ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 10, Issue 6,June 2024

Research on Disease Prediction Models Based on Big Data

Xuecheng Liu[1]，Yun Li[2]
College of Mathematics and Statistics, Taishan University, 271000, Tai'An, ShanDong, China
Email:controllxc@126.com

***Abstract*—This article takes the clinical diagnosis and treatment data generated by community clinics as the research object, conducts basic review, and then screens, merges, and calculates them. Due to the heterogeneity and incompleteness of medical and health big data in community clinics, data preprocessing operations are required. After data preprocessing, the target dataset is formed and stored on the big data platform. Finally, this article establishes a disease prediction model based on these big data.**

***Index Terms*—big data, prediction models, data cleaning, data collection**

I. Introduction

The current healthcare industry is facing many challenges, including the growing demand for health management among the elderly, a shortage of medical staff, and the complexity of medical process supervision. To address these issues, this article conducted a study on a disease prediction model based on big data [1-4]. This study can achieve early prediction of patient diseases, deeply analyze the causes of diseases, and assist medical staff in better tracking and decision-making of patient conditions, thereby providing strong support for the development of the medical and health industry. To ensure that the data meets the requirements of the model and that each physical examination data is true, complete, and standardized, the following preprocessing operations are performed on the data: data organization, data cleaning, and then data storage and analysis are carried out to provide data support for establishing a disease prediction model based on big data.

II. Data collection

A. Data Source

The data used in this study was sourced from community clinics and health checkups, including diagnostic tables, basic information tables, patient ID tables, physiological indicator tables, medication tables, and other information of the visiting patients. In addition, a treatment record data statistics table is also established, including information on the number of visits, regional distribution, age group statistics, gender statistics, etc.

B. Data Collection

Developing a unified standard for collecting medical basic data, creating a sound mechanism for collecting medical data, and ensuring the integrity of medical basic data, is the only way for the medical basic data collection system to be designed according to standardized principles. From the perspectives of metadata standardization and data exchange standardization, we comprehensively consider the quality control of all collected medical data to ensure its accuracy, completeness, consistency, completeness, effectiveness, and timeliness. Support the expansion of collection sources and data collection items, ensuring that the basic expansion in the later stage will not affect further research in the future.

III. Data cleaning

A. Data Format Transformation

Data format conversion is responsible for converting heterogeneous data formats into a unified standard format in the data center, mainly including conversion definition, completing the definition of the original data format, the definition of the original data source, the definition of the target data format, the definition of the target data, and other configurations. It also completes the data conversion processing according to the rules of the conversion definition, and finally converts the original data into the data format required by the data analysis system. [5] discussed about diabetic retinopathy from retinal pictures utilizing cooperation and information on state of the art sign dealing with and picture preparing. The Pre-Processing stage remedies the lopsided lighting in fundus pictures and furthermore kills the fight in the picture. Although the Disease Classifier step was used to identify arising wounds and other data, the Division stage divides the image into two distinct classes. The methodology for ensuring red spots, exhausting and recognizing evidence of vein-lobby hybrid focuses was also developed in this work, using the hidden data, shape, size, object length to expansiveness distribution as contained in the general fundus picture in the problem area. Besides the Diabetic Retinopathy (DR) analysis, two graphical user interfaces (GUIs) were produced throughout this project.

B. Formatted Data Cleaning

The raw data from the business department is first cleaned to remove invalid data. Then, corresponding data cleaning and filtering rules are configured for future use. Finally, the formatted data is filtered based on the data cleaning and filtering rules to form a basic database. [6] discussed about Nanorobots Control Activation For Stenosed Coronary Occlusion, this paper presents the study of nanorobots control activation for stenosed coronary occlusion, with the practical use of chemical and thermal gradients for biomedical problems. The recent developments on nanotechnology new materials allied with electronics device miniaturization may enable nanorobots for the next few years. New possibilities for medicine are expected with the development of nanorobots. It may help to advance the treatment of a wide number of diseases: cardiovascular problems, neurosurgery, cancer, diabetes and new cell therapies.

In this step, this article establishes a collection log. Based on the medical basic database, statistical analysis is conducted on information data from different collection channels in the medical database, comprehensively evaluating the timeliness, accuracy, legality, security, and other aspects of data collection. This provides statistical data support for the effective innovation of subsequent data collection methods and daily maintenance measures.

IV. Data Storage

A. *Data Resource Management*

This article constructs a unified data resource utilization and management system, which unifies the cataloging, hierarchical maintenance, cross network interaction, secure sharing, global management, and on-demand configuration of independent data resources, achieving interoperability and unified management of information resources.

B. *Data Update Management*

In the process of updating data in the basic database, it is necessary to merge and integrate the newly processed data with the existing data in the basic database according to the set update rules. During the update process, information such as the update time, data volume, operator, data source, and update status should be recorded for later maintenance and tracking.

V. Data Analysis

A. *Data Mining*

Using the cluster segmentation and outlier analysis algorithm in data mining, the collected and integrated high-quality medical and health datasets are subjected to in-depth analysis and computational operations, in order to extract the inherent correlation patterns of the data.

B. *Data calculation*

Using distributed computing to solve the problem of large amounts of medical and health data, saving overall computing time, and greatly improving computing efficiency. On the basis of designing a data analysis method, this article utilizes a distributed computing framework to compute the dataset.

C. *Predictive analysis ability*

Based on data mining algorithms and data calculation results, combined with the advantages of various prediction models, a comprehensive prediction model based on fuzzy integration is established to make judgments on future situations and provide predictive analysis.

VI. Clinical Treatment Plan Decision-Making

By utilizing the big data based disease prediction model proposed in this article, doctors can determine preliminary diagnosis based on the patient's characteristics, the next auxiliary examinations/tests needed, and adopt the most cost-effective treatment plan. When patients seek medical treatment, doctors input patient characteristic parameters into the model, which automatically compares with historical data to draw a series of conclusions. Big data analysis methods are selected for data analysis and processing to calculate the relationship between patients, treatment plans, and treatment outcomes, thereby obtaining the best treatment plan for different patients.

This article establishes a disease prediction model based on big data. By integrating high-quality medical and health datasets, the goal of warning and predicting population disease indicators is achieved. Doctors can also input patient data and compare the analysis and calculation results of the optimal treatment plan selection in the system to obtain the best treatment plan recommendation for the patient.

VII. Acknowledgment

The work is supported by 2022 Tai 'an City Science and Technology Innovation Development Project (Policy Guidance): Research on Disease Prediction Models Based on Big Data (2022NS333).

Xuecheng Liu is a lecturer. He obtained his first degree of Bachelor of Management at the University of Jinan and Master of Science Degree at GuiZhou University. His major fields of study are Network Information Security.

REFERENCES

[1] H.Q. Dong "Research on Data Governance System of Digital Health Industry," *Journal of Modern Information*, vol. 45, pp. 1-19, May 2024.

[2] J.J. Fan, "Research on the Model of Public Data Entering the Data Element Market," *Journal of Information Resources Management*, vol. 14, pp. 68-81, Feb. 2024.

[3] M. Xu, "Analysis of Smart Medical Development Environment in China Based on PEST Model," *Journal of Medical Informatics*, vol. 45, pp. 35-39, Mar 2024.

[4] A.R. Wang, "Exploration of Clinical Specialty Capability Evaluation System Based on Medical Big Data," *Chinese Health Quality Management*, vol. 31, pp. 5-7, Mar 2024.

[5] Christo Ananth, D.R. Denslin Brabin, Jenifer Darling Rosita, "A Deep Learning Approach To Evaluation Of Augmented Evidence Of Diabetic Retinopathy", Turkish Journal of Physiotherapy and Rehabilitation, Volume 32,Issue 3, December 2021,pp. 11813-11817.

[6] Christo Ananth, R.K. Shunmuga Priya, T.Rashmi Anns, S.Kadhirunnisa, "Nanorobots Control Activation For Stenosed Coronary Occlusion", International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE), Volume 2, Special Issue 13, March 2016, pp: 60-76.