Innovative Design of Medical Big Data Platform Integrating Machine Learning and Knowledge Graph

Jun Wang[1(✉)] and Ai-Rong Yu[2]

[1] Nanjing Vocational College of Information Technology, Nanjing 210007, China
intraweb@163.com
[2] The Army Engineering, University of PLA, Nanjing 210007, China

Abstract. With the rapid development of medical information technology, the amount of medical data continues to increase, and the data structure becomes increasingly complex. How to efficiently process and utilize this data to improve the quality and efficiency of medical services has become an important issue. This article proposes an innovative design for a medical big data platform that integrates machine learning and knowledge graph, using large-scale language models and deep learning models to conduct deep analysis and mining of medical text, images, and other data; Adopting a knowledge graph based medical data integration method to build a sustainable medical big data ecosystem. By transforming medical data from different sources and categories into a unified knowledge representation, the integration, storage, management, analysis, and mining of medical data can be achieved. The research results will provide more accurate, faster, and effective data decision-making support for applications such as hospital management, clinical treatment, and scientific research and teaching.

Keywords: Medical big data · Platform Design · Machine Learning · Knowledge Graph

1 Preface

With the rapid development of medical informatization and digitization, the aggregation and utilization of medical big data has become a hot issue in the medical field. In China, due to the large number of patients, medical big data has the characteristics of large-scale and multimodal. However, compared to Western countries, China's massive clinical medical data has not yet been fully utilized in the support of medical AI technology [1]. In this context, medical big data platforms that integrate machine learning and knowledge graph technology are gradually receiving attention. Machine learning technology can help doctors discover potential patterns and trends from a large amount of medical data, improving the accuracy and efficiency of medical diagnosis and treatment. And knowledge graph technology can semantically model and organize knowledge in the medical field, thereby achieving knowledge sharing and reuse. Therefore, combining machine learning and knowledge graph technology can bring broader prospects and

F. Huang et al. (Eds.): ICAIE 2023, AHCS 15, pp. 620–627, 2024.
https://doi.org/10.2991/978-94-6463-242-2_76

deeper impacts to the analysis and application of medical big data. In this context, this article proposes an innovative design scheme for a medical big data platform that integrates machine learning and knowledge graph technology [2, 3]. Based on machine learning technology, massive medical data is analyzed and mined, while knowledge graph technology is used to semantically model and organize knowledge in the medical field, thereby achieving knowledge sharing and reuse. In addition, this article further proposes an adaptive learning mechanism that can continuously optimize and update the knowledge graph based on actual application scenarios, improve the adaptability and flexibility of the platform, and effectively verify the application effect of the platform in the medical field by constructing a multi-layer system of microservices. The research results not only provide a new approach and method for data analysis and application in the medical field, but also have certain promotion and application value. In the future, we will further improve the functionality and performance of the platform, explore more in-depth application scenarios, and make more active contributions to the aggregation and utilization of medical big data [4].

2 Development and Current Situation of Medical Big Data

In recent years, with the continuous and mature development of emerging technologies such as cloud computing, big data, the Internet of Things, mobile internet, and artificial intelligence, the integration of traditional medical industry with these emerging technologies has been accelerated. Among them, the new medical formats represented by health and medical big data have continuously stimulated the development of the medical industry. In order to further promote and standardize the development and application of the healthcare big data industry, in June 2016, the State Council issued the first official document of the healthcare big data industry - "Guiding Opinions on Promoting and Standardizing the Development of Healthcare Big Data Applications", proposing to establish 100 regional clinical medical data demonstration centers by 2020 [5, 6].

Health and medical big data plays an extremely important role in big data. On the one hand, the awakening and deepening of human health awareness have continuously raised the level of demand for medical health, stimulating the deep application of big data technology in the medical field. On the other hand, the popularity of mobile/internet healthcare, automated analysis and detection devices, and wearable devices has made patients, doctors, enterprises, and governments direct creators of data, generating massive amounts of medical data every day, providing a crucial foundation for the development of medical big data.

With the improvement of the living standards of the Chinese people, the increasing aging population, and the strengthening of residents' awareness of health management, the demand for medical and health services in China is constantly increasing, [7] and the massive amount of data is showing explosive and geometric growth. With the increasingly close combination of clinical research and big data, key data technologies such as data lake and master data management will be gradually applied in a large scale, and the integrated integration mode will lead to changes in traditional industry empowerment. Big data is driving the development of the whole medicine.

3 Analysis of the Characteristics of Medical Big Data

Medical big data refers to the massive, diverse, complex, precise, secure, heterogeneous, and closed data generated in the field of healthcare. With the continuous development of information technology and the deepening of medical informatization construction, medical big data has become an important resource in the field of healthcare, with broad application prospects and value. By analyzing medical big data, we found that it has six typical attribute characteristics [8, 9], as shown in Fig. 1.

(1) Massive capacity. The volume of medical big data is enormous, which may come from clinical data, laboratory data, imaging data, genetic data, vital sign data, medical record data, health record data, etc. For example, in the United States, the data volume of the healthcare system reached 150 exabytes in 2011. With the continuous development of Internet of Things technology and the emergence of medical devices, health wearable devices, etc., the monitoring of various vital sign data has become practical and possible, and the increase in data volume is faster.
(2) Diversity. Medical big data is not only vast in quantity, but also diverse in types. These datasets may come from different medical institutions, different medical services, and different medical fields, including diagnosis, treatment, rehabilitation, disease prediction, health management, and other aspects. At the same time, data forms may also include structured, semi-structured and unstructured data, such as medical records, medical orders, images, etc.
(3) Fastness. With the development of data collection methods, the speed of medical big data collection, processing, and analysis has also become faster. For example, during the outbreak of the epidemic, the speed of collecting and analyzing case data is crucial for the control and treatment of the epidemic.
(4) Accuracy. Due to the close relationship between medical data and human health, illness, and life, any mistake may lead to erroneous conclusions. Therefore, it is necessary to ensure the content integrity and constraint integrity of the data during data processing.

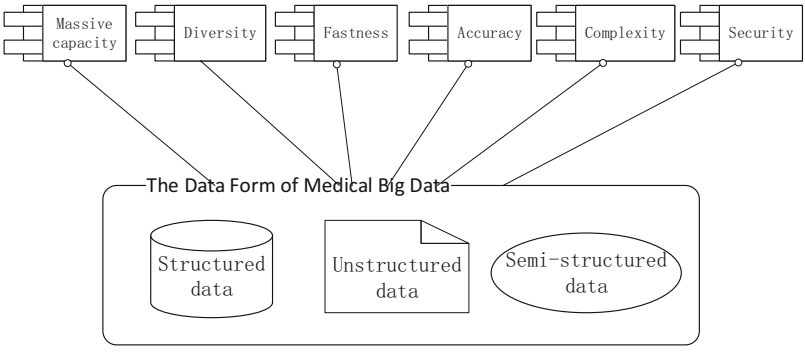

Fig. 1. Analysis of the characteristics of medical big data

(5) Complexity. Due to the fact that the medical field contains a large number of medical professional terms, including over 30000 disease names alone, as well as tens of thousands of diagnostic, surgical, and drug names, as well as a large number of professional terms such as imaging and laboratory examinations, the understanding and application of these terms require the knowledge and skills of medical professionals. On the other hand, there are complex interrelationships between different data types, such as patient genetic data, imaging data, and medical record data, which may be interrelated. Deep data mining and analysis are needed to discover the inherent relationships among them. These pose enormous challenges to the integration and analysis of medical data.
(6) Security. Due to the sensitive data in medical big data, such as the patient's name, age, gender, contact information, medical history, physical examination, etc., the leakage of this information can pose a threat to the patient's privacy and security. Therefore, the security of medical big data is one of its very important attributes [10, 11].

4 Design of a Platform Framework Integrating Machine Learning and Knowledge Graph

4.1 Functional Design

In response to the rapid growth of medical data and the rapid development of information technology, a medical big data platform integrating machine learning and knowledge graph has been constructed. By using large-scale language models and deep learning models to deeply analyze and mine medical text, images, and other data, and combining knowledge graph based medical data integration methods, a sustainable medical big data ecosystem has been constructed. The specific functions are shown in Fig. 2.

4.1.1 Data Storage and Management

Data storage and management utilize big data technology to efficiently store and manage data. First, the data is divided into structured data and unstructured data, which are stored in relational database and distributed file system respectively. Secondly, achieve fast query and retrieval of data, supporting multiple data formats such as CSV, JSON, etc.

4.1.2 Data Collection and Preprocessing

This function is responsible for collecting medical data from multiple data sources, and the collected data is cleaned and standardized to meet the needs of subsequent analysis. For data involving patient privacy, desensitization technology is used for processing.

4.1.3 Knowledge Graph Construction and Reasoning

The knowledge graph construction and inference module includes sub modules such as entity extraction, attribute extraction, relationship extraction, and knowledge representation. Through natural language processing technology, entities, attributes and

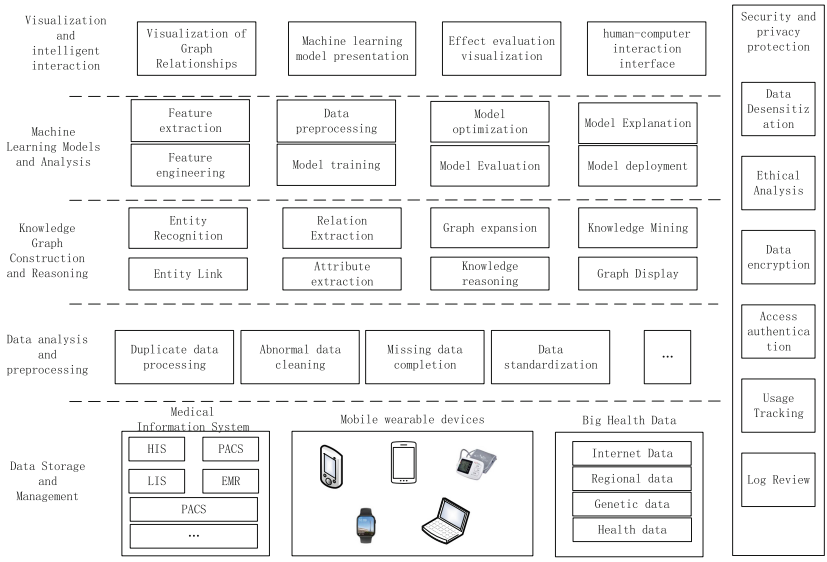

Fig. 2. Main function division of the platform

relationships are extracted from medical data, and knowledge maps are constructed using knowledge representation technology. Based on the knowledge graph, association analysis and inference can be conducted to explore potential knowledge.

4.1.4 Machine Learning Models and Analysis

Machine learning models and analysis are responsible for intelligent analysis of medical data. According to different analysis tasks, machine learning algorithms such as prediction, classification, and clustering can be used. In addition, for complex problems, deep learning techniques such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Long Short Term Memory Networks (LSTM) can be used. Through machine learning and deep learning technologies, efficient mining and analysis of medical data can be achieved.

4.1.5 Visualization and Intelligent Interaction

The visualization and intelligent interaction module provides a user-friendly interface to display data, knowledge graphs, and analysis results. Users can quickly understand the data situation, knowledge structure, and analysis results through an intuitive visual interface. At the same time, it supports users to obtain the required information through intelligent interactive methods, such as through question answering systems, voice recognition, and other methods.

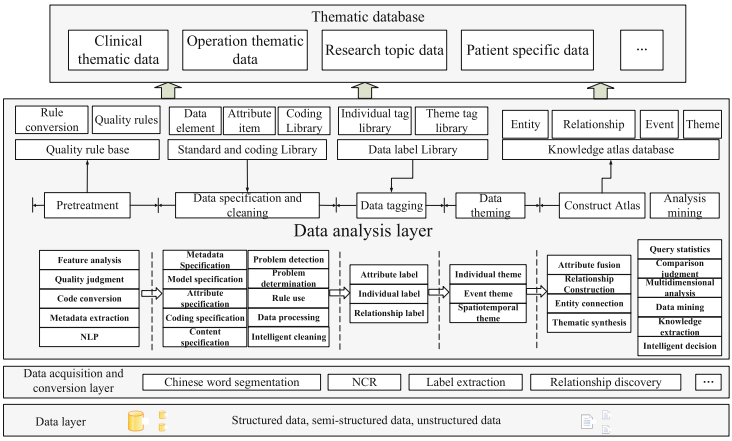

Fig. 3. Platform Technology System Architecture Design

4.1.6 Security and Privacy Protection

The security and privacy protection module is responsible for ensuring the security of data during storage, processing, and transmission, and protecting patient privacy. Specifically, it includes functions such as data encryption, access control, and auditing. Data encryption technology can ensure the security of data during transmission, access control can limit the access rights of different users to data, and audit function can record user operations to ensure data security and traceability.

4.2 Technical System Design

Its technical system is built using a microservice architecture, as shown in Fig. 3.

The system closely combines the data collection, analysis and mining layer with the upper service application, and realizes the data extraction, loading and crawling of different categories, sources and targets in the data layer. It needs to provide support for mainstream databases and provide effective parsing means for unstructured files such as JPG, CSV, XML and HTML. The natural language processing layer realizes the digital processing of electronic medical record data. In the analysis and mining layer, the combination of machine learning model and expert experience decision-making is used to deal with the problems of data standardization, data quality, information integrity, relationship network construction and so on. The updated data forms different subject databases according to the purpose and organization form to provide data services for clinical management, operation decision-making, scientific research analysis, patient portrait and other functions.

5 Conclusion

At present, the rapid accumulation of medical big data provides enormous opportunities for the healthcare industry, while also bringing challenges in data security, data sharing, and data integration analysis. In response to these challenges, artificial intelligence

technologies such as machine learning and knowledge graph have brought hope for the effective utilization of medical big data. Machine learning technology can achieve automatic learning and prediction based on large-scale medical data, enabling personalized diagnosis and treatment plan formulation. Knowledge graph can achieve medical knowledge inference and new knowledge discovery through relationship prediction and path search. Combining machine learning and knowledge graph technology with medical big data can achieve more accurate medical diagnosis, personalized treatment plan formulation, and deeper understanding of disease mechanisms. In the future, with the further accumulation of data and the improvement of algorithm models, medical big data analysis platforms based on machine learning and knowledge graphs are expected to have a greater impact and truly achieve intelligent healthcare. However, there are also issues such as data privacy protection and artificial intelligence bias, which require the joint improvement and protection of relevant legal systems and technical levels.

Acknowledgments. This work was supported by the high-level talent research initiation fund project of Nanjing Vocational College of Information Technology (YB20221502) and the Industry Guidance Committee project of the Ministry of Industry and Information Technology (GXHZWZ13058).

References

1. Alhussain, T. (2018). Medical Big Data Analysis Using Big Data Tools and Methods. JOURNAL OF MEDICAL *IMAGING AND HEALTH INFORMATICS*, 8(4), 793–795.
2. Chen, Q., & Wang, W. (2021). Analysis on the Application of Big Data Technology in Medical and Health Industry. *Journal of Physics: Conference Series*, 1883, 12135–12136.
3. Dingkun, L., Zhou, Y., Li, L., Xiaolin, W., Bifei, Q., & Yaning, L. (2019). *Practical Data Mid-Platform Design and Implementation for Medical Big Data.*
4. Khan, F., Prasad, B. V. V. S., Syed, S. A., Ashraf, I., & Ramasamy, L. K. (2022). An Efficient, Ensemble-Based Classification Framework for Big Medical Data. *Big Data*, 10(2), 151–160.
5. Li, G., Liu, Y., Zhao, H., Cai, H., & IEEE. (2018). Research on Application of Healthcare Data in Big Data Era *2018 INTERNATIONAL CONFERENCE ON ROBOTS & INTELLIGENT SYSTEM (ICRIS 2018)* (377–379).
6. Li, M., Wang, C., Yan, L., Wei, S., & IEEE. (2019). Research on the Application of Medical Big Data *14TH INTERNATIONAL CONFERENCE ON COMPUTER SCIENCE AND EDUCATION (ICCSE 2019)* (478–482).
7. Liu, J., Zhang, Y., Xing, C., & IEEE. (2017). Medical Big Data Web Service Management Platform *2017 11TH IEEE INTERNATIONAL CONFERENCE ON SEMANTIC COMPUTING (ICSC)* (316–321).
8. Shah, F., Li, J., Shah, Y., Shah, F., & IEEE. (2017). Broad Big Data Domain via Medical Big Data *2017 4TH INTERNATIONAL CONFERENCE ON SYSTEMS AND INFORMATICS (ICSAI)* (732–737).
9. Wang, Y., & Ren, S. (2017). Survey on Visualization of Medical Big Data. *Journal of Frontiers of Computer Science & Technology*, 11(5), 681–699.
10. Zhang, Z., Yang, Z., Huang, Y., & Zhan, J. (2021). Big Medical Data and Medical AI Standards: Status Quo, Opportunities and Challenges. *Medical Journal of Peking Union Medical College Hospital*, 12(5), 614–620.

11. Zhao, H., Li, G., Feng, W., & IEEE. (2018). Research on Visualization and Application of Medical Big Data *2018 INTERNATIONAL CONFERENCE ON ROBOTS & INTELLIGENT SYSTEM (ICRIS 2018)* (383–386). International Conference on Robotics and Intelligent System (ICRIS).