Data Mining
in
Healthcare

Table of Contents

ABSTRACT 3
1. Introduction 4
2. Background 4
3. Data Mining Process 5
4. Data Mining Techniques 6
4.1 Clustering 6
4.1.1 Partitioned Clustering 7
4.1.2 Hierarchical Clustering 7
4.1.3 Density Based Clustering 7
4.2 Association 8
4.3 Classification and Regression 8
5. Applications 8
5.1 Effective Management 9
5.2 Hospital Ranking 9
5.3 Improved Treatments and Patient care 9
5.4 Customer Relationship Management 10
5.5 Medical Device and Pharmaceutical Industries 10
5.6 Fraud and abuse detection 11
6. Challenges 11
6.1 Interdisciplinary Nature of Data 11
6.2 Social and Security Issues 12
6.3 Integration of Mining Methodology 12
6.4 Complex Data and Performance issues 12
7. Future Directions 13
7.1 Personalized Healthcare 13
7.2 Research Purpose Data Collection 13
7.3 Automation of Data Mining process 13
7.4 Standardization of clinical vocabulary 14
7.5 Improved Data Sharing 14
7.6 Integrating Mining in Healthcare system 14
8. Conclusion 15
9. References 16
10. Appendix 20

Abstract

Big data analytics in healthcare is turning heads all over the world as digitization of health information is generating enormous amounts of data in the form of patient records, billing systems, medical claims, etc. The demand for value-based patient care has thus increased the popularity of data mining tools and techniques to manage and provide better medical services and treatments to patients. Data mining in healthcare has shown incredible breakthroughs in improving patient outcomes and safety, detecting health insurance fraud, reducing healthcare costs and many more. In this paper, we present a brief introduction of data mining process and explores various data mining tools and techniques used in healthcare, its applications and related works. It concludes by highlighting the benefits, challenges and the future directions of data mining in the healthcare industry. This paper elucidates the data mining techniques that can process huge volume of clinical data available today which will help to improve patient care, reduce medical costs and aid many other decision-making processes in the healthcare sector.

1. Introduction

Healthcare industry is one of the largest service sectors in the current technologically driven world. The explosion of clinical information from different sources in the form of electronic health records, insurance claims, laboratory results, medications, etc. are overwhelming the healthcare sector. Moreover, the COVID-19 pandemic has brought all nations globally to a situation where there is a pressing demand for effective, safe and efficient healthcare services for patients. To solve such healthcare challenges, the complexity and volume of data has become a barrier for traditional methods to be efficient. With the advent of data mining techniques, it is now possible to look through the generated large amounts of medical data from different sources and derive valuable hidden insights. As big data is gathering momentum in the healthcare field, hospitals and facilities are now taking interest in using data mining techniques to make sense of raw data by exploring and finding relationships and patterns and predicting outcomes out of healthcare data to stay ahead of the competition. This technique can thus be used to solve many healthcare challenges, provide better and affordable patient care thereby improving patient satisfaction and detect healthcare fraud by tracking down unfamiliar patterns in the data.

2. Background

Data Mining came into existence in the middle of 1990s. Generally, the term “Mining” is the process of extraction of something valuable from the earth’s surface like coal, diamond, gold, etc. In that context, Data mining refers to the process of extraction of useful and hidden information from bulk data that could potentially drive business

decisions. However, while coal or diamond mining results in the extraction of the valuable metals respectively, the result of the extraction process in Data mining is not data instead it is the patterns and knowledge that we gain from data. That is why Data Mining is also known as Knowledge Discovery or Knowledge Extraction. Today, Data mining has evolved into a very powerful tool with the capability to reshape any industry due to its endless possibilities to extract previously unknown trends from data. Several research works have been conducted on healthcare data to solve pressing challenges by applying data mining techniques. For example, Moore in 2015 with the help of data mining techniques introduced a model based on information from the mRNA analysis to predict a patient's risk of cancer. Milioli et al. using data mining in 2016 was able to classify breast cancer through a genetic data bank which paved a clear way to understand more about cancer diagnosis [2]. Paydar et al. in 2016, data mining techniques developed a model to predict malignancy risk for Thyroid nodules using ANN [7].

3. Data Mining Process

Knowledge Discovery Process is an iterative, structured and interactive procedure used for identifying useful and potential business decision-driving patterns from large and complex datasets. This model gives an overview on the techniques that can be used to extract useful knowledge from historical data, analyze and then predict future outcomes from it. Data Mining is the main part of the Knowledge Discovery Process which is composed of various stages as shown in Figure 3.1 in the appendix. The first stage is data selection where for example, health records data is collected from different sources. The second stage is preprocessing of the selected data where noise and irrelevant

information is removed and data from multiple sources are combined. Data cleaning and data integration are the vital processes in the data preprocessing stage. This stage is followed by transformation of the data into appropriate format accepted by the data mining block. The fourth and most important stage is Data Mining where suitable data mining techniques are applied on the transformed data for extracting valuable information and finally the evaluation stage where the performance of the model is interpreted, and the identified patterns are presented to the user using various knowledge representation methods like data visualization tools.

4. Data Mining Techniques

Data Mining techniques can be mainly classified into descriptive and predictive categories. The goal of the descriptive task is to analyze historical data to learn about what is happening in a business whereas the predictive task uses the past and present data to forecast and create models thereby allowing the businesses to make predictions about the future. The descriptive approach comprises of clustering and association models whereas the predictive approach is mainly based on the classification algorithm, which is further broadly divided into many classes like neural networks, decision trees, kNN, etc.

4.1. Clustering

The clustering methodology can be considered as one of the most important unsupervised learning tools with the goal of finding a pattern in a collection of data which is unlabeled. This means grouping similar data points together in the same cluster and different data points to their respective clusters based on similar features. Table 4.1

shows different clustering types, their pros and cons and related work in the healthcare field.

4.1.1. Partitioned Clustering

In Partitioned clustering, datapoints are grouped into different groups based on their similarity. However, the algorithm needs to know the number of clusters required in advance. It is further divided into k-means clustering technique and k-medoids. A related work that was done using k-means clustering approach is grouping the patients into two risk categories (high risk and low risk of having a heart disease) based on the measures of blood pressure and cholesterol levels.

4.1.2. Hierarchical Clustering

In hierarchical clustering, a hierarchy of clusters is built using the top-down (divisive method) or bottom-up (agglomerative method) approach. Here, an advantage over k means clustering is that the algorithm does not require to define the number of clusters in advance. A related work has come up with a way using hierarchical clustering to group patients based on their length of stay in the hospital which enhances the capability of hospital resource management.

4.1.3. Density Based Clustering

The density-based clustering covers the disadvantages shown by the other two types. It is the most widely used clustering technique as it can even find a cluster completely formed inside a different cluster and can handle outliers. A related work would be clustering of skin wound image using DBSCAN.

4.2. Association

Association process observes frequently appearing patterns and correlations from datasets. It is also known as market-basket analysis. For example, if a customer is buying a sanitizer, then the chance of him/her buying a mask is high. This information helps the vendor to further incorporate products and enhance their sales. Also, association has a high significance in the healthcare field where the relationships between patient symptoms, diseases and health conditions can be tracked. For example, the Utah Bureau of Medicaid used association rules to track down fraud in the doctors' prescription and treatment methods. Thus, this approach is found to be highly effective to identify fraud in medical claims made by the respective authority.

4.3. Classification and Regression

They are two of the most extensively used supervised learning methods of Data mining in the healthcare sector. It is a predictive method that is used to assign a label to an unknown datapoint. The target variable is predicted by creating a model from the independent features. These independent features are discrete in classification tasks and continuous in regression analysis. Table 4.3 shows different classification types, their pros and cons and related work in the healthcare field.

5. Applications

Data mining in healthcare has varied uses and is practiced in several domains of healthcare industry. The more the improvement and inventions in data mining techniques the better expansion of its applications in healthcare sector are observed. We would like to discuss some of the prominent applications of data mining in healthcare sector.

5.1. Effective Management

Data Mining tools help in building models that would support decision making process in utilizing both human and technical resources related to hospitals and patients. The techniques involved in data mining help in identifying the disease along with its severity which further gives an idea to prioritize and allocate the required equipment's and treatment to the patients in a quick and efficient manner. Blue cross has developed a system using data mining algorithms to provide service at minimal cost and improve results of disease treatment.

5.2. Hospital Ranking

Determining the rank of a hospital in a particular disease treatment or service is an important task as public usually refer to such resources for primary care consultation. Data mining tools are used in predicting the rank of the hospitals and this is done with the help of data related to the ability of treating high risk patients by a hospital. By using the nineth revision of international classification of diseases codes, rebuilding the details of patients and applying mining techniques will give an idea of how the risk factors for ranking of hospitals are calculated based on the success rates and mortality rates of patients.

5.3. Improved Treatments and Patient care

Data mining algorithms and tools enable the process of comparing two or more treatments provided for a same disease or medical condition and enable to examine the results such as which treatment performed better at condition and whatever the side effects related to other treatments. This helps in providing treatments precisely to the condition of patient. Huge amount data is available and collected as electronic health

records which when analyzed using data mining tools and techniques by healthcare services detects the needs or requirements of patients at presents and in the future which increases patient's satisfaction. Authors Hallick, Milley and Kolar have found and mentions that data mining techniques are helpful providing information to the patients for protection and prevention from diseases.

5.4. Customer Relationship Management

The techniques used in data mining enable capturing of patterns, specifications and requirements of the patients which can further help in division of customers into different categories based on common characteristics. Further, the health care services can be provided specific to each category of customers which would result in customer satisfaction and their retention. There are examples of data mining techniques where healthcare providers develop a specific index value that would track how customers are reacting to a specific service and evaluate its effect. Customer potential management corporation has created an index that indicated the usage of healthcare service and further this was utilized by OSF Saint Joseph Medical Centre which resulted in better communications and increase in revenue of the medical center. Rafalski explained how Sinai heath model used data mining for customer relationship management.

5.5. Medical Device and Pharmaceutical Industries

Development in medical device industry plays a crucial role in healthcare for providing and tracking vital signs of the patients. Data mining algorithms enable real time analysis of data streaming from bio sensors and medical gadgets, transferring the the reports on to wireless devices like mobile, laptops based on resources like battery and memory space. Data mining also supports pharmaceutical industries in controlling the inventory

level of the medicines and decision making by using algorithms to check required amount of stock availability. A review by Hadi, Mobyen, and Amy Loutfi helps us understand how data mining is used on data from sensors to monitor the health condition.

5.6. Fraud and abuse detection

Data mining techniques help in determining the fraudulent or irregular patterns observed in insurance claims or prescriptions provided by physicians. This is done by setting up norms for proper transactions (training model) and if any abnormal behavior is observed then it would be detected by the mining algorithm. Health care insurers, hospitals have developed data mining models that would determine the irregular activities for example, Texas and Utah Medicaid Fraud and Abuse systems saved millions of dollars annually from fraud and abuse medical claims. Australian Health Insurance also used mining techniques on huge data and reported millions of dollars.

6. Challenges

Healthcare data comes with many challenges beginning from collecting the data from multiple platforms to application of right mining technique to the data. Although the data mining techniques have established their importance in applications of healthcare it has to overcome the following issues for a practical application.

6.1. Interdisciplinary Nature of Data

Healthcare data is acquired from different systems like hospitals, insurance companies, health reports, doctors handwritten prescriptions. All these systems might have specific terminology within their field which is hard to understand for people coming from other departments. Researchers trying to apply mining techniques usually find the data tough

to preprocess as they are expected to select information(data) that is useful for their experimentation from the pool of raw data context which would require domain expert knowledge.

6.2. Social and Security Issues

While the data is being utilized for creating better healthcare solutions some patients could have a privacy concern in using their medical records. Many patients have an outlook that the data mining process would fail, and their data would be misused by other people. As the approach might involve sharing of data with experts or panel of researchers for decision making purposes the collection of customers information for determining patterns and behavior can be considered exceeding confidentiality and becoming an important issue.

6.3. Integration of Mining Methodology

In case the healthcare data collected is a mix of both structured and unstructured data i.e., contains all forms of textual, numerical, images data from laboratory reports then it would be a hard task to apply a common technique to process all the data. It involves multiple steps of cleaning, processing and transformation of data before considering as an input to the model. Diverse data with high dimensionality might require usage of multiple techniques together or one after other to obtain results which would be a challenging process.

6.4. Complex Data and Performance issues

As discussed, in above challenge the real-world data could be diverse (heterogeneous) in nature and is collected from various platforms in the form of text, spatial data, audio, video, reports, objects, graphical and mining all this data on one machine at the same

time is not practical. The efficiency of data mining process is observed based on its performance and ability to acquire information from large datasets leading to challenges in features like size and scalability of the software being used. Therefore, it leads to the importance of adopting parallel, distributed techniques and big data solutions.

7. Future Directions

There is a tremendous potential in using data mining for healthcare. However, the data being used should be clean and easy to understand for various researchers to create new approaches and solutions, so we shall also discuss the future aspects that should be taken into consideration for better application of data mining.

7.1. Personalized Healthcare

Proper maintenance and use of electronic health records can be beneficial in developing personalized healthcare services to improve patient experience as they collect both clinical and demographic information of the users.

7.2. Research Purpose Data Collection

Usually, the healthcare data is collected for understanding a medical condition or for implementation of better healthcare service. If there are specific objectives and requirements set prior to the collection of data then it would be easy to acquire structured data as per the research requirement avoiding complex data with missing values, outliers and noise.

7.3. Automation of Data Mining process

Heath care professionals are mostly the end users of applications that are developed using data mining who might have limited knowledge in familiar with handling the

analytical systems. Therefore, it is recommended to automate the data mining process so that it wouldn't require any human intervention but in practical the automation would be challenging as the same processes cannot be used for all the applications and data.

7.4. Standardization of clinical vocabulary

To avoid the issues related to terminology of healthcare data few authors have proposed to have a meta data and data dictionary of data required for data mining process which also improves the quality of healthcare data. Further data mining could have an advantage from data visualization but there should be more empirical studies to study the interpretation techniques of healthcare data patterns.

7.5. Improved Data Sharing

To overcome the data privacy, legal issues the health care service providers should collect data about preferences of their customers if they agree or disagree to share their data. The health care providers should also create awareness among their customers and public that their confidential or personal data like (SSN) would be encrypted and will not be accessible by people who are not related to healthcare sector.

7.6. Integrating Mining in Healthcare system

There were only few articles from review that explained how the datamining process would be integrated into the health care system for decision making. So, it is unclear to understand the effect of knowledge discovery process (KDD) using data mining on the real-world decision framework is time consuming or leads to increase in workload. Hence more studies related to the integration of entire system and its impact on the working environment would be interesting to understand.

8. Conclusion

This paper elucidates an overview of data mining techniques and applications in healthcare. It also cites related research works done in this field and the challenges to apply data mining techniques in the medical industry. The paper starts with the background, definition and a review on data mining process. It further surveys the various data mining techniques in healthcare and then compare the differences between each. Finally, highlights the applications and the challenges the industry is facing along with the future directions. All along, related works are cited and specific application areas of data mining in healthcare are illustrated.

9. References

1. Sohail M.N., Jiadong R., Uba M.M., Irshad M. (**2019**) A Comprehensive Looks at Data Mining Techniques Contributing to Medical Data Growth: A Survey of Researcher Reviews. In: Patnaik S., Jain V. (eds) Recent Developments in Intelligent Computing, Communication and Devices. Advances in Intelligent Systems and Computing, vol 752. Springer, Singapore. Link
2. Pooja H, Dr. Prabhudev Jagadeesh M P, A Collective Study of Data Mining Techniques for the Big Health Data available from the Electronic Health Records, **2019** IEEE.
3. Malik, M.M., Abdallah, S. & Ala'raj, M. Data mining and predictive analytics applications for the delivery of healthcare services: a systematic literature review. *Ann Oper Res* **270,** 287–312 (**2018**). Link
4. A Systematic Review on Healthcare Analytics: Application and Theoretical Perspective of Data Mining Md Saiful Islam 1, Md Mahmudul Hasan 1, Xiaoyi Wang 1, Hayley D. Germack 1,2,3 and Md Noor-E-Alam 1, **2018** Link
5. Ogundele I.O et al., "A Review on Data Mining in Healthcare", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 7, Issue 9, September **2018**
6. Hiller, J.S. Healthy Predictions? Questions for Data Analytics in Health Care. Am. Bus. Law J. 2016, 53, 251–314. Healthcare 2018, 6, 54 43 of 43 159.
7. D.Usha Rani, "A survey on Data Mining Tools and Techniques in Medical Field", International Journal of Advanced Networking & Applications (IJANA), Volume: 08, Issue: 05 Pages: 51-54 (**2017**) Special Issue, TECHSA-17

8. Estape, E.S.; Mays, M.H.; Sternke, E.A. Translation in Data Mining to Advance Personalized Medicine for Health Equity. Intell. Inf. Manag. **2016**, 8, 9–16. Link
9. Milioli HH, Vimieiro R, Tishchenko I, Riveros C, Berretta R, Moscato P. Iteratively refining breast cancer intrinsic subtypes in the METABRIC dataset. BioData Min. **2016**; 9:2
10. Sheenal Patel and Hardik Patel, "Survey of Data Mining techniques used in healthcare domain", International Journal of Information Sciences and Techniques (IJIST) Vol.6, No.1/2, March **2016**
11. Karimi, S.; Wang, C.; Metke-Jimenez, A.; Gaire, R.; Paris, C. Text and data mining techniques in adverse drug reaction detection. ACM Comput. Surv. **2015**, 47, 56. Link
12. Bandyopadhyay, S.; Wolfson, J.; Vock, D.M.; Vazquez-Benitez, G.; Adomavicius, G.; Elidrisi, M.; Johnson, P.E.; O'Connor, P.J. Data mining for censored time-to-event data: A Bayesian network model for predicting cardiovascular risk from electronic health record data. Data Min. Knowl. Discov. **2015**, 29, 1033–1069.
13. Mohammad Hossein Tekieh1, Bijan Raahemi, "Importance of Data Mining in Healthcare: A Survey", IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, **2015.**
14. Parvez Ahmad, Saqib Qamar and Syed Qasim Afser Rizvi, "Techniques of Data Mining in Healthcare: A Review", International Journal of Computer Applications (0975 –8887), Vol 120, **2015**
15. Elahe Parva et al., "The Necessity of Data Mining in Clinical Emergency Medicine; A Narrative Review of the Current Literatrue", Bull Emerg Trauma 2017;5(2):90-9

16. Moore AC, Winkjer JS, Tseng TT. Bioinformatics Resources for MicroRNA Discovery. Biomark Insights. **2015**;10(Suppl 4):53–8
17. Paydar S, Pourahmad S, Azad M, Bolandparvaz S, Taheri R, Ghahramani Z, et al. The Evolution of a Malignancy Risk Prediction Model for Thyroid Nodules Using the Artificial Neural Network. Middle East Journal of Cancer. **2015**;7(1):47–52
18. Yang, J.-J.; Li, J.; Mulder, J.; Wang, Y.; Chen, S.; Wu, H.; Wang, Q.; Pan, H. Emerging information technologies for enhanced healthcare. Comput. Ind. **2015**, 69, 3–11.
19. Martin, C.M.; Félix-Bortolotti, M. Person-centred health care: A critical assessment of current and emerging research approaches. J. Eval. Clin. Pract. 2014, 20, 1056–1064.
20. Santos, R.S.; Malheiros, S.M.; Cavalheiro, S.; De Oliveira, J.P. A data mining system for providing analytical information on brain tumors to public health decision makers. Comput. Methods Prog. Biomed. 2013, 109, 269–282
21. Divya Tomar and Sonali Agarwal, “A survey on Data Mining approaches for Healthcare”, International Journal of Bioscience and Bio-Technology Vol.5, No.5 (2013)
22. M. Durairaj and V. Ranjani, “Data Mining Applications in Healthcare Sector: A Study”, 2013
23. Lu, R.; Lin, X.; Shen, X. SPOC: A secure and privacy-preserving opportunistic computing framework for mobile-healthcare emergency. IEEE Trans. Parallel Distrib. Syst. 2013, 24, 614–624 160.

24. Boris Milovic and Milan Milovic, “Prediction and Decision Making in Health Care using Data Mining”, International Journal of Public Health Science (IJPHS) Vol. 1, No. 2, December 2012
25. Shen, C.-P.; Jigjidsuren, C.; Dorjgochoo, S.; Chen, C.-H.; Chen, W.-H.; Hsu, C.-K.; Muller, R.; Robson, B.; Apte, C.; Weiss, S.; et al. A data-mining framework for transnational healthcare system. J. Med. Syst. 2012, 36, 2565–2575.
26. Lee, T.-T.; Liu, C.-Y.; Kuo, Y.-H.; Mills, M.E.; Fong, J.-G.; Hung, C. Application of data mining to the identification of critical factors in patient falls using a web-based reporting system. Int. J. Med. Inf. 2011, 80, 141–150.
27. Duan, L.; Street, W.N.; Xu, E. Healthcare information systems: Data mining methods in the creation of a clinical recommender system. Enterp. Inf. Syst. 2011, 5, 169–181
28. https://healthcareinamerica.us/how-data-mining-is-changing-health-care-27c1e9b3b372
29. Illhoi Yoo et al., “Data Mining in Healthcare and Biomedicine: A Survey of the Literature”, 2011
30. https://www.baianat.com/articles/mine-gold-in-your-data

Appendix

Figure 3.1 Various stages in the Knowledge Discovery Process

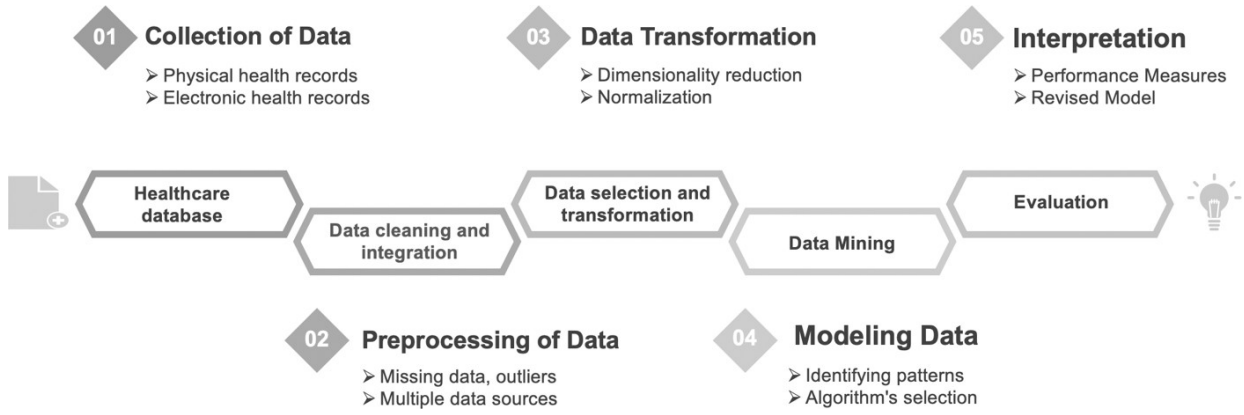

Figure 4 Different Data Mining Techniques in Healthcare

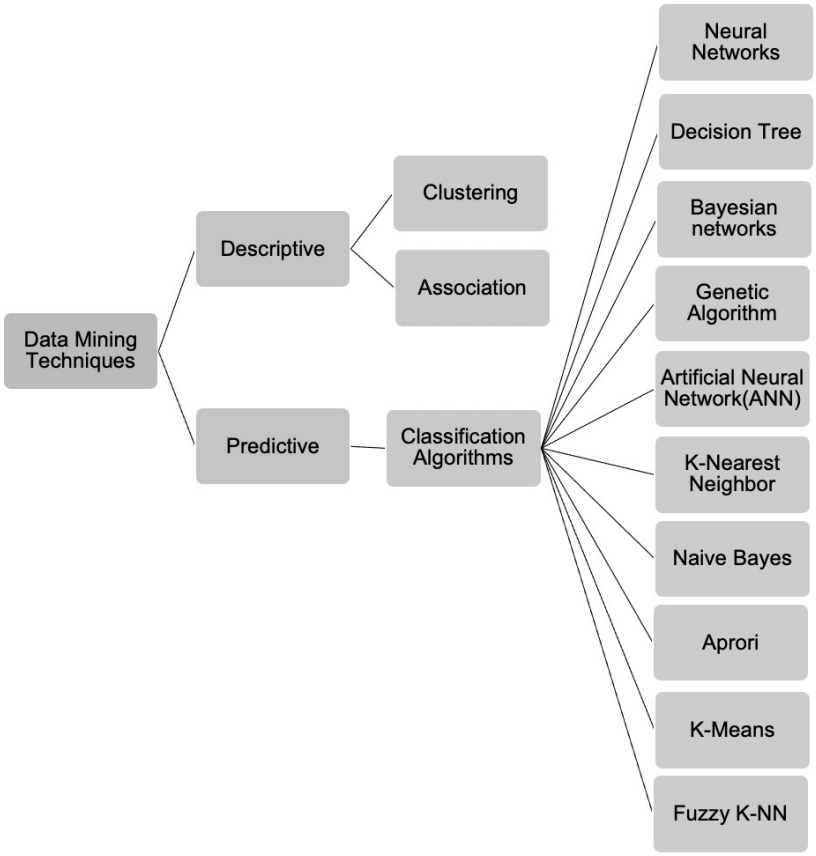

Table 4.1 Clustering types and their characteristics

	Partitioned Clustering (e.g., K-Means)	Hierarchical Clustering	Density Based Clustering
Function	Partition given 'n' data points into 'k' clusters based on similarity measure	decompose the data points either using bottom-up approach or top-down approach.	Uses DBSCAN and OPTICS to discover clusters based on the basis of density connectivity analysis
Pros	Less complex method and efficient	No need to define the number of clusters in advance	No need to define the number of clusters in advance and easily handle cluster with arbitrary shape
Cons	- Can handle only spherical shaped cluster - Requires number of clusters in advance	Can handle only spherical shaped cluster	Suitable for discovering cluster of arbitrary shapes and handles outliers
Related work	grouping of person based on high blood pressure and cholesterol level into high risk and low risk of having heart disease.	Grouping the patients based on their length of stay in the hospital that enhance the capability of hospital resource management	Clustering of Skin Wound Image using DBSCAN

Table 4.3 Classification types and their characteristics

Methods	Advantage	Disadvantage	Related Work
K-NN	Easy to implement	Sensitive to noise	classification of cardiovascular disease in order to generate early warning system
Decision Tree	Handles both numerical and categorical data	Restricted to one categorical output variable	characterize skin diseases in adults and children
SVM	Less overfitting issues	Training process takes more time	breast cancer diagnosis using hybrid SVM based strategy
Neural Network	Able to handle noisy data	Overfitting	analyzing chest diseases using ANN
Bayesian Belief Network	Better accuracy for Huge datasets	Dependency among variables leads to inaccurate results	analyze the psychiatric patient data using BBN in making significant decision regarding patient health s

Figure 5 Applications of Datamining in Healthcare

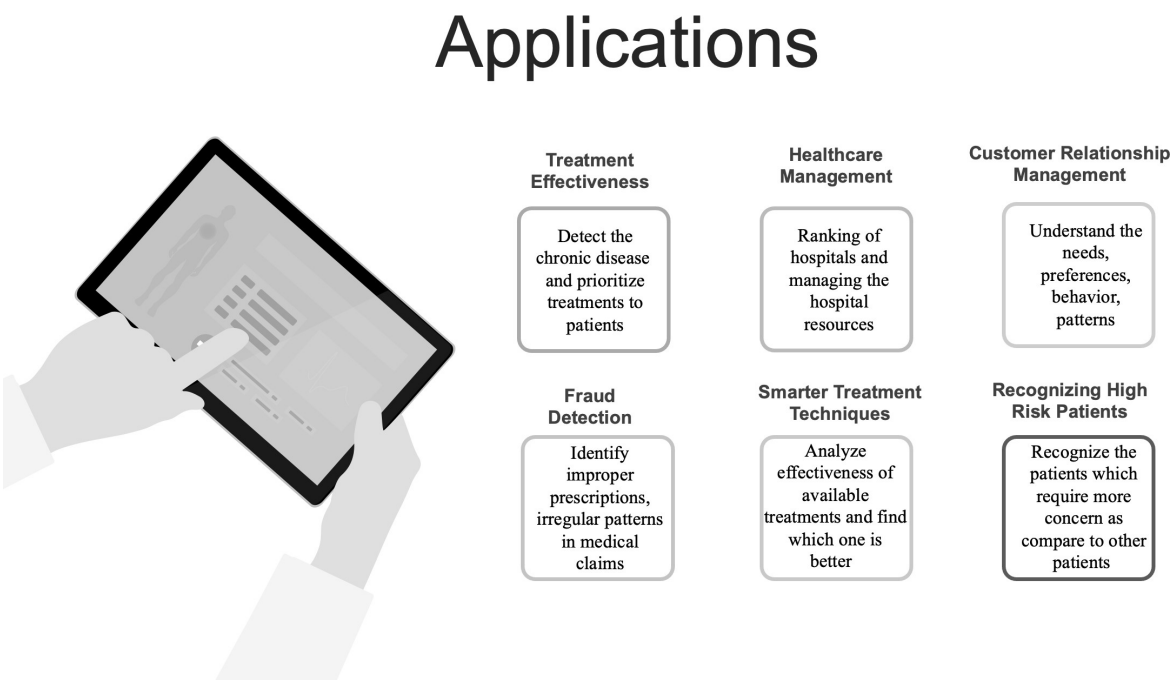

Figure 6 Challenges of Datamining in Healthcare

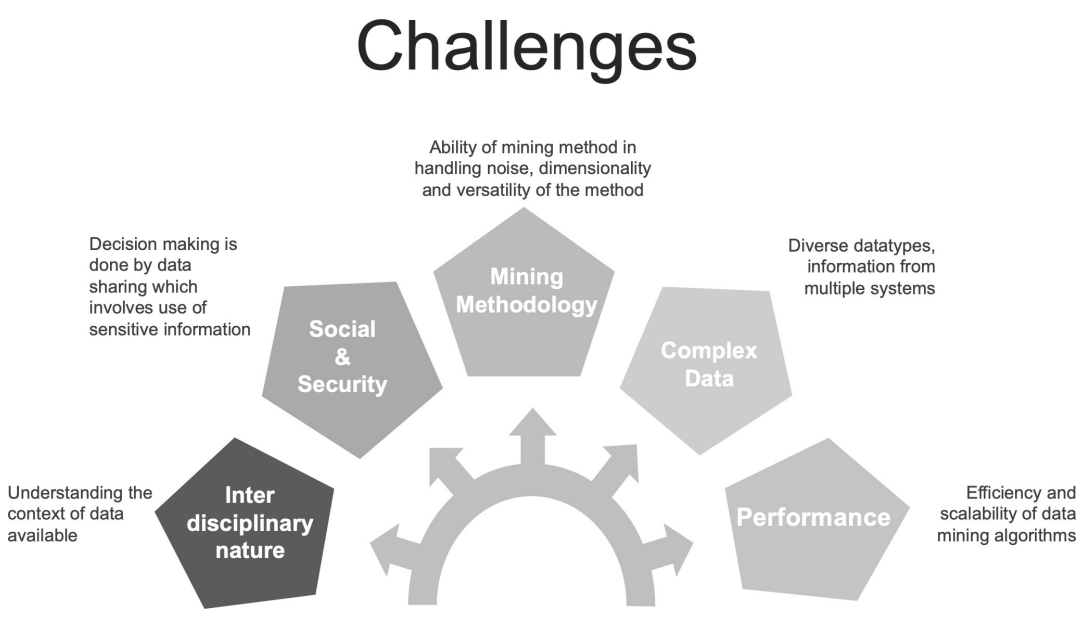

Figure 7 Future Directions of Datamining in Healthcare