Journal of Biostatistics and Epidemiology

J Biostat Epidemiol. 2019;5(4): 235-245

Review Article

An Application-Based Review of Recent Advances of Data Mining in Healthcare

Saeed Shirazi [1,*], Hamed Baziyad [1], Hamed Karimi [2]

[1]Department of Information Technology, Faculty of Industrial and Systems Engineering, Tarbiat Modares University, Tehran, Iran.
[2]Department of Algorithms and Computation, Faculty of Engineering Science, School of Engineering, University of Tehran, Tehran, Iran.

ARTICLE INFO

Received 15.06.2019
Revised 16.07.2019
Accepted 24.08.2019
Published 03.12.2019

Key words:
Data mining;
Health care;
Learning paradigm;
Supervised;
Unsupervised;
Knowledge discovery

ABSTRACT

Background: Data mining as an integral part of the knowledge discovery in database (KDD) has gained significant attention over the past few years. By and large, data mining is the process of finding interesting structures in a considerably voluminous amount of data. Owing to its methods and algorithms supporting variable types of data, the data mining approach has been applied in many scientific areas, including the healthcare industry.
Regarding this matter, in this paper, we elaborate on the latest papers, including data mining techniques and algorithms in the healthcare field of research.
Results: We present a data mining review based on the newest researches. Afterward, we categorize data mining papers in healthcare based on supervised and unsupervised learning paradigms as well as classifying them in terms of their applications in the healthcare domain.
Conclusions: In every healthcare application, we propose some summary points of the papers. At last, we delve into the absence and hence, the necessity of existing some novel methods in healthcare domains in this researches.

Introduction

A data mining session exploits one or several algorithms to identify unusual patterns within data. The knowledge obtained from a data mining session is a generalized model of the data. The healthcare field of study is one of the essential application domains for data mining techniques. Generally, all the healthcare data are stored in electronic format by healthcare organizations. Primarily, healthcare data contains all the information due to patients as well as the other roles involved in healthcare industries. The amount of this type of data is increased at a very rapid rate. Due to increasing growth in the size of electronic healthcare data, a type of complexity exists in it. In other words, it is notable that healthcare data becomes very complicated. By using the traditional methods, it becomes arduous to fetch the meaningful information from it. However, regarding the advancement in the field of statistics, mathematics, and every other discipline, it is now feasible to extract meaningful patterns from it. Data mining is useful in such a situation where extensive collections of healthcare data are available. (1)

In this paper, we elaborate on novel data mining papers in the healthcare field. Considering the researcher's interest in data mining applications and methods in healthcare, there are many papers published in the past few years regarding this matter. However, there are some novel surveys concerning data mining fields and techniques with healthcare-related areas. For instance, Jothi

* Corresponding Author Email: s.shirazi@modares.ac.ir

Please cite this article in press as Shirazi S, Baziyad H, Karimi H. An Application-Based Review of Recent Advances of Data Mining in Healthcare. J Biostat Epidemiol. *2019; 5(4): 235-245*

et al. proposed a review of data mining in healthcare in the year 2015. (2)

Nevertheless, considering the growing attention regarding data mining techniques and algorithms in the healthcare field, researchers should be knowledgeable about the latest applications and methods in this matter. Moreover, an application-based paper for the usage of data mining applications in healthcare could be informative for further researches of scientists. Resultantly, for a proper concentration on educational issues of researches in this matter, we gathered 97 papers regarding data mining and healthcare. We propose an application-based review of data mining in healthcare with the concentration on 24 papers accepted or published after the year 2014 which are most relevant to the issue. The papers for investigation in this paper have been retrieved from google scholar using "data mining" and "healthcare" keywords together in different styles of writing.

In the following sections, firstly, we delve into data mining's meaning and basic concepts. Then, we elaborate on data mining categories and knowledge discovery stages. Afterward, we present considered papers regarding the field of healthcare in which they are applied. These fields consist of disease areas, public health, health insurance, and so forth.

Data mining an overview

Nowadays, data sizes are generally growing. The necessity of comprehending so much information, which are complicated enough and enriched data sets, has now increased in all the varied domains of science, technology, and business. The web-based systems, society, applications, businesses, and their related services, and networks in the field of science or engineering, among others, are increasingly producing data at a fast pace of growth because of the improvement of robust storage and connection tools. This extensive data growth does not entirely let useful information or structured knowledge be understood or extracted spontaneously, i.e., with this massive amount of data, the ability to extract useful knowledge latent in them and act on them is becoming incredibly important in today's competitive world. This action is called Data Mining (DM), the process of applying a computer-based information system (CBIS), including novel techniques, for discovering knowledge from data. (2)

Data mining is generally the process of finding significant structure in data. The structure may take many forms, including a set of rules, a tree, a graph or network, one or several equations, and so forth. The structure can be, for example, part of a sophisticated visual dashboard or as simple as a list of political candidates and an associated number representing voter sentiment based on Twitter feeds. (3)

Data mining finds its origins in the realms of statistics, mathematics, machine learning, artificial intelligence, and business. The term first appeared in the academic community around 1995. The phrase knowledge discovery in databases (KDD) was coined in 1989 to emphasize that knowledge can be derived from the data-driven discovery and is frequently used interchangeably with data mining.

In addition to performing data mining, a typical KDD process model includes a methodology to extract and preprocess data as well as making decisions about the actions of the data mining approach. When a particular application involves the analysis of large volumes of data stored in several locations, data extraction and preparation become the most time-consuming parts of the discovery process.

The final goal of using data mining is to apply what has been discovered in new situations. Several data mining techniques exist. Nevertheless, all data mining methods use induction-based learning. Learning based on induction is the process of forming general concept definitions by observing specific examples of concepts to be learned. (3)

The main purpose of both data science and KDD is to extract useful data knowledge. However, there are two apparent distinctions. The first one is that data science is usually combined with large volumes of data, whereas KDD has its origins in the database community. Secondly, data mining is a required stage within the KDD process model. Finally, along with data science, we have seen the advent of big data, cloud computing, and distributed data mining.

Data mining task categories

As discussed above, data mining is an advanced technology for analyzing big data sizes. Generally, data mining analytics can be classified

into two separate categories: supervised and unsupervised analytics. Supervised analytics, such as boosting and bootstrap aggregating, are powerful for predictive modeling. The knowledge representations of supervised analytics are regression or classification models, which describe the quantitative or qualitative relationships between input and output variables. The success of supervised analytics is dependent on two factors, i.e., domain expertise and training data. Domain expertise is crucial for developing functional models. It is especially crucial for specifying a model architecture, selecting model inputs, and tuning model parameters. However, the involvement of domain expertise will typically reduce the value of big data, as only a small subset of variables is used in model development. Training data refers to a set of observations where the input and output variables are both available. The quality of training data has a significant influence on the model's strength and reliability. It is worth mentioning that collecting high-quality training data can be costly, time-consuming, and sometimes even not possible in practice. (4)

On the contrary, unsupervised analytics focus on discovering the intrinsic structure, correlations, and associations in data. The success of implementing unsupervised analytics is not subject to the availability of training data, as there is no discrimination between inputs and outputs. The prominent advantage of unsupervised analytics is the ability to discover previously unknown knowledge. (4) Supervised analytics adopts a backward approach in data analysis, which means the mining target (e.g., the model output) is pre-defined. Unsupervised analytics adopts a forward approach in data analysis. All the data are taken as inputs, and the mining target is not explicitly defined. The ultimate goal is to reveal thriller relationships in data if any. In such a case, the value of big data can be best realized, and the knowledge discovered might be valuable for practical applications. Despite the strength of supervised learning in prediction, unsupervised learnings are more applicable and practical in finding new knowledge with respect to limited background knowledge.

Stages of knowledge discovery in database

Vast amounts of data in the world, i.e., raw data, are originally intractable for human or manual applications. Therefore, the analysis of such data is now a requirement. The process of this analysis of data is also prominent in exploiting the results achieved by extracted knowledge. One of the central and significant steps in following the process of analyzing raw data is the preprocessing step.

Data preprocessing includes data preparation, combined by integration, cleaning, normalization and transformation of data; and data reduction tasks; such as for instance-selection, feature selection, discretization, and so on. The result expected after reliable chaining of data preprocessing tasks is a final dataset, which can be considered valid and beneficial for further data mining algorithms. (5)

Mostly, data mining is to solve problems by analyzing data present in real databases. Nowadays, it is qualified as science and technology for exploring data to discover already present unknown patterns. Many people distinguish DM as a synonym of the KDD process, while others view DM as the primary step of KDD. By and large, the primary and essential steps of knowledge discovery considering data mining step as the most outstanding step are as follows: (5)

Problem Specification: Designating and arranging the application domain, the relevant background knowledge obtained by experts, and the final objectives pursued by the end-user.

Problem Understanding: Including the comprehension of both the selected data to approach and the expert knowledge associated with achieving a high degree of reliability.

Data Preprocessing: This part includes operations for data cleaning (such as stemming, lemmatization, and handling the removal of noise and inconsistent data like stop words), data integration (multiple data sources may be combined into one), data transformation (data is transformed and merged into forms which are appropriate for specific DM tasks) and data reduction, including the selection and extraction of both features and examples in a database.

Data Mining: It is an essential process where the methods are used to fetch valid patterns of data. This stage includes the choice of the most suitable DM task (such as clustering, classification,

regression or association), the choice of the DM algorithm itself belonging to one of the previous families, and ultimately, the embedding the selected algorithm to the problem, by tuning essential parameters and validation procedures.
Evaluation: Assessing and interpreting the extracted patterns based on particular measures.
Result Exploitation: The last step uses the knowledge directly, combining the knowledge with another system for more processes or merely reporting the extracted knowledge through visualization tools.

Review of data mining in healthcare

In the following, we elaborate on data mining techniques that have applied in healthcare fields in papers that have retrieved regarding this issue. We categorize these papers base on the fields of healthcare that they are applied to.

Data mining methods in healthcare

In this paper, we concentrate on researches applying data mining methods and algorithms in healthcare. Moreover, owing to a thorough understanding, also, we investigate survey researches. Not considering survey papers, other researches have applied both supervised unsupervised data mining methods with a tendency to supervised methods. About 56 percent of papers applied supervised methods, while about 44 percent of them used unsupervised methods. Besides, among data mining methods, classification and clustering methods were of noticeable attention. Besides, decision trees, regression, and feature selection were mostly used algorithms in these papers. The data mining tasks and the number of papers assigned to each task are shown in Table 1.

Table 1. Number of papers, assigned to related data mining tasks

Task	Count
Classification	7
Prediction	2
Clustering	2
Association	2
Time-series analysis	1

Disease areas

Among the researches we investigate, some papers concentrate on disease areas in healthcare. For instance, in their paper, Patel at al. proposed various techniques like classification, clustering, and association as some prominent and significant data mining approaches, and some related studies to assess, analyze, and predict human disease were illustrated. (6) For a proper understanding of tasks in this matter, we classify papers based on the diseases in the following sections.

Kidney disease

In their work, Vijayarani et al. predict kidney diseases (Acute Nephritic Syndrome, Chronic Kidney disease, Acute Renal Failure, Chronic Glomerulonephritis) using SVM and Naïve Bayes classification algorithms to find an efficient algorithm. (7) As a result of the experiments, the authors discussed that the SVM algorithm provides more accurate results than the Naïve Bayes classifier algorithm for this matter.

Heart disease

In this research, Shafique and Campus tried to discover significant patterns from data derived by heart patients. They used three algorithms with two different scenarios. These implemented algorithms are Naïve Bayes, Decision Tree, and Neural Network (8) Resultantly, based on their experiments, the Naïve Bayes classification algorithms had the highest accuracy among all other algorithms with the accuracy of 82.914%.

Diabetes

The fundamental nature of DM is combined with long-term complications and the various number of health disorders. In their research, Saravana et al. used the predictive analysis algorithm in the Hadoop environment via the Map-Reduce approach to predict the common types of diabetes, complications along with it, and the

type of appropriate treatment that was provided. (9) Based on the author's analysis, their system provides an efficient way to
cure and care for the patients with proper results in such terms as affordability and availability. Moreover, Kavakiotis et al. proposed a kind of systematic review of the machine learning applications, data mining techniques, and tools in the area of diabetes research regarding below: (10)

- Prediction and Diagnosis, which this category is the most popular one among the others
- Diabetic Complications
- Genetic background and Environment
- Health care and Management

Another research proposes an adaptive distributed data mining technique based on an ensemble strategy which is privacy-preserving and also, can perform well to ultimately obtain knowledge from numerous healthcare facilities without having access to the sensitive data of a patient. (11) Their proposed framework can prevent the "negative impact" during integration from multiple sources in comparison to other existing works. Moreover, In the case of classifying patients in diabetes, Perveen et al. proposed a method to assess the efficiency of some data mining techniques. (12)
The purpose is classifying patients with diabetes mellitus precisely; using factors of diabetes risk along with three various ordinal adults groups in CPCSSN:

- Young adults.
- Middle-aged adults.
- Adults are older than 55.

Then, it is to determine the best ensemble framework for the J48 decision tree that would help efficiently identify the diabetes patients and especially with high accuracy. As their analysis revealed, the results indicated that the AdaBoost ensemble method outperformed than bagging as well as a standalone J48 decision tree.

Malaria

Johansson et al. applied classification trees to estimate the relation between RDT results and antibiotic over-treatment using recursive partitioning based on the model, and they learned the impact of 38 other inputs as variables at facility-level, patient-level, and provider-level. (13) As their investigation revealed, integrated paediatric fever management was sub-optimal for completed assessments and antibiotic targeting despite standard compliance with malaria treatment guidelines.

Brain tumor and Acute ischemic stroke

The authors, in their paper, propose a fully optimized hybrid outstanding tumor feature selection algorithm combined with decision tree and bagging, in order to construct an interpretable set of simplified diagnosis rules for the tumor classification task. (14) Their experiments revealed that the proposed approach outperforms the conventional techniques applied in brain tumor classification problems to overcome the imbalanced features of medical data. Another research presented an approach in order to develop a stroke severity index (SSI) using administrative data. (15) they investigated seven predictive features and developed three models. Based on their results, the k-nearest neighbor model performed slightly better than the multiple linear regression model and the regression tree model.

Quality of life

The authors, in this paper, compared related work that develops some decision support and prediction systems in the clinical and medical area, with respect to studies in the realm of quality of life. (16) The authors in this paper provide explanations such as the relationship between the cost to gain information and the knowledge resulting from that measurement to explain the non-systematic evaluation of HRQoL in researches.

Public health

The current state of social media mining and its recent advances for health monitoring were summarized in a research paper. Then, the

authors focused on samples of promising research fields, technical challenges, and societal implications and considerations. After that, they provided a review of some significant researches in related areas. (17)

Health insurance

The field of data mining, which is partitioned into two separate and different learning techniques, i.e., supervised and unsupervised learning, was used for detecting fraudulent claims in another research. However, because each of the learning techniques above has its advantages and disadvantages, a new hybrid approach to identify fraudulent claims in the health insurance industry had presented via combining the benefits of both methods. (18) Considering the advantages and disadvantages of the classification and clustering techniques, ECM was chosen by the authors as the clustering method because the data flows in continuously, and the ability to cluster dynamic data and the SVM as classification method since it provides the scalability and usability.

Smart environment

In their research, Sprint et al. proposed an algorithm that can determine and detect treatment changes that are steady and firm with the medical literature for these cases. Their results recommend that the changes can be spontaneously detected using BCD. Their proposed smart home, change detection approach, and activity recognition algorithms are of beneficial data mining techniques to find out the behavioral influences of significant health conditions. (19)

Moreover, Yassine et al. presented frequent pattern mining, cluster analysis, and prediction in order to calculate and analyze energy usage fluctuations sparked by occupants' behavior. Since people's habits are approximately determined via routines of every day, finding these routines lets us discover abnormal activities that may demonstrate people's problems in taking care of themselves, in cases like not using a shower/bath or not preparing food. (20) They have demonstrated the applicability of their proposed model to accurately detect multiple appliance usage and make short and long term predictions at high accuracy.

Mobile health application

An exploratory study examined behavioral engagement with a weight loss app, Lose It!, and characterized higher versus lower engaged groups. (21) Their results demonstrate the importance of customization for behavioral engagement owing to the fact that most engaged subgroup, customized their recipes and exercises and utilized more custom features of the app. Furthermore, in a classification paper, the authors' purposes were to analyze financial health app data to determine successful weight loss subgroups by exploratory analyzes, and also, to verify the stability of the results. (22) This study demonstrated the probability of identifying distinct subgroups in "messy" commercial app data, and the identified subgroups to be replicated in the independent samples.

Healthcare systems and problems

In their paper, Lin et al. proposed a new method for satisfying the requirements of big medical data in precision medicine, which combines emotion computing and emerging communication technologies. (23) Their system effectively monitors and records the vital physical signal of users, as well as increasing the storage capacity and improving the security of stored healthcare data. The authors in another research concentrated on big data problems, their perspectives, and their solutions regarding the healthcare industry. (24) the quality of patient care at optimal cost. Their paper also investigated various analytical tools that researchers could apply to leverage benefits from the considerable set of healthcare data

Related reviews of data mining and healthcare

In their chapter, a brief overview of a vast range of different analytics and visualization components developed by Hu et al. had given as well as some examples of medical insights concluded by those components, and also, some new directions taken by the authors. (25) Jothi et al. illustrate evaluation approaches and achieved results of some persuasive papers in their research. Besides, a summary of the papers' findings is proposed as a conclusion of this paper. (2) Another research gives an overview of studies in which data mining techniques were used to detect health care fraud and abuse, using unsupervised and supervised data mining methods. (26) Another review paper discusses the field of big data, particularly medical big data and its applications, how they can be analyzed, and big medical data prevalent challenges. (27)

At last, In their paper, Rojas et al. proposed a literature survey on the performance of the mining process in the healthcare field of study. The study range of this review can cover about 74 papers, along with their case studies, and each of them was assessed and analyzed concerning eleven main aspects. (28) Figure 1 shows the healthcare fields and the number of papers assigned to each field. Furthermore, table 2 presents an overview of every individual paper investigated in this review paper.

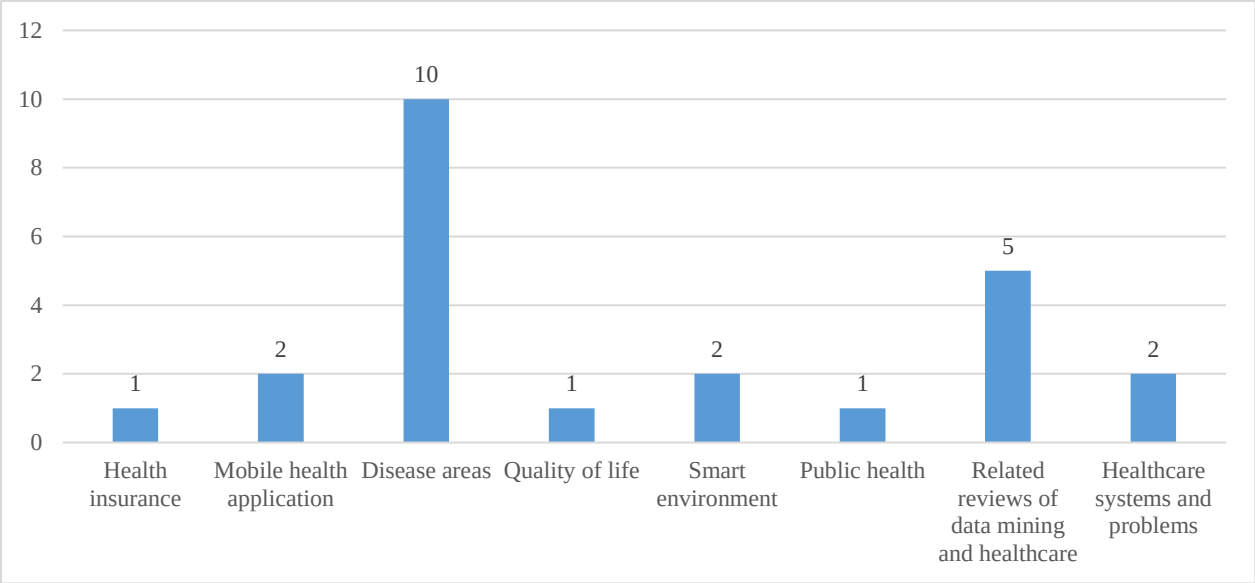

Figure 1 Number of papers, assigned to related healthcare fields

Table 2. Summarization of every paper investigated in this article

Type	Method	Journal/Conference	Year of publication	Keywords	Fields	About paper	Authors	Supervised/ Unsupervised	Reference
research	applying algorithm, SVM, Naïve Bayes	International Journal on Cybernetics & Informatics	2015	data mining, disease prediction, SVM, Naïve Bayes, Glomerular Filtration Rate (GFR)	kidney diseases	In this work, the authors predict kidney diseases (Acute Nephritic Syndrome, Chronic Kidney disease, Acute Renal Failure, Chronic Glomerulonephritis) using SVM and Naïve Bayes classification algorithms to find an efficient algorithm.	S. Vijayarani, S.Dhayanand	supervised	(7)
research	survey	International Journal of Information Sciences and Techniques	2016	data mining, health care, classification, clustering, association	healthcare, diseases	In this paper, various techniques like classification, clustering, and association are proposed as some prominent and significant data mining approaches, and also some related studies to assess, analyze and predict human disease are illustrated.	Sheenal Patel, Hardik Patel	survey	(6)
conference	survey	2015 10th Iberian Conference on Information Systems and Technologies (CISTI)	2015	quality of life, data mining, support systems decision	quality of life	The authors, in this paper, compared related work that develops some decision support and prediction systems in the clinical and medical area, with respect to studies in the realm of quality of life.	Joaquim Gonçalves, Luís Paulo Reis, Brígida Mónica Faria, Victor Carvalho, Álvaro Rocha	survey	(16)
chapter book	overview	Healthcare Information Management Systems	2016	data-driven healthcare analytics, learning health system, practice based evidence, real-world evidence, clinical decision support, machine learning,	healthcare	In this chapter, a brief overview of a vast range of different analytics and visualization components developed by the authors is given as well as some examples of medical insights concluded by those components, and also, some new directions taken by the authors.	Jianying Hu, Adam Perer, Fei Wang	overview	(25)

				data mining, data visualization					
Case study	applying algorithm, feature selection, decision tree	IEEE Access	2017	brain tumor, morphological features, ANNIGMA, MRMR, feature selection, classification	brain tumor	The authors in their paper, propose a fully optimized hybrid outstanding tumor feature selection algorithm combined with decision tree and bagging, in order to construct an interpretable set of simplified diagnosis rules for the tumor classification task.	Shamsul Huda, John Yearwood, Herbert F. Jelinek, Mohammad Mehedi Hassan, Giancarlo Fortino, Michael Buckland	supervised	(14)
research	applying algorithm, decision tree, Weka, neural network, Naïve Bayes	International Journal of Innovation and Applied Studies	2015	heart diseases, decision tree, Weka, neural network, Naïve Bayes	heart diseases	In this research, the authors tried to discover significant patterns from data derived by heart patients. They used three algorithms with two different scenarios. These implemented algorithms are Naïve Bayes, Decision Tree, Neural Network.	Umair Shafique, Fiaz Majeed, Haseeb Qaiser, Irfan Ul Mustafa	supervised	(8)
research	classification	Malaria journal	2016	antibiotic resistance, IMCI, malaria, diagnosis, child health, fever case management	fever, malaria	In this research, classification trees estimated the relation between RDT results and antibiotic over-treatment using recursive partitioning based on the model, and also, they learned the impact of 38 other inputs as variables at facility-level, patient-level, and provider-level.	Emily White Johansson, Katarina Ekholm Selling, Humphreys Nsona, Bonnie Mappin, Peter W. Gething, Max Petzold, Stefan Swartling Peterson, Helena Hildenwall	supervised	(13)
conference	review	The Third Information Systems International Conference	2015	data mining, data mining in healthcare, health informatics	healthcare	In this review paper, evaluation approaches and achieved results of some impressive papers are illustrated. In addition, a summary of the papers' findings is proposed as a conclusion of this paper.	Neesha Jothi, Nur'Aini Abdul Rashid, Wahidah Husain	review	(2)
research	Review of fraud and abuse	Global Journal of Health Science	2015	health care, data mining, KDD, business intelligence, insurance claim, fraud	healthcare	This research gives an overview of studies in which data mining techniques were used to detect health care fraud and abuse, using unsupervised and supervised data mining methods.	Hossein Joudaki, Arash Rashidian, Behrouz Minaei-Bidgoli, Mahmood Mahmoodi, Bijan Geraili, Mahdi Nasiri, Mohammad Arab	review	(26)
research	a systematic review of the applications of machine learning, data mining techniques, and tools in the field of diabetes research	Computational and Structural Biotechnology Journal	2016	machine learning, data mining, diabetes mellitus, diabetic complications, disease prediction models, biomarker(s) identification	diabetes	This paper is a kind of systematic review of the machine learning applications, data mining techniques, and tools in the area of diabetes research regarding: a) Prediction and Diagnosis, which this category is the most popular one among the others, b) Diabetic Complications, c) Genetic background and Environment e) Health care and Management	Ioannis Kavakiotis, Olga Tsave, Athanasios Salifoglou, Nicos Maglaveras, Ioannis Vlahavas, Ioanna Chouvarda	review	(10)
research	applying algorithm, predictive analysis algorithm	2nd International Symposium on Big Data and Cloud Computing (ISBCC'15)	2015	healthcare industry, Hadoop/Map Reduce, big data, predictive analysis	diabetes	The fundamental nature of DM is combined with long term complications and various number of health disorders. In this research, the authors used the predictive analysis algorithm in Hadoop environment via the Map-Reduce approach to predict the common types of diabetes, complications along with it and the type of provided appropriate treatment.	Saravana kumar N M, Eswari T, Sampath P, Lavanya S	unsupervised	(9)
research	review	Kidney research and medical practice	2017	big data, epidemiology, data mining, healthcare, statistics	healthcare	This review paper discusses the field of big data, particularly medical big data and its applications, how they can be analyzed, and big medical data prevalent challenges.	Choong Ho Lee, Hyung-Jin Yoon	review	(27)
research	proposing method	Information Sciences	2015	privacy-preserving, data mining,	diabetes	This research proposes an adaptive distributed data mining technique based on an ensemble	Yan Li, Changxin Bai, Chandan K. Reddy	unsupervised	(11)

				ensemble learning, electronic health record boosting, machine learning, healthcare		strategy which is privacy-preserving and also, can perform well in completely obtain knowledge from numerous healthcare facilities without having access to the sensitive data of a patient.			
research	building a new system	IEEE Acces	2016	emotion computing, healthcare big data; 5G, cloud computing, software-defined networking	healthcare	This paper proposes a new method for satisfying the requirements of big medical data in precision medicine, which combine the emotion computing and emerging communication technologies together.	Kai Lin, Fuzhen Xia, Wenjian Wang, Daxin Tian, Jeunjeun Song	unsupervised	(23)
conference	big data problems	2nd International Conference on Innovations in Information Embedded and Communication Systems	2015	analytics, big data, cloud, data management, healthcare, open source	healthcare	The authors in this research concentrate on big data problems, their perspectives, and their solutions regarding the health care industry.	Prabha Susy Mathew, Anitha S. Pillai	unsupervised	(24)
conference	review	Conference: Proceedings of the Pacific Symposium	2016	social media, data mining, natural language processing, public health	public health	In this paper, the current state of social media mining and its recent advances for health monitoring were summarized. Then, the authors focused on samples of promising research fields, technical challenges, and societal implications and considerations. After that, they provided a review of some significant researches.	Michael J. Paul, Abeed Sarker, John S. Brownstein, Azadeh Nikfarjam, Matthew Scotch, Karen L. Smith, Graciela Gonzalez	review	(17)
research	classification	Journal of Medical Internet research	2016	weight loss, mobile health, mobile app, data mining, classification	mobile app	In This classification paper, the authors' purposes were to analyze financial health app data to determine successful weight loss subgroups by exploratory analyzes, and also, to verify the stability of the results.	Katrina J Serrano, Mandi Yu, Kisha I Coa, Linda M Collins, Audie A Atienza	review	(22)
research	applying algorithm, Ensemble method; base learner; bagging; Adaboost and decision tree	Procedia Computer Science	2016	diabetes Mellitus, ensemble method, base learner, bagging, Adaboost and decision tree	diabetes	In this study, the purpose is to assess the efficiency of some data mining techniques to classify patients with diabetes mellitus precisely, using factors of diabetes risk along with three various ordinal adults groups in CPCSSN: (i) Young adults. (ii) Middle-aged adults. (iii) Adults are older than 55. Then, it is to determine the best ensemble framework for J48 decision tree that would help efficiently identify the diabetes patients and especially, with high accuracy.	Sajida Perveena, Muhammad Shahbaza, Aziz Guergachib, Karim Keshavjeec	supervised	(12)
conference	fraud detection	International Conference on Communication , Information & Computing Technology	2015	data mining, health insurance fraud, supervised, unsupervised	health insurance	In this paper, the field of data mining which is partitioned into two separate and different learning techniques, i.e., supervised and unsupervised learning is used for detecting fraudulent claims. However, owing to the fact that each of the aforementioned learning techniques has its own advantages and disadvantages, a new hybrid approach to identify fraudulent claims in the health insurance industry is presented via combining the benefits of both methods.	Vipula Rawte, G Anuradha	supervised-unsupervised	(18)
research	literature review	Journal of Biomedical Informatics	2016	healthcare processes, process mining, case studies, literature review	healthcare	In this paper, the authors proposed a literature survey on the performance of the mining process in healthcare field of study. The study range of this review can cover about 74 papers along with their case studies, and	Eric Rojas, Jorge Munoz-Gama, Marcos Sepúlveda, Daniel Capurro	review	(28)

						each of them was assessed and analyzed with respect to eleven main aspects.			
research	Validation methods	Transactional, behavioral medicine	2017	mobile health application, mobile health technology, smartphone app, data mining, big data, user engagement, classification, regression tree	mobile health application	This exploratory study examined behavioral engagement with a weight loss app, Lose It! and characterized higher versus lower engaged groups.	Katrina J. Serrano, Kisha I. Coa, Mandi Yu, Dana L.Wolff-Hughes, Audie A. Atienza	supervised	(21)
conference	change detection approach	2016 IEEE International Conference on Smart Computing	2016	pervasive computing, machine learning, smart environments, time series analysis	smart environment	In this research, the authors' algorithm can determine and detect treatment changes that are steady and firm with the medical literature for these cases. Their results recommend that the changes can be spontaneously detected using BCD. Their proposed smart home, change detection approach, and activity recognition algorithms are of beneficial data mining techniques to find out the behavioral influences of major health conditions.	Gina Sprint, Diane Cook, Roschelle Fritz, Maureen Schmitter-Edgecombe	unsupervised	(19)
research	Regression, knn, feature selection	Journal of Clinical Epidemiology	2015	Acute ischemic stroke, disease severity, administrative data, data mining, prediction model, outcomes research	acute ischemic stroke,	This research presented an approach in order to develop a stroke severity index (SSI) using administrative data.	Sheng-Feng Sung, Cheng-Yang Hsieh, Yea-Huei Kao Yang, BPharm, Huey-Juan Lin, Chih-Hung Chen, Yu-Wei Chen, Ya-Han Hu	supervised	(15)
research	frequent pattern	IEEE Access	2017	Big data, smart cities, smart homes, smart cities, health care applications, Behavioral Analytics, Frequent Pattern, Cluster Analysis, Incremental Data-Mining, Association Rules, Prediction	smart environment	In this paper, We propose using frequent pattern mining, cluster analysis, and prediction are presented in order to calculate and analyze energy usage fluctuations sparked by occupants' behavior. Due to the fact that people's habits are approximately determined via routines of every day, finding these routines lets us discover abnormal activities that may demonstrate people's problems in taking care of themselves, in cases like not using shower/bath or not preparing food.	Abdulsalam Yassine, Shailendra Singh, Atif Alamri	unsupervised	(20)

Discussion

In this paper, we presented an application-based data mining review of healthcare. At first, we classified paper based on supervised and unsupervised data mining techniques that the researchers apply in their papers. However, the lack of semi-supervised methods in this researches is noticeable. Semi-supervised learning concentrates on both labeled and unlabeled data and how computers and natural systems can study these two types of data to learn. Traditionally, learning has been restricted by either unsupervised methods such as clustering and outlier detection on unlabeled data or supervised learning consisted of classification and regression that are applied on labeled data. Semi-supervised learning tries to take advantage of both labeled and unlabeled data to evaluate if this could change the learning process. (29) Resultantly, Semi-supervised learning proposes algorithms to benefit both types of data. However, based on our study, this paradigm has not been used in any healthcare-related research. Furthermore, we investigated healthcare applications, in which the papers above have focused. These applications consisted of areas of diseases, smart environments, and so forth. Nevertheless, some crucial fields of healthcare are missing in these papers. Among these, the areas referring to patients, physicians, and the relation between these two could be an exciting

area owing to their importance in the healthcare field.

Conclusion

In this paper, we have delved into the latest papers concerning data mining methods and applications in the healthcare field. First, we classified 24 papers regarding this matter based on supervised and unsupervised learning paradigms. Then, papers classified based on the healthcare field that they were focused on. These fields mostly consisted of healthcare diseases, smart environments, and mobile health applications. At last, based on our knowledge, we elaborated on methods and fields missing in case of the latest researches regarding this issue. In case of further researches, regarding the growing number of research papers in fields of data mining and healthcare, application-based researches could be done focusing on merely one or two specific healthcare fields, and data mining approaches applied in those over long periods.

References

1.P. Ahmad, "Techniques of Data Mining In Healthcare : A Review," vol. 120, no. 15, pp. 38–50, 2015.
2.N. Jothi, N. Aini, A. Rashid, and W. Husain, "Data Mining in Healthcare – A Review," *Procedia - Procedia Comput. Sci.*, vol. 72, pp. 306–313, 2015.
3.R. J. Roiger, *Data Mining A Tutorial-Based Primer (Second Edition)*, Second. Taylor & Francis Group, 2016.
4.C. Fan, F. Xiao, Z. Li, and J. Wang, "Unsupervised data analytics in mining big building operational data for energy efficiency enhancement : A review," *Energy Build.*, vol. 159, pp. 296–308, 2018.
5.S. García, *Intelligent Systems Reference Library 72 Data Preprocessing in Data Mining*. .
6.S. Patel and H. Patel, "SURVEY OF DATA MINING TECHNIQUES USED IN HEALTHCARE DOMAIN Sheenal," *Int. J. Inf. Sci. Tech.*, vol. 6, no. 1, pp. 53–60, 2016.
7.D. S. Vijayarani S, "Data Mining Classification Algorithms for Kidney Disease Prediction," *Int. J. Cybern. Informatics*, vol. 4, no. 4, pp. 13–25, 2015.
8.U. Shafique and L. Campus, "Data Mining in Healthcare for Heart Diseases Data Mining in Healthcare for Heart Diseases," no. March, 2015.
9.N. M. Saravana, T. Eswari, P. Sampath, and S. Lavanya, "Predictive Methodology for Diabetic Data Analysis in Big Data," *Procedia - Procedia Comput. Sci.*, vol. 50, pp. 203–208, 2015.
10.I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine Learning and Data Mining," *Comput. Struct. Biotechnol. J.*, 2016.
11.Y. Li, C. Bai, and C. K. Reddy, "A distributed ensemble approach for mining healthcare data under privacy constraints," *Inf. Sci. (Ny).*, 2015.
12.S. Perveen, M. Shahbaz, A. Guergachi, and K. Keshavjee, "Performance Analysis of Data Mining Classification Techniques to Predict Diabetes," *Procedia - Procedia Comput. Sci.*, vol. 82, no. March, pp. 115–121, 2016.
13.E. W. Johansson *et al.*, "Integrated paediatric fever management and antibiotic over - treatment in Malawi health facilities : data mining a national facility census," *Malar. J.*, pp. 1–12, 2016.
14.S. Huda, J. Yearwood, H. F. Jelinek, M. M. Hassan, and M. Buckland, "A hybrid feature selection with ensemble classification for imbalanced healthcare data : A case study for brain tumor diagnosis," vol. 3536, no. c, pp. 1–13, 2016.
15.S. Sung *et al.*, "Developing a stroke severity index based on administrative data was feasible using data mining techniques," *J. Clin. Epidemiol.*, 2015.
16.J. Gonçalves, L. P. Reis, and V. Carvalho, "Data Mining and Electronic Devices applied to Quality of Life Related to Health Data."
17.M. Scotch, K. L. Smith, and G. Gonzalez, "Social media mining for public health monitoring and surveillance," pp. 468–479, 2016.
18.V. Rawte, "Fraud Detection in Health Insurance using Data Mining Techniques," 2015.
19.G. Sprint, D. Cook, R. Fritz, and M. Schmitter-edgecombe, "Detecting Health and Behavior Change by Analyzing Smart Home Sensor Data," pp. 1–3, 2016.
20.A. Yassine, S. Singh, and A. A. Member, "Mining Human Activity Patterns from Smart Home Big Data for Healthcare Applications," vol. 3536, no. c, pp. 1–10, 2017.
21.K. J. Serrano, K. I. Coa, M. Yu, D. L. Wolff-hughes, and A. A. Atienza, "Characterizing user engagement with health app data: a data mining approach," *Transl. Behav. Med.*, pp. 277–285, 2015.
22.K. J. Serrano, M. Yu, K. I. Coa, L. M. Collins, and A. A. Atienza, "Mining Health App Data to Find More and Less Successful Weight Loss Subgroups," *J. Med. Internet Res.*, vol. 18, pp. 1–11, 2016.
23.K. Lin, F. Xia, W. Wang, D. Tian, and J. Song, "System Design for Big Data Application in Emotion-aware Healthcare," *IEEE Access*, vol. 3536, no. c, pp. 1–9, 2016.
24.P. S. Mathew, "Big Data Solutions in Healthcare : Problems and Perspectives," 2015.
25.J. Hu, A. Perer, and F. Wang, "Data Driven Analytics for Personalized Healthcare."
26.H. Joudaki, A. Rashidian, B. Minaei-bidgoli, M. Mahmoodi, and B. Geraili, "Using Data Mining to Detect Health Care Fraud and Abuse : A Review of Literature," vol. 7, no. 1, pp. 194–202, 2015.
27.C. H. Lee and H. Yoon, "Medical big data : promise and challenges," vol. 2017, no. 1, pp. 3–11, 2017.
28.E. Rojas, J. Munoz-gama, M. Sepúlveda, and D. Capurro, "Process mining in healthcare : A literature review," *J. Biomed. Inform.*, vol. 61, pp. 224–236, 2016.
29.A. B. G. Xiaojin Zhu, *Introduction to Semi-Supervised Learning*, First. Morgan & Claypool Publishers, 2009.