INTELLIGENT PREDICTIVE METHODS USING DATA MINING IN HEALTH CARE : A SURVEY

Babitha Thamby, S. Sheeja*

Abstract

In every field of knowledge data mining has its permanent role for extracting data from large number of repositories. Healthcare field is producing large amounts of data so that data mining has a lot to do in this particular field of knowledge, Since the globe is going through pandemic situation the human life value has very much importance since the survival rate of human beings could do a lot of with global economy and all other sectors with which humans are very much connected. Data Mining has very much efficient in analyzing, extracting data and finding patterns for knowledge discovery [1] in the medical field as it producing millions of digital and non-digital data in every second. Since the electronic health records (EHR) is growing exponentially, data mining is having its own signature in healthcare services. The heterogeneous big data from healthcare sector has very much to do with the organization data. By using different approaches for knowledge discovery like historical data analysis, pattern finding and other methods for the diagnosis of diseases can be done and can also reduce the treatment cost. Also mining has a lot to do with the various sub sectors in healthcare including prediction, clinical decision support, survival rate etc. using machine learning algorithms. There are some difficulties in getting and handling the medical data since it has a lot of issues regarding privacy, heterogeneity etc.

Key words: Data Mining, healthcare, knowledge discovery, prediction, applications.

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

I INTRODUCTION

The digital world has signed its signature in the healthcare domain by keeping medical records and other related works in the repository and it became an inextricable now. Other than searching and booking doctor appointments, now the stage has reached in the area of prediction, clinical decisions, drug discovery, survival analysis and even post drug -administration studies. Since the globe is facing many new diseases including pandemics like covid-19, the data mining has very much to do with the clinical big data that is accumulating per second even in the new normal. The invaluable advantages of mining in healthcare big data have led us to the predictive analytics and lot of opportunities in this area.

Machine learning in healthcare is now has a prominent role and it can help both patients as well as clinical experts in variety aspects. The common healthcare use cases for ML are clinical decision support, drug impacts and the development of care recommendations in clinical level. The performance of ML in automatic diagnosis and prediction has equal role that was given by the experienced radiologist. Various ML algorithms were used to detect diseases including cancer so that the progression of such diseases can be cured. In United States nearly 1.2 billion medical documents are available per year and they are helpful in mining of data for more accurate clinical results.

The data that is accumulated in medical field has many issues like noisy data, missing data, irrelevant attributes and heterogeneous behaviour. In mining we clean the gathered data using different methods, transforming it and analysing large volumes of data, extracting knowledge from it, making

some understandable patterns for decision making and it is known as Knowledge Discovery in Databases (KDD) [2]. Beginning from the very much initial stage of data collection from different sources, pre-processing the data and transforming the data can be done into a desired format. Different Data Mining techniques are applied on the processed data for extracting valuable information. And Finally, the evaluation is done [3].

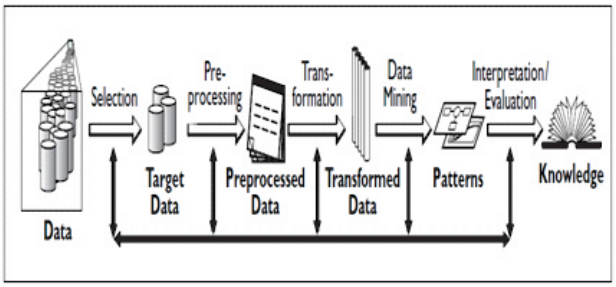

Fig. 1 Various levels of Information Extraction.

A. Selection of Data

For Knowledge discovery the original massive data will be taken into account. Depends on the application, relevant data will be taken for building models.

B. Pre-processing the Data

The healthcare data will be noisy or missing. So proper methods like regression can be considered for data pre-processing for making it more reliable.

C. Transformation of Data

Since the healthcare sector needs more accurate data, the transformation of data will be more application-specific or project-specific. Aggregation or consolidation of data will be done with care for reducing the data that is valid and appropriate for mining.

D. Data Mining

Using intelligent algorithms data patterns can be identified in this stage. Knowledge discovery can be done and it will be helpful in clinical decision making.

E. Data Interpretation or Evaluation

The available data patterns will be evaluated and will make decisions based on it and the knowledge will be presented.

Many open-source mining tools are available nowadays. This survey paper is meant to give a brief idea about the applications of mining in healthcare and related areas. The methods have extensive role in the particular area of medicine as people are very much prone to new diseases including pandemics like Covid-19 around the world. Data analytics in this area spread in the following categories. They are Descriptive, diagnostic, predictive, and prescriptive analytics [4].

A. 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.

B. Diagnostic Analytics:

It searches the reason for the phenomenon or situation occurred and asking for the facts that triggers the matter. For example, if the descriptive analytics is showing the progressing trend of cervical cancer then the diagnostic analysis will deal with the reason for the happening of the disease and will be helpful to give all the data required.

C. Predictive Analytics:

It is used to make predictions about future events which are unknown. Predictive analytics use statistics, machine learning, modelling, and artificial intelligence for future outcome prediction. It is capable of predicting the

complication of severe diseases. Data mining techniques like clustering, classification, association etc. can be used for a better prediction from the raw data.

D. Prescriptive Analytics:

Leading to optimal decision-making by giving remedial solutions. It will be helpful for long term business operations. We can take advantage of future opportunities or can prevent a future risk. Example, anyone can refuse or deny a drug which is harmful to him.

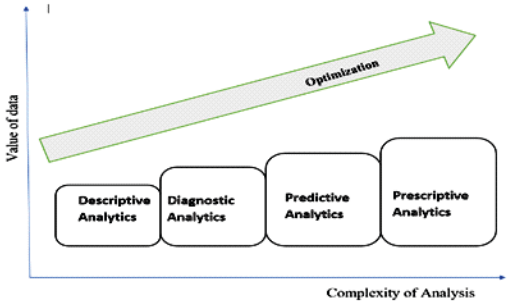

Fig 2. Various levels of Data analysis using big data.

II DATA MINING METHODS IN HEALTHCARE

A large volume of data is taken into account and from these big data, proceeding with various procedures and we are taking some useful or understandable patterns or doing knowledge discovery [5].

A. Classification

The function used to sort data items to a target class. We are finding out to which category of data; our findings belong to. In the beginning we will have a training phase. We are training our created models for getting accurate results. During second phase we will test the accuracy of our classification rules by using any testing model on test data. The output will be a category like Yes/No, Spam/Not Spam, possible/not possible, 0/1etc.The algorithm 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].

B. Regression

It is used for identification and analyzation of the relationship between variables as by the presence of the other factor and is helpful in defining the probability of specific variable. It is helpful in projecting certain costs, depending on other factors such as competition, availability and consumer demand. It is helpful in giving the exact relationship between two or more variables in the given data set.

C. Clustering

Here we are partitioning set of data objects into subsets. It is helpful if we do not have much information about various types of data objects active in a population. The data or objects with similarities are grouped together and those with dissimilarities will be considered as another group. Each cluster will be given a cluster-ID for the smooth and efficient management of complex datasets using machine learning. Cancerous cells can be identified with the clustering as we are grouping cancerous and non-cancerous cells.

D. Association

It is the process of finding a relation between attributes. It produces If…. then patterns. Example if a person buys bread, then there will be 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. The association between clinical data values showing a prominent forecasting of diseases including cancer and sometimes the pandemics too.

E. Outlier Detection

There data which is showing deviation from other data in a dataset is identified and such observations from the dataset will be excluded gradually. The deviation of data or errors

may be occurred by natural, entry time, measurement, sampling or processing. It is applicable in the case of data discrepancies in healthcare sector with the purpose of cleaning medical data.

III APPLICATION OF DATA MINING TECHNIQUES IN HEALTHCARE

A. Diagnosis

Data mining is helpful in decision making with knowledge discovery. The number of inputs will be large. It can analyse various pathological signals and medical images.

B.Treatment

Based on modelled historical performance, the available best treatment plans can be selected from a pool of methods.

C. Prediction / Prognosis

Outcome of treatment plans can be improved with the analysis of risk assessment and accurate predictive analytics.

D.Measuring Treatment Effectiveness

The outcomes of various drug therapies can be assessed.

E. Patient Satisfaction and Economic Stability

Use of non-beneficial drugs can be eliminated. Since the outcome of therapies could be done, harmful procedures can be avoided.

F. Biological Analysis

Analytical tasks can be automated like blood analysis and thereby tracking the level of glucose, ion level in body fluids, also detection of pathological condition.

G.Hospital Management

Optimize allocation of resources and assist in future planning for improved services.

H. Fraud Detection

Helpful in identification of uncommon patterns. Data mining applications fraud detection can identify unusual prescriptions, duplicate drugs, manipulation in medical and insurance claims etc.

I. Drug & Vaccine Findings

The drug discovery based on the disease study has a major role in data mining.

IV LITERATURE REVIEW

In this review an extensive study was done on some areas of healthcare which uses data mining.[7] used mining methods for infection control and public health surveillance.One year data of Pseudomonas aeruginosa infection was taken from Birmingham Hospital and new patterns were identified which helped in the control of infection. In [8] a model that helped to predict the stay duration of a patient during a single visit, the time taken from hospital admission until his discharge, and it helped to provide better facilities and manage the resources in efficient way. In [9] Electronic Health records were taken into get an idea of admissions and overcrowding in the peak hours of hospital timings. Hospital data from Northern Ireland were taken to compare different contrasting machine learning (ML) algorithms in predicting the risk of admission and mainly three algorithms were used, logistic regression, decision trees (DT) and gradient boosted machines (GBM). Another study was done by [10] for the detection of breast and colon cancer using SVM classifier. Cervical cancer detection was done by using different classification techniques such as Naive Bayes (NB), Random Forest Decision Tree (RF), Sequential Minimal Optimization (SMO),C4.5 Decision Tree (C4.5), k-Nearest Neighbors (kNN), Multilayer Perceptron (MLP) Neural Network and Simple Logistic Regression (SLR) have been taken to classify the healthy sets and infected cancer sets. The Primary liver cancer known as Hepatocellular carcinoma

(HCC) survival rate was analysed. In the article [11], using Naïve Bayes (NB), Decision Tree and Support Vector Machine (SVM) for predicting the heart disease of diabetic patients.

Since the pandemics are threatening the entire population, the tracking and post pandemic infection study have great importance in this human era. In [12] a study was done to get the impact of traditional Chinese medicine on pandemics like Covid-19. In [13] study was done to find out the post-covid19 syndrome [14]. The research is still proceeding with the help of clinical data collected from various hospitals [15]. Outcome of different vaccinations were also done with the help of big data and mining. The Covid-19 vaccine Pfizer and other vaccines like sputnik, covishield, covaccine have undergone for many studies even after clinical trials [16] and revealed that people with significant allergic reactions showing reactions in their body. But eventually it was a safe side for the entire population since the recovery rate of the pandemic increased very much by the timely vaccination. The various genetical changes occurring to the novel corona virus also taken into study since by every three to four months new variations of the virus is coming and is having fast spreading capability. Health insurance sector has impacts by mining techniques,[17] reveals that fraud detection in the insurance industry is highly capable of alarming unauthorized activities using data mining methods. Mining methods used in [18] to detect sentiment analysis during the Covid-19 pandemic using social media.

V CHALLENGES IN HEALTHCARE DATA MINING

One of the big challenges of data mining in healthcare is to get the data that is relevant as well as accurate. Sometimes it is difficult for the hospitals and clinics to get data of quality. The medical data is still complex and heterogeneous, sometimes the patients having a problem of privacy to provide their health data. In the case of healthcare finance, they are keeping the data. But sometimes it is not that much quantitative. Unwillingness of 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, doctors review, serum levels, laboratory results, patient's family history, contact details in case of pandemics etc. Before the data mining process, the data need to be collected. Successfully building a data warehouse is an economic burden. The other issue is inconsistent or non-standardized data, corrupted or missing data, data ownership, ethical issue and legal issues [20].

Data mining Applications in healthcare	➢ Efficient Hospital management ➢ Drug administration ➢ Sentiment analysis of patients ➢ Survival rate analysis ➢ Health insurance ➢ Fraud detection ➢ Disease diagnosis & prognosis ➢ Post disease study
Advantages of mining	➢ Organisational knowledge-based data ➢ Data Mining Mitigates Potential Drug Interactions ➢ Improves Patient Outcomes and Safety Precautions ➢ Prediction of diseases ➢ Fraud detection is efficient. ➢ Fast management of data. ➢ Target oriented decision making ➢ Cost efficient ➢ Matches Specialist to Patient ➢ Effective treatment ➢ Finding hidden patterns in existing diseases
Issues faced with data	➢ Heterogeneous data ➢ Missing data ➢ Incorrect/incomplete data ➢ Data privacy & security ➢ Large volume of data ➢ Training for expert systems

Table1. Various applications, advantages and issues with Data mining in health care.

VI 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. In this paper we discussed 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 clearer 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 me 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. The future of healthcare sector may very much depend data mining for minimizing healthcare costs, predicting treatment plans as well as best practices, drug discovery, measuring drug effectiveness, detect fraudulent insurance and medical claims. The ultimate aim will be to provide best available patient care to improve the quality of human life..

REFERENCES

[1] Available at: https://www.ceine.cl/wp-content/uploads/2012/12/KDD.png.

[2] P. Chandrakala et al., “Influence of data mining techniques in healthcare research,” Turk. J. Comput. Math. Educ. (TURCOMAT), vol. 12, no. 14, pp. 1303-1311, 2021.

[3] A.Almansoori et al., “Critical review of knowledge management in healthcare,” Recent Adv. Intell. Syst. Smart Appl., pp. 99-119, 2021

[4] Md. Islam, “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,. Multidisciplinary Digital Publishing Institute, vol. 6, no. 2, 2018, P. 54.

[5] J. Santos-Pereira et al., “Top data mining tools for the healthcare industry,” J. King Saud Univ. Comput. Inf. Sci., 2021

[6] S. L. Nalawade and R. V. Kulkarni, “DzApplication of data mining in health care,dz,” Int. J. Sci. Eng. Res. (IJSR), vol. 5, no. 4, pp. 262-268, 2016.

[7] S. E. Brossette et al., “Association rules and data mining in hospital infection control and public health surveillance,” J. Am. Med. Inform. Assoc., vol. 5, no. 4, pp. 373-381, 1998.

[8] A. Azari et al., “Healthcare data mining: Predicting hospital length of stay (PHLOS),” Int. J. Knowl. Discov. Bioinformatics, vol. 3, no. 3, pp. 44-66, 2012

[9] B. Graham et al., “Using data mining to predict hospital admissions from the emergency department,” IEEE Access, vol. 6, pp. 10458-10469, 2018

[10]M. L. Abdelnabi, “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, vol. 408, no. 3, 2020.

[11]A. Kumar et al., “Comparative Analysis of Data Mining techniques to predict heart disease for diabetic patients”

in, Communications in Computer and Information Science Intl. Conf. on Adv. in Comput. and Data Sci. Singapore: Springer, pp. 507-518, 2020

[12]K. Zhang, "Is traditional Chinese medicine useful in the treatment of COVID-19?," Am. J. Emerg. Med., vol. 38, no. 10, p. 2238, 2020

[13]A. Sarker and Y. Ge, "Mining long-COVID symptoms from Reddit: Characterizing post-COVID syndrome from patient reports," JAMIA Open, vol. 4, no. 3, p. ooab075, 2021

[14]J. Hall et al., "Identifying patients at risk of post-discharge complications related to COVID-19 infection," Thorax, vol. 76, no. 4, pp. 408-411, 2021

[15]P. Radanliev et al., "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 Metab. Syndr., vol. 14, no. 5, pp. 1121-1132, 2020

[16]E. Mahase, "Covid-19: People with history of significant allergic reactions should not receive Pfizer vaccine, says regulator," BMJ, vol. 371, m4780, 2020

[17]G. Saldamli et al., "Health care insurance fraud detection using Blockchain" in Seventh Intl. Conf. on Softw. Defined Syst. (SDS). IEEE, 2020, pp. 145-152.

[18] H. Yin et al., "Detecting topic and sentiment dynamics due to COVID-19 pandemic using social media" in, Lecture Notes in Computer Science Intl. Conf. on Adv. Data Min. and Appl. Cham: Springer, pp. 610-623, 2020

[19]G. Comandé and G. Schneider, "Regulatory challenges of data mining practices: The case of the never-ending lifecycles of "health data"," Eur. J. Health Law, vol. 25, no. 3, pp. 284-307.

[20]A. Sharma et al., "Literature Review and challenges of Data Mining techniques for Social Network Analysis," Adv. Comp. Sci. Technol., vol. 10, no. 5, pp. 1337-1354, 2017.

[21]R. J. Oskouei et al., "Data mining and medical world: Breast cancers' diagnosis, treatment, prognosis and challenges," Am. J. Cancer Res., vol. 7, no. 3, p. 610-627, 2017.

[22]L. Wang and C. A. Alexander, "Big data analytics in medical engineering and healthcare: Methods, advances and challenges," J. Med. Eng. Technol., vol. 44, no. 6, pp. 267-283, 2020

[23]S. M. Ayyoubzadeh et al., "Predicting COVID-19 incidence through analysis of google trends data in iran: Data mining and deep learning pilot study," JMIR Public Health Surveill., vol. 6, no. 2, p. e18828, 2020.