International Journal of Innovative Research in Computer Science & Technology (IJIRCST)
ISSN: 2347-5552, Volume-12, Special Issue-1, March-2024
https://doi.org/10.55524/CSISTW.2024.12.1.69
Computing for Sustainable Innovation: Shaping Tomorrow's World (CSISTW-2024) on 16th-17th February, 2024
Organized by B.N. College of Engineering and Technology, Lucknow, India

Using and Investigating Big Data Mining in Clinical Medicine

Deepika Saxena[1] and Dr Raj Thaneeghaivel V[2]

[1] Ph.D. Research Scholar, Department of Computer Application, SRK University, Bhopal, Madhya Pradesh, India
[2] Professor, Department of Computer Application, SRK University, Bhopal, Madhya Pradesh, India

Correspondence should be addressed to Deepika Saxena; deepika.mist@gmail.com

ABSTRACT- To examine big data mining ideas,technology, and applications in clinical medicine. PubMed and the Chinese Hospital Knowledge Database were used to find English and Chinese publications from 1975 to 2015 that discussed big datamining ideas, technologies,and practical applications in clinical medicine. Selection of Studies: original research on the theory and technology of big data mining and its uses in the medical industry. Fuzzy theory, rough set theory, cloud theory, Dempster-Shafer theory, artificial neural network, genetic algorithm, inductive learning theory, Bayesian network, decision tree, pattern recognition, high-performance computing, and statistical analysis are some of the fundamental theories and technologies of big data mining that were covered in this review. Big data mining's use in clinical medicine was examined in the Big data mining has the potential to play an important role in clinical medicine.

KEYWORD- Big Data; Clinical Medicine; Data Mining.

I. INTRODUCTION

The internet, cloud computing, and mobile internet revolutions, in particular, have led to continual scientific and technological developments that have resulted in an incredible rate of growth in data. The age of "Big Data," which refers to massive amounts of data that are beyond the capacity of standard computing tools to gather, organize, or analyze, has quietly ushered in for humankind. Big data is defined by the four Vs: volume, variety, velocity, and value.[1] The term "medical big data" describes the application of big data in the medical domain following innovative methods of storing, retrieving, exchanging, analyzing, and presenting information related to human health and medicine. The growth of medical big data can be attributed to advancements in medical technologies and hospital information systems.

As of the end of November 2013, there were 962,000 medical facilities in the country, made up of 923,000 other facilities and 24,000 hospitals. 6.9 billion People sought medical attention at all of the country's hospitals. Medical data from a medium-sized hospital can currently build to several trillion bytes each year. Furthermore, data mining technology is always developing to make it possible to extract data from databases for study. Data mining technology is more useful than the original decision support system when it comes to data integration.

A. Data Mining Theories:

- **Fuzzy Theories:** Fuzzy sets and fuzzy reasoning, which ascertain the association of pertinent items by identifying the membership function and reasoning in accordance with predetermined rules, are the mathematical cornerstones of fuzzy theory. Fuzzy theory works particularly well with concepts that follow specific principles and are frequently applied in clinical medicine. A fuzzy hierarchy technique was employed by Kim afar et al. To examine the variables influencing nurses' satisfaction with hospital information systems. They took into account 22 secondary criteria in addition to the three primary factors of information, system, and service quality. In the end, they discovered that the most significant index (holding 58% of the weight) was information quality. In a different type of example, Kuo et al. predicted the prognosis of prostate cancer using two-stage fuzzy neural networks. This approach investigated the connection between clinical characteristics and outcomes.
- **Cloudy Theories:** Fuzzy theory is extended by cloud theory. For a qualitative concept articulated in terms of linguistic values and its numeric representation, it is seen as an uncertain transformative model. By combining the fuzziness and randomnessof qualitative notions, it circumvents the drawbacks of the conventional fuzzy set theory and offers a potent technique for the information processing of qualitative and quantitative integration. A new cloud-based health management paradigm was created by Liu and Xiao.

  Through "horizontal crossing" and "vertical stratification," they were able to bring together components of the health system, and they subsequently leveraged these links to share and work with all types of clouds. A brand-new system of health management based on identification, survey, and regulation was put in place. The system oversees health networks and takes proactive measures before.
- **Rough Set Theory:** Rough set theory is a mathematical method for explaining and resolving unclear problems. This concept has become one of the most useful tools in the data mining industry. Its primary feature is that, in order to solve problems outside of the data set, it doesn't require any prior knowledge, making it easy to use and straightforward. The disadvantage is that it can be difficult to deal with continuous attributes directly. It is widely used in

evaluation, reduction, analysis, and finding of data. Gil-Herrera et al.used a crude set theory-based meta-analysis technique to help individual patients find homogeneous groupings. This approach uses differentiated and non distinguished rough set theories to categorize patients into homogeneous groups. A total of 1111 patients were divided into This approach resulted in three patient categories with significantly reduced statistical heterogeneity (16.8%, 0 and 0 respectively). This approach will ultimately aid in the personalization of medical decision-making and has significant potential for the automation and standardization of the detection process as well as for meta-analysis of the management of intermittent peritoneal dialysis. In order to facilitate the early diagnosis of breast cancer, Li et al. proposed utilizing the Ada Boost support vector machine (SVM) classification algorithm in combination with cluster boundary sampling preprocessing techniques (CBS Ada Boost SVM). The algorithm used a machine learning technique to diagnose unknown photo data. In parallel, a genetic algorithm based on a crude set reduction approach was used to remove redundant traits. To evaluate the effectiveness of the proposed methods, the accuracy, confusion matrix, and receiver were computed on a digital data switching matrix. Operational characteristic curves gave the physicians vital information for the early diagnosis of breast cancer.

- **Artificial neural network ANN:** Is an algorithm developed from simulations of the human brain? It learns statistical law of data in a way similar to the mind's memory and induces a data model that can describe the features of a sample. Then, the data model that has been learned is used to classify new data. It has characteristics include association functions, self-organization, high fault tolerance, massively parallel processing, and adaptive capabilities. An artificial neural network (ANN) system is more intelligent than a fuzzy logic system. This approach represents a significant advancement in the medical field. In order to assess patients' heart health prognoses, Sun karia et al. Recently developed a heart rate classification system based on an ANN. They selected 46 patients' electrocardiogram (ECG) data from a hospital outpatient information system, entered each patient's ECG data into the classification algorithm, and then compared the results with those from heart disease specialists. The patients' true cardiac health was in excellent agreement with the algorithm's and the experts' findings.
- **Genetic algorithm:** The first genetic algorithm was proposed by Holland in 1975.This method encodes and computes data by simulating biology evolution in nature and induces selection, crossover, and mutation operations to filter data and eliminate invalid information through the rule of survival of the fittest; in the end, the required data are kept. Genetic algorithms are easy to parallel and have been widely used in classification and optimization problems. Johnson et al.used this algorithm to forecast the progression of Alzheimer's disease. Compared to the stepwise selection method in their research, the genetic algorithm was better at predicting the development from a healthy state to mild cognitive impairment, and then to the final evolution into Alzheimer's disease. The results suggested that, in models of disease progression prediction, the contribution of combined variables is more important than that of a single variable.
- **Decision Tree:** Biggest In machine learning, a decision tree is a prediction model. It uses a set of rules to categorize data and uses knowledge gain from information theory to determine the field of the data in the database. Next, a decision tree node is created, and the tree's branch is arranged in accordance with the various field values. The lowest nodes of the tree and each branch's branches are finally established. An easily comprehensible feature that can produce rules is displayed in a decision tree. Furthermore, it is a well-liked data mining technique because of the minimal calculation volume. These days, the classification and regression tree, C4.5, iterative dichotomizer 3 (ID3), and supervised learning in quest are among the frequently utilized decision tree techniques. Right now, the ID3 approach is the most widely used of all. This approach was successfully applied by Yang et al.to investigate risk assessments of public health incidents. Additionally, in order to increase the accuracy of mean corpuscular hemoglobin concentration (MCHC) detection, Yaël et al.Analyzed blood samples using a decision tree algorithm. The values (rbc o and hgb o) were obtained using a novel fluorescent flow cytometry, and correlation analysis was performed on the hemoglobin and red blood cell levels. The outcomes demonstrated that aberrant MCHC levels could be correctly identified by rbc o and hgb o. Data learning has the potential to enhance biological interpretive advice and facilitate quick follow- up, particularly through the exploration of novel parameters.

II. PROPOSED USE DATA MINING TECHNOLOGIES

A. Bayesian Network

Based on graph theory and probability analysis, a Bayesian network (BN) is a model for expressing and reasoning about uncertain knowledge. A graph of assigned complex causal interaction networks serves as its representation. The arc connecting the variables in the network indicates the direct causal relationship between the events, and each node in the network represents a variable. Because of BN's conditional independence, taking into account the finite variables that are connected to each variable can significantly reduce the difficulty of solving a problem. Its performance is similar to that of decision trees and neural networks in some cases. This method was employed by Lee et al.To forecast the likelihood of radiation pneumonia. They investigated 19 patients out of the 54 nonsmall cell lung cancer patients who had three-dimensional conformal radiation therapy. Serum levels of related biomarkers (transforming growth factor, angiotensin converting enzyme, α 2 microspheres, and interleukin 6) were measured during the first and intermediate stages of treatment. The findings indicated that 200 cases was the ideal number to predict radiation pneumonia using the BN method. Comparing the BN

model to multivariate logistic regression, the optimum performance record area under the receiver operating characteristic curve (AUC) was 0.83, a substantial increase. When used to probabilistic inference on radiation pneumonia, it was clear that the BN methodology offered the flexibility to describe hierarchical interactions between radiation pneumonia factors. Furthermore, patients with type 1 diabetes were predicted to have low glycemic index and hypoglycemia risk using this method by Dain et al. According to the findings, type 1 diabetes patients using insulin pumps had poor metabolic control and a higher risk of hypoglycemia if their daily insulin dose basal rate was higher than 50%. Different tools, nevertheless, such a bolus calculator, short-term modifications to basal rates, and methodical educational instruction, most likely lessened the risk of hypoglycemia. These outcomes gave experience with insulin pump setup.

B. Data Mining Applications in Clinical Medicine:

In the medical field, very large volumes of data are generated, which have a broad prospect in clinical applications. The feasibility of some applications of clinical data mining was explored as described below.

C. Assessing Disease Risks:

The prognosis and therapeutic intervention techniques are significantly impacted by the assessment of disease risk. Clinicians can choose the best course of treatment for a patient by quickly and accurately assessing the patient's condition. The prognosis of diseases and risk variables has intricate relationships. Numerous diseases might share the same risk factors, and multiple risk factors can coexist in a single disease. As a result, there is more correlation than causation in the association between risk factors and diseases. Thus, in order to effectively treat a disease for its tertiary prevention, data mining can be utilized to assess the disease risks and identify the critical characteristics related with the disease prognosis. Scholars in this nation are continually researching this. To assess the risk of unstable angina, Dong et al.[25] created an algorithm based on genetic and fuzzy theory (genetic fuzzy system). Age, sex, blood pressure, creatinine, aspartate transaminase, lactate dehydrogenase, creatine kinase, MB isoenzyme of creatine kinase, Troponin T, and the prior history of cardiovascular disorders were among the relevant indexes of 54 patients that they assessed. When the computed results were contrasted with the opinion of a senior clinician, they were discovered to be extremely consistent. On the other hand, the data mining method produced faster and more convenient results.

D. Practical Drug use Guidance:

When it comes to developing big data mining applications for the medical field, the US has led the way. Its government implemented the findings of big data research—Pillbox is one example of one such product—in the field of medicine. Pillbox is a big data-driven drug advising tool that holds all the details of several medications, such as their color, shape, dosage, and interactions with other medications. This device can accurately tell patients on a wide range of medications, particularly uncommon clinical medications. It can also assist patients in comprehending the properties of medications and provide practical advice on how to use them. Because Pillbox is easy to use, even older people can benefit from it. Pillbox is supposed to make it easier to distinguish between various medications. More significantly, it can lessen adverse effects and aid in understanding how a medicine works. Utilizing data from the World Health Organization VigiBase, Japanese researchers Suzuki et al.[30] found that the frequency of drug-induced liver damage can vary depending on the combination of medicines used. To test for drug hepatotoxicity, they employed acetaminophen, isoniazid, valproic acid, and amoxicillin clavulanic acid. They also assessed the frequency of liver events with these four medications when concomitant. It was concluded that the hepatic safety of drugs can be altered by comedications. The types and mechanisms of experimental medications used in autoimmune myocarditis were compiled by Han et al.,Who also thoroughly examined the benefits and features of different treatment options. The utilization of big data mining technologies can offer significant technological assistance in forecasting the outcomes of experimental medications in clinical settings.

E. Strengthening Medical Management:

Big data mining has a brilliant future in medical management. It can help to analyze medical operation indicators of hospitals for a period of time, including medical information management, medical quality management, medical supplies management, financial management, medical dispute management, doctors work performance, hospital management, decision-making management, hospital resource management, etc., to help hospital administrators provide data support for medical decision-making.

F. Benefiting Evidence-Based Medicine:

Evidence-based medicine is now the primary focus of clinical practice. Medical professionals can look up the most recent research developments for each unique clinical situation to find a solution and prevent experience-related rashness. Thanks to technological advancements in data mining, it is now possible to include substantial amounts of medical research-based information into data warehouses. This facilitates the implementation of more effective search methods and yields more precise search results. The foundation of the Ningbo Wisdom Health Care System project was the evidence-based medicine building initiative. This project's data exchange and sharing is now complete. Eight municipal medical institutions, including Yinzhou, Haishu, Jiangdong, Jiangbei, and Fenghua counties, were docked with the city, urban, and district platforms. This allowed for the collection and exchange of diagnosis, treatment, and health care information, including the majority of patients' electronic medical records and medical data for the entire city. In order to diagnose and treat patients more quickly and precisely, doctors now have a direct practical basis for doing so thanks to the mining of the wealth of health information.

III. CONCLUSION

A new era of easily accessible guidelines and characteristics for a wide range of objects from the quantity of fundamental data is made possible by big data

mining technology. With the unrelenting advancement of science and technology, the field of medical data mining is promising and will grow into a potent tool for the treatment and prevention of human disease, despite the challenges and issues that still need to be resolved in its infancy.

IV. FUTURE PROSPECTS

At present, the theories and technologies of big data mining are being explored for improvement. Each mining theory and technology has its own characteristics and limitations. Fuzzy theory is more frequently studied because it can deal with incomplete data and is suitable for data with a few rules, but the disadvantage is the incompleteness. Rough set theory is simple and does not need transcendent information except data that need to be processed. Its disadvantage is that it cannot deal with continuous attributes. Cloud theory is characterized by virtualization, high extension, and service diversity although its security is questionable. The D-S theory can satisfy conditions that are weaker than Bayesian probability and express uncertainties directly.

CONFLICT OF INTERREST

The authors declare that they have no conflict of interest.

REFERENCES

1. Wu ZJ, Guo Q. Research on the development strategy of China's health management industry in the era of big data. Health Econ Res 2014;6:14-6.
2. Yu GP, Bao XY, Huang XT, Liu W, Xu BB, Yu N, et al. Medical and health big data: Types, characteristics, and relevant issues. J Med Inform 2014;35:9-12. doi: 10.3969/j.issn.1673- 6036.2014.06.002.
3. Kimiafar K, Sadoughi F, Sheikhtaheri A, Sarbaz M. Prioritizing factors influencing nurses' satisfaction with hospital information systems: A fuzzy analytic hierarchy process approach. Comput Inform Nurs 2014;32:174-81. doi: 10.1097/CIN.0000000000000031.
4. Kuo RJ, Huang MH, Cheng WC, Lin CC, Wu YH. Application of a two-stage fuzzy neural network to a prostate cancer prognosis system. Artif Intell Med 2015;63:119-33. doi: 10.1016/j.artmed.2014.12.008.
5. Gil-Herrera E, Tsalatsanis A, Kumar A, Mhaskar R, Miladinovic B, Yalcin A, et al. Identifying homogenous subgroups for individual patient meta-analysis based on rough set theory. Conf Proc IEEE Eng Med Biol Soc 2014;2014:3434-7. doi: 10.1109/EMBC.2014.6944361.
6. Li P, Bi T, Huang J, Li S. Breast cancer early diagnosis based on hybrid strategy. Biomed Mater Eng 2014;24:3397-404. doi: 10.3233/BME-141163.
7. Haque. H. "A Deep Learning Model in the Detection of Alzheimer Disease" Vol.12 No.10 (2021), 4013-4022
8. Liu KH, Xiao LF. New health management model based on cloud platform and big data. J Public Health Prev Med 2014; 25:89-91.
9. Haque. H." Alzheimer's disease: Resting-State Brain Networks and Deep Learning Methods" Issue: 7 | Pages: 15961-15971(2021)
10. Meng XM, Ling PL, Gong XH. Research on techniques and tactics intelligent decision support system of net antagonistic event competitions project. Comput Eng 2012;21:148-52. doi: 10.3969/j.issn.1000- 3428.2012.21.040.