reporting of Medpar® data. Proc AMIA Symp. 2002:992.
[55] Rafalski E. Using data mining/data repository methods to identify marketing opportunities in health care. J Consum Market. 2002;19(7):607-13.
[56] Song G. Application of data mining technology in the CRM of pharmaceutical industry. 2018 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS); 2018 Jan 25-26; Xiamen, China. USA: IEEE; 2018. p. 61-4.
[57] Goodall CR. Data mining of massive datasets in healthcare. J Comput Graph Stat. 1999;8(3):620-34.
[58] Lavrac N, Bohanec M, Pur A, Cestnik B, Debeljak M, Kobler A. Data mining and visualization for decision support and modeling of public health-care resources. J Biomed Informat. 2007;40(4):438-47.
[59] Kniesner TJ, Leeth JD. Data mining mining data: MSHA enforcement efforts, underground coal mine safety, and new health policy implications. J Risk Uncertainty. 2004;29(2):83-111.
[60] Ortega PA, Figueroa CJ, Ruz GA. A medical claim Fraud/Abuse detection dystem based on data mining: A case study in Chile. Proceedings of the 2006 International Conference on Data Mining; 2006 Jun 26-29; Las Vegas, USA. USA: CSREA Press; 2006. p. 224-31.
[61] Liou FM, Tang YC, Chen JY. Detecting hospital fraud and claim abuse through diabetic outpatient services. Health Care Manag Sci. 2008;11(4):353-8.
[62] He H, Graco W, Yao X. Application of genetic algorithm and k-nearest neighbour method in medical fraud detection. In: McKay B, Yao X, Newton C.S, Kim JH, Furuhashi T, editors. Simulated Evolution and Learning. Berlin: Springer; 1998. p. 74-81.
[63] Yang WS, Hwang SY. A process-mining framework for the detection of healthcare fraud and abuse. Expert Syst Appl. 2006;31(1):56-68.
[64] Chandola V, Sukumar SR, Schryver JC. Knowledge discovery from massive healthcare claims data. Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining; 2013 Aug 11-14; Chicago, USA. USA: Association for Computing Machinery; 2013. p. 1312-20.
[65] RM SP, Maddikunta PK, Parimala M, Koppu S, Reddy T, Chowdhary CL, Alazab M. An effective feature engineering for DNN using hybrid PCA-GWO for intrusion detection in IoMT architecture. Comput Comm. 2020;16(11):139-49.
[66] Benzaid C, Lounis K, Al-Nemrat A, Badache N, Alazab M. Fast authentication in wireless sensor networks. Future Generat Comput Syst. 2016;55:362-75.
[67] Islam SR, Kwak D, Kabir MH, Hossain M, Kwak KS. The internet of things for health care: a comprehensive survey. IEEE Access. 2015;3:678-708.
[68] Dimitrov DV. Medical internet of things and big data in healthcare. J Healthc Inform Res. 2016;22(3):156-63.
[69] Alazab M, Tang M. Deep learning applications for cyber security. Switzerland: Springer; 2019.
[70] Azab A, Layton R, Alazab M, Oliver J. Mining malware to detect variants. 2014 Fifth Cybercrime and Trustworthy Computing Conference; 2014 Nov 24-25; Auckland, New Zealand. USA: IEEE; 2014. p. 44-53.
[71] Farivar F, Haghighi MS, Jolfaei A, Alazab M. Artificial intelligence for detection, estimation, and compensation of malicious attacks in nonlinear cyber-physical systems and industrial IoT. IEEE Trans Industr Inform. 2019;16(4):2716-25.