ANALYSIS OF DATA MINING IN HEALTHCARE

Babitha Thamby, S. Sheeja*

Abstract

Data Mining has its prominent role in every sector by using the large amount of data from repositories. Healthcare is the most relevant area of data mining since it produces large volumes of data in every second. In this pandemic situation, we have to keep our eyes on medical data since the growth of all other sectors depends on the quality life style of human beings. Digital data in healthcare sector has a prominent role in the Data Mining area since Data Mining has very much efficient in analyzing, extracting data and finding patterns for knowledge discovery. Data Mining is having its own signature in healthcare services due to the exponential growth of electronic health records. Big data from healthcare sector is heterogeneous in nature and it has a lot to do with the organization data. Using Data Mining approaches for knowledge discovery can reduce the cost of treatment and also provide better treatment by giving better clinical decision support.

Keywords: Data Mining, healthcare, big data, knowledge discovery.

I. INTRODUCTION

The digital world technologies have put its signature in the healthcare domain by keeping medical records and related works in the repository and now it became an inextricable. Early days searching and booking clinical appointments by the public is done. But now the stage has reached in the prediction, clinical decisions, drug discovery, survival analysis and even post drug -administration studies by healthcare professionals. Since the world is going through many new diseases including pandemics, Data Mining has lot to do with the clinical big data that is accumulating each second even in the new normal. Making these exponential growths of data into opportunities are the difficult task. The invaluable advantages of Data Mining in medical big data have led to the extensive scope of the same.

The data that is accumulated in healthcare sector has many issues like missing data, irrelevant attributes and heterogeneous behaviour. Data Mining deals analysing large volumes of data, extracting knowledge[1] from it, making patterns for decision making and we call it as Knowledge Discovery in Databases (KDD) [2].Beginning from the initial stage of data collection from different sources, pre-processing the selected data and transforming the data into desired format, then Data Mining techniques are applied on the processed data for extracting valuable information. Finally, evaluation is done [3].

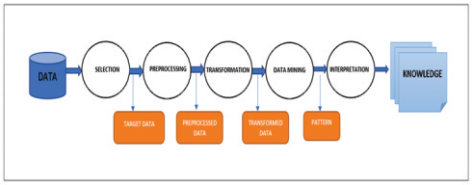

Fig1. Various stages of information extraction.

Many open-source data mining tools are available nowadays. The aim of this paper is to give a brief idea about the application of Data Mining in the field of healthcare and related areas. The extensive role of mining has its own signature in the medical area since people are prone to new diseases including pandemics like Covid-19 around the globe. Data analytics in healthcare widely spread in the following categories. They are descriptive, diagnostic, predictive and prescriptive analytics [4].

Department of Computer Science,
Karpagam Academy of Higher Education, Coimbatore, Tamil Nadu, India
*Corresponding Author

Descriptive Analytics: Historical data is analysed to get a better business prediction. It is a statistical method using data aggregation and mining on the collected data. For example, we can make a statistical representation about the success rate of online classes based on the data of different students and can be used to analyse the learner's engagement and learner's performance. The findings can identify the areas of improvement so quickly.

Diagnostic Analytics: It is a step more than the descriptive model. It searches for the reasons why the problem or situation occurred and also asking for the factors that triggers the matter. For example, if the descriptive analytics are showing the increasing trend of cervical cancer, then diagnostic analysis deals with the reason for the occurrence of the disease and it will helpful to provide all the requisite data.

Predictive Analytics: The method is used to make predictions about unknown future events. Predictive analytics uses many concepts from data mining, statistics, modelling, machine learning, and artificial intelligence for future outcome prediction. It can predict the complication of severe diseases. For a better prediction from the raw data, we can use Data Mining techniques like clustering, classification, association etc.

Prescriptive Analytics: It provides remedial decisions based on leading to optimal decision-making. It can be helpful in long term business operations. We can take advantage of future opportunities or can prevent a future risk. For example, one can deny any medicinal drug if it is harmful in the coming future.

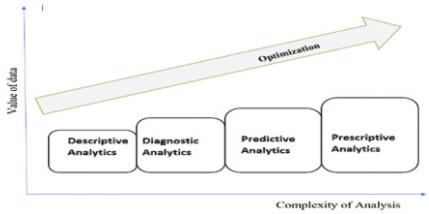

Fig 2. Various levels of Data analysis using big data.

II. DATA MINING METHODS IN HEALTHCARE

A huge mass of data are taken into consideration and from these big data, we are proceeding with various procedures and finally we are taking some decisions or formulating some useful information or knowledge discovery[5].

2.1. Classification

The mining function is used to assign data items to a target class. We can find out to which category of data the findings belongs. At first, we will have a training phase in which we are training our created models to get accurate results. In the second phase, we are testing the accuracy of our classification rules by testing the model on test data. The output variable of classification will be a category not a value like yes or no, spam or not spam, possible or not possible, 0 or 1 etc. Sometimes it will be multiple categorical like categories of different types of music. The algorithm which is used for the dataset classification is known as classifier.This method has been used in healthcare applications. K-Nearest Neighbor (K-NN), Decision Tree (DT), Support Vector Machine (SVM) are some of the algorithms used for classification [6].

2.2 Regression

Regression analysis is used to identify and analyse the relationship among variables because of the presence of the other factor. It is used to define the probability of the specific variable. Regression, primarily a form of planning and modelling. For example, we might use it to project certain costs, depending on other factors such as availability, consumer demand and competition. Primarily, it gives the exact relationship among two or more variables in the given data set.

2.3 Clustering

Clustering is the process of partitioning a set of data objects into subsets. This method is helpful when we do not have that

much information about various types of data objects active in a population. The data or objects with similar characteristics are grouped together and those with dissimilarities will be considered as another group. Prior to classification many datasets use clustering in order to group the data since clustering mainly deals with the descriptive analytics. A cluster-ID will be given to each cluster for the efficient management of complex dataset using machine learning. Identification of cancerous cells can be done with the clustering since we are grouping the cancerous cells and non-cancerous cells.

2.4 Association

Association is the process of finding a relation or association with between attributes. For instance, if a person buys bread, then there is a chance of buying butter/jam with it. The association rule could be developed from such relations and it is more helpful in healthcare sector since the association between clinical data values are often shows a prominent forecasting of various diseases including cancer and sometimes the pandemics too.

2.5 Outlier detection

In outlier detection, the data which is showing deviation from other data in a dataset is identified. The exclusion of such observations from the dataset will be done gradually. The deviation of data or errors may be occurred by natural, entry time, measurement, sampling or processing. This method is applicable in the case of data discrepancies in healthcare sector with the aim of cleaning the data in medical databases.

III. LITERATURE REVIEW

In this review, an extensive study was done on some areas of healthcare where Data Mining is applied.

Data Mining rules were applied by [7] for infection control and public health surveillance. One year data of Pseudomonas aeruginosa infection was collected from Birmingham Hospital. New interesting patterns in surveillance data were identified which helped in the control of infection. In the article [8] a model that helped to predict a patient's length of stay during a single visit, the time from hospital admission until discharge, and it helped the healthcare workers to provide better facilities and manage the resources efficiently. [9] Electronic data was used to get a better knowledge of admissions and overcrowding in the peak hours of hospital timings. The data was taken from hospitals from Northern Ireland to compare contrasting machine learning algorithms in predicting the risk of admission and mainly 3 algorithms were used namely logistic regression, decision trees and Gradient Boosted Machines (GBM). Another study was done by [10] for the detection of breast and colon cancer by using SVM classifier. Cervical cancer detection was done by using different classification techniques such as Naive Bayes (NB), C4.5 Decision Tree (C4.5), Sequential Minimal Optimization (SMO), Random Forest Decision Tree (RF), k-Nearest Neighbors (kNN), Multilayer Perceptron (MLP) Neural Network and Simple Logistic Regression (SLR); they have been used to classify the healthy and infected cancer sets. Primary liver cancer known as Hepatocellular carcinoma (HCC) survival rate was analysed by using a novel protein. In the article [11], using Naïve Bayes (NB), Support Vector Machine (SVM) and Decision Tree for predicting the possibility of heart disease for diabetic patients.

Since pandemics have come to threaten the entire globe, the tracking and post disease study have great importance in this human era [12] where the powerful impact of traditional Chinese medicine on pandemics like Covid-19 are threatening. In[13] study was done to find out the post-covid19 syndrome[14]. The study is still going on with the help of clinical big data collected from different hospitals [15]. The vaccine related studies are also done with the help of big data and mining. The Covid-19 vaccine Pfizer has undergone for many studies even after clinical trials [16] and

revealed that people with significant allergic reactions could not take the vaccine. Health insurance sector has impacts by mining techniques,[17] showing that fraud detection in the insurance industry is highly capable of alarming unauthorized activities by using data mining methods. In [18] the mining methods were used to detect sentiment dynamics during the Covid-19 pandemic by using social media.

IV. CHALLENGES IN DATA MINING

One of the major challenges of Data Mining is to get the accurate and significant data. Sometimes, it is difficult for the hospitals and clinics to get quality data. The field of healthcare is still complex and heterogeneous; sometimes the patients are not willing to provide their health data and it is a case of privacy. In the case of healthcare finance, they are keeping the data. But, sometimes it is not that much quantitative. Data sharing is another issue faced by the data mining techniques. Clinical data is not shared by the organisations since it is the violation of patient's privacy terms [19].

The data is facing its own limitations like heterogeneous and high volume like administration data like booking and contact details, doctor consultation, laboratory results, doctors review, serum levels, patient's family history, contact details in case of pandemics etc. Before the data mining process, the data need to be collected. One of the approaches is to successfully build a data warehouse and it is a process causing economic burden. The other issue is inconsistent or non-standardized data, corrupted or missing data, data ownership, ethical issue and legal issues [20].

Applications	• Hospital management • Fraud detection • Disease diagnosis & prognosis • Post disease study • Drug administration • Sentiment analysis of infected patients • Survival analysis • Health insurance
Advantages	• Organisational knowledge-based data • Target oriented decision making • Cost efficient • Effective treatment • Finding hidden patterns in existing diseases • Prediction of diseases • Fraud detection is efficient • Fast management of data
Issues	• Heterogeneous data • Missing data • Incorrect/incomplete data • Data privacy & security • Large volume of data • Training for expert systems

Table1. Applications, advantages and issues of Data Mining in health care.

V.CONCLUSION

Applications of Data Mining brought many advantages like resources optimisation, better hospital management, better treatment, clinical decision support, diagnosis and prognosis of diseases [21], health insurance policy planning, drug discovery and post drug administration studies, etc. The paper described the data analytics methods, the steps in knowledge discovery, application of Data Mining in healthcare and the issues faced in the data handling [22]. By doing some literature reviews, it is clear that we can do a lot with these voluminous medical data by exploring the Data Mining methods. It will be helpful for the entire society since the world is still facing pandemics and related issues [23]. Researches should be more to the side of healthcare since the growth of entire world depends upon healthy human beings. Medical organizations should provide quality data for doing any research. By providing a quality and enough data for research, Data Mining can do better discovery of knowledge from medical data.

REFERENCES

[1] Sarma, Jhumpa. "Analysis of Data Mining Tools Used in Healthcare Domain." (2021).

[2] Chandrakala, P., A. Sumithra, A. Saranya, and R. Bagavathi Lakshmi. "Influence of Data Mining Techniques in Healthcare Research." Turkish Journal of Computer and Mathematics Education (TURCOMAT) 12, no. 14 (2021): 1303-1311.

[3] Almansoori, Afrah, Mohammed AlShamsi, Said A. Salloum, and Khaled Shaalan. "Critical review of knowledge management in healthcare." Recent Advances in Intelligent Systems and Smart Applications (2021): 99-119.

[4] Islam, Md Saiful, Md Mahmudul Hasan, Xiaoyi Wang, and Hayley D. Germack. "A systematic review on healthcare analytics: application and theoretical perspective of data mining." In Healthcare, vol. 6, no. 2, p. 54. Multidisciplinary Digital Publishing Institute, 2018.

[5] Santos-Pereira, Judith, Le Gruenwald, and Jorge Bernardino. "Top Data Mining Tools for the Healthcare Industry." Journal of King Saud University-Computer and Information Sciences (2021).

[6] S. L. Nalawade and R. V. Kulkarni, DzApplication of Data Mining in Health Care,dz International Journal of Science and Research (IJSR), vol. 5, no. 4, pp. 262-268, 2016.

[7] Brossette, Stephen E., Alan P. Sprague, J. Michael Hardin, Ken B. Waites, Warren T. Jones, and Stephen A. Moser. "Association rules and data mining in hospital infection control and public health surveillance." Journal of the American medical informatics association 5, no. 4 (1998): 373-381.

[8] Azari, Ali, Vandana P. Janeja, and Alex Mohseni. "Healthcare data mining: predicting hospital length of stay (PHLOS)." International Journal of Knowledge Discovery in Bioinformatics (IJKDB) 3, no. 3 (2012): 44-66.

[9] Graham, Byron, Raymond Bond, Michael Quinn, and Maurice Mulvenna. "Using data mining to predict hospital admissions from the emergency department." IEEE Access 6 (2018): 10458-10469.

[10] AbdElNabi, Mohamed Loey Ramadan, Mohammed Wajeeh Jasim, Hazem M EL-Bakry, Hamed N. Taha, and Nour Eldeen M Khalifa. "Breast and colon cancer classification from gene expression profiles using data mining techniques." Symmetry 12, no. 3 (2020): 408.

[11] Kumar, Abhishek, Pardeep Kumar, Ashutosh Srivastava, VD Ambeth Kumar, K. Vengatesan, and Achintya Singhal. "Comparative Analysis of Data Mining techniques to predict heart disease for diabetic patients." In International Conference on Advances in Computing and Data Sciences, pp. 507-518. Springer, Singapore, 2020.

[12] Zhang, Kai. "Is traditional Chinese medicine useful in the treatment of COVID-19?." The American journal of emergency medicine 38, no. 10 (2020): 2238.

[13] Sarker, Abeed, and Yao Ge. "Mining long-COVID symptoms from Reddit: characterizing post-COVID syndrome from patient reports." JAMIA open 4, no. 3 (2021): ooab075.

[14] Hall, Jocelin, Katherine Myall, Jodie L. Lam, Thomas Mason, Bhashkar Mukherjee, Alex West, and Amy Dewar. "Identifying patients at risk of post-discharge complications related to COVID-19 infection." Thorax 76, no. 4 (2021): 408-411.

[15] Radanliev, Petar, David De Roure, and Rob Walton. "Data mining and analysis of scientific research data records on Covid-19 mortality, immunity, and vaccine development-In the first wave of the Covid-19 pandemic." Diabetes & Metabolic Syndrome: Clinical Research & Reviews 14, no. 5 (2020): 1121-1132.

[16] Mahase, Elisabeth. "Covid-19: People with history of significant allergic reactions should not receive Pfizer vaccine, says regulator." BMJ: British Medical Journal (Online) 371 (2020).

[17] Saldamli, Gokay, Vamshi Reddy, Krishna S. Bojja, Manjunatha K. Gururaja, Yashaswi Doddaveerappa, and Loai Tawalbeh. "Health Care Insurance Fraud Detection Using Blockchain." In 2020 Seventh International Conference on Software Defined Systems (SDS), pp. 145-152. IEEE, 2020.

[18] Yin, Hui, Shuiqiao Yang, and Jianxin Li. "Detecting topic and sentiment dynamics due to COVID-19 pandemic using social media." In International Conference on Advanced Data Mining and Applications, pp. 610-623. Springer, Cham, 2020.

[19] Comandé, Giovanni, and Giulia Schneider. "Regulatory Challenges of Data Mining Practices: The Case of the Never-ending Lifecycles of 'Health Data'." European Journal of Health Law 25, no. 3 (2018): 284-307.

[20] Sharma, Anu, M. K. Sharma, and R. K. Dwivedi. "Literature Review and challenges of Data Mining techniques for Social Network Analysis." Advances in Computational Sciences and Technology 10, no. 5 (2017): 1337-1354.

[21] Oskouei, Rozita Jamili, Nasroallah Moradi Kor, and Saeid Abbasi Maleki. "Data mining and medical world: breast cancers' diagnosis, treatment, prognosis and challenges." American journal of cancer research 7, no. 3 (2017): 610.

[22] Wang, Lidong, and Cheryl Ann Alexander. "Big data analytics in medical engineering and healthcare: methods, advances and challenges." Journal of medical engineering & technology 44, no. 6 (2020): 267-283.

[23] Ayyoubzadeh, Seyed Mohammad, Seyed Mehdi Ayyoubzadeh, Hoda Zahedi, Mahnaz Ahmadi, and Sharareh R. Niakan Kalhori. "Predicting COVID-19 incidence through analysis of google trends data in iran: data mining and deep learning pilot study." JMIR public health and surveillance 6, no. 2 (2020): e18828.