Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 3, pp. 1422–1428, DOI 10.12694/scpe.v26i3.4158

INNOVATION OF PRECISION MEDICAL SERVICE MODEL DRIVEN BY BIG DATA

FUJUN WAN[*] XINGYAO ZHOU[†] CHONGBAO REN[‡] AND YUCHEN ZHANG[§]

Abstract. This paper proposes a precision medical service system driven by big data. The PCA-GRA-BK algorithm, which combines principal component analysis (PCA), grey association analysis (GRA) and Bayesian classifier (BK), is adopted. The algorithm extracts critical information from massive medical data, identifies patient characteristics, predicts disease risk, and provides personalized treatment plans. First, the system uses PCA technology to reduce the dimensionality of the original medical data and extract the most representative principal components to reduce data redundancy and retain critical information. Then GRA method was used to analyze the correlation between different medical indicators to determine the main factors affecting health status. Finally, the BK algorithm updates the probability model based on prior knowledge and current data to predict patients' disease risk accurately. A simulation modeling environment is constructed and the PCA-GRA-BK algorithm is tested in this environment to verify the effectiveness of the system. The experimental results show that the algorithm has excellent performance in the accuracy of disease prediction and personalized treatment recommendation. Compared with traditional medical decision support systems, this system has shown significant advantages in extensive data processing capabilities and precision medical services.

Key words: Big data-driven; Precision medicine; PCA-GRA-BK algorithm; Simulation modeling; Personalized treatment.

1. Introduction. In the wave of the digital age, the healthcare field is undergoing an unprecedented transformation. The rise of big data technology has supported the realization of precision medical services. The core concept of precision medicine is individual differences, which emphasize the development of personalized prevention and treatment strategies based on a patient's genetic background, lifestyle and environmental factors. The practice of this concept is inseparable from efficient data processing and analysis technology [1].

In recent years, principal component analysis (PCA), a standard data dimensionality reduction method, has been widely used in the pre-processing stage of medical big data. PCA can transform multiple variables into a few comprehensive variables, thus simplifying the data structure and improving the efficiency of subsequent analysis [2]. Grey correlation analysis (GRA), on the other hand, shows unique advantages in dealing with small samples and uncertainties. By calculating the correlation degree among various factors, GRA reveals the key factors that significantly impact the target [3]. Bayes classifier (BK) is a classification method based on probability statistics. The Bayes classifier (BK) can constantly update the model according to existing data to improve prediction accuracy [3]. The organic combination of these three algorithms to form the PCA-GRA-BK algorithm is expected to provide a comprehensive and efficient set of analytical tools for precision medicine.

Domestic and foreign scholars have made some progress in researching precision medical service systems. Literature [4] proposes a disease prediction model based on deep learning, which can extract features from electronic medical records to achieve early diagnosis of chronic diseases. Literature [5] developed a personalized drug recommendation system based on cloud computing, which used patients' historical medication data to recommend the most appropriate drug combinations. However, most of these studies focus on applying a single algorithm or technology, and lack consideration of the comprehensive performance of the entire precision medical service system.

The research content of this paper aims to build an extensive data-driven precision medical service system based on the PCA-GRA-BK algorithm. First, the paper will elaborate on the principles and steps of the PCA-GRA-BK algorithm and the advantages of their application in medical data processing [6]. Secondly, the

[*]China National Institute of Standardization, Beijing 100000, China
[†]China National Institute of Standardization, Beijing 100000, China (Corresponding author, 189310283930163.com)
[‡]China Special Equipment Inspection & Research Institute, Beijing 100000, China
[§]China National Institute of Standardization, Beijing 100000, China

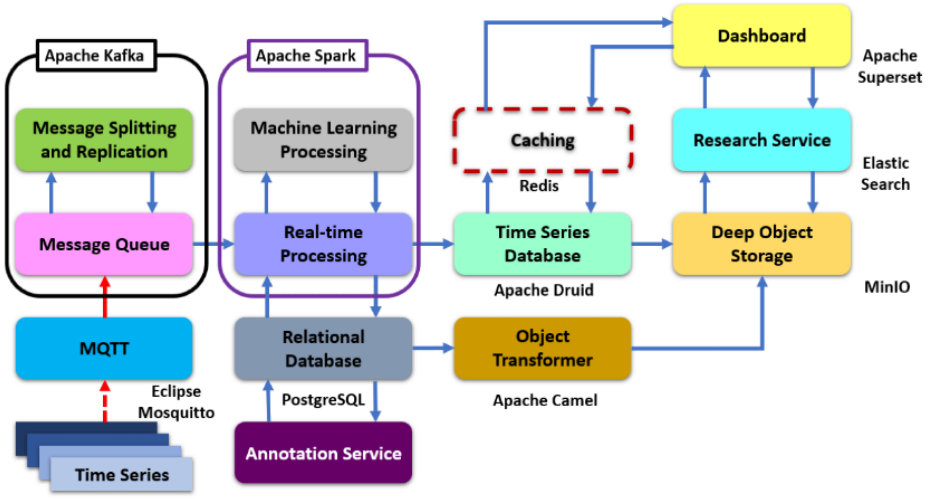

Fig. 2.1: Architecture diagram of distributed cache system for medical big data.

algorithm's performance in the precision medical service system is evaluated through simulation modeling. This paper will simulate different medical scenarios, collect and analyze medical data of virtual patients, and verify the effectiveness of the PCA-GRA-BK algorithm in disease prediction and treatment plan recommendation. Finally, this paper will discuss the potential application of this system in the actual medical environment, as well as the challenges and future development direction.

2. Smart medical data management under big data.

2.1. Research on medical data collection cleaning rules, data warehouse and interface standards. There are many types of medical big data. They exist in relational databases, as well as in unstructured and unstructured parts. Scholars can update and access it in real-time [7]. In addition, due to the characteristics of the medical field itself, it needs to update the status of the data in real time as time goes by. Then, the status of the data ontology is judged in real-time so that all kinds of data can be saved in a unified form. This provides efficient data analysis for both senior and back-office staff. This paper will build an ETL model based on Sqoop. Unified structured data migration between the relational database and the Hadoop platform [8]. Secondly, it uses semi-structured data and unstructured data transfer functions on Hadoop.

A general data interface standard for external applications is proposed. It includes interface format, language, load balancing design, etc. This project plans to develop an intelligent medical cloud computing platform based on Hadoop [9]. In addition, the project adds medical information processing components to improve the operating efficiency of the cloud computing platform. In this way, the real-time acquisition of user physiological parameters, reasonable allocation of resources and directional analysis of display results are realized. Regarding data storage, this paper presents a distributed file management system and a cache database structure. Then, the traditional relational database is supported.

2.2. Design of extensive medical data warehousing and management system.. The storage and management model of medical big data should also meet data warehouse requirements [10]. Then, build a topic-oriented, integrated, changeable and decision-making data warehouse. A distributed Redis architecture based on Zoo Keeper is proposed (Figure 2.1). FIG. 2.1 shows the medical image data processing method. The design is carried out with hierarchical thinking. This architecture divides the overall architecture into two levels. The data layer mainly deals with the specific Redis database and completes the packaging processing of medical services and medical data in the service layer. ZooKeeper is a highly stable performance that ensures efficient cluster load balancing [11]. Zoo Keeper configures multiple backup nodes for Redis hosts. Multiple backups are performed to the Redis host via the Redis backup device to ensure the availability of the Redis

cluster. Obtain Redis host slice information from ZooKeeper. The routing method is constructed to solve the problem in the Redis cluster.

3. Medical service data mining analysis and decision-making information service system. The prediction model, association model and service model are studied based on the data model of distributed cache. Integrate it with the needs of intelligent health services to extract practical information from massive data. The scale of data accepted by the system during operation and maintenance is tons, with the rapid data growth [12]. The method studied in this subject can provide an early warning model for developing future diseases and provide a scientific basis for government departments and medical institutions. This project will be based on significant data architecture and medical subject data. They use cutting-edge technologies such as core performance index analysis, cluster gap analysis, data multidimensional analysis, data report analysis, data instrument analysis, etc., for extensive health data analysis and decision support. This project presents a technique for fast storage, indexing and querying massive data. With the continuous expansion of data scale, efficient data storage, indexing and query have become the core problems of data warehousing, and the solution of these problems depends on good data organization and optimization algorithms. A suitable query method is critical to a database. This paper uses collaborative filtering technology to analyze and manage the medical big data stored in the data warehouse. A hospital personalized service platform is constructed using the HL7 communication protocol [13]. The HL7 recommendation message connects the platform with other related software platforms in the hospital.

4. Smart health and medical extensive data display system. In innovative medicine, whether it is patients, doctors, or managers, they want to be able to present the valuable information hidden in big data. In this paper, function mining in intelligent health systems is studied. (1) Enable the hospital to promptly grasp the current medical development trend and adjust the indicators promptly. The medical big data cloud platform enables all hospitals to share medical resources. At the same time, information exchange and collaborative sharing can be conducted promptly [14]. It enables users to obtain service perspectives at multiple levels to achieve the diversity of service models. In addition, the data available on the platform can be used to integrate parts of the medical business. Establish a new service model to save operating costs. (2) Analyze various reports, charts and analysis results. In this way, decisions are made and implemented according to the needs of decision-makers and government staff. (3) Patients conduct a comprehensive analysis and prediction of their case data. Patients can choose the appropriate doctor to consult and get guidance, treatment and reference on the platform. (4) Physicians can also evaluate patients based on patient and platform information. In this way, a personalized treatment plan is developed for the patient.

The ultimate goal of intelligent health management supported by big data is to achieve good interaction on the platform. This paper uses the Zoo Keeper distributed cache framework to build a data representation system for an intelligent health system. Reasonable query for user needs. This project builds a message subscription mechanism based on distributed distribution to meet the diverse processing needs of medical big data [15]. For the different data sources of medical information available, A medical Data processing strategy combining offline batch processing and online real-time computing is proposed (4.1 Cited in How can Big Data Analytics Support People-Centred and Integrated Health Services: A Scoping Review). This project intends to adopt data hierarchical and shunt methods to prolong the data calculation process to minimize the time delay of massive and complex medical big data processing. Then, it is divided into three steps: log parsing, product distribution and new operation. Flink technology and SparkStreaming technology are used in data collection. Flink is a kind of offline parallel stream data processing technology with high throughput and low latency, which is well adapted to the initial log analysis characteristics of medical big data. SparkStreaming is a micro-batch processing method. It can divide the incoming real-time data into several small batches to ensure a stable response during the newly added operations. The added computing delay is minimized to make full use of the efficiency of the computing engine.

5. Gray correlation analysis. Each index's correlation degree is obtained using the grey correlation degree method. In this way, the comprehensive level of each index is constructed. The process is as follows.

1). In studying the grey correlation degree, people must first find the reference sequence reflecting the system's characteristics [16]. Secondly, it is necessary to find out the contrast sequence essential to the whole

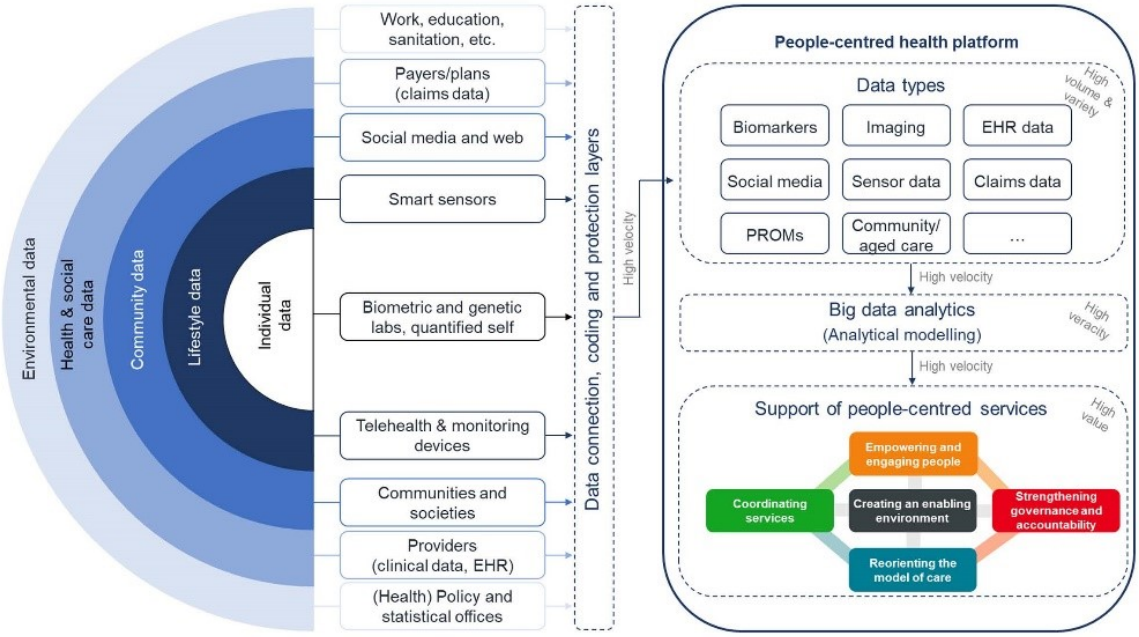

Fig. 4.1: Medical big data processing strategy.

system. A set of data that can reflect the characteristics of a system is called a reference sequence, which is used to measure and compare the degree of correlation between the sequences. Sensitivity is usually treated as A reference sequence $F = F(t) \mid 1, 2, \cdots, n.F(t)$ represents the number that corresponds to the reference sequence. A series of factors is called a relative sequence. It can be expressed in terms of $G_i = G_i(t) \mid t = 1, 2, \cdots, n, i = 1, 2, \cdots, m.g_i(t)$ represents the t th value in the i th comparison series, and n represents the number of QI features.

2). Because the measurement units are not necessarily the same, it is necessary to average them first when conducting gray correlation analysis. Take the average value of each value in the sequence so that the processed data value is close to the order of 1 .

$$g_i'(t) = \frac{g_i(t)}{g_i} \tag{5.1}$$

$\bar{g}_i$ is the average of series i. $g_i(t)$ means that the t data in the i order is averaged. $g_i(t)$ is the t data after the average processing of the i data.

3). the quantity difference between the various data is reduced in the standardization process. This makes the calculation of the grey correlation coefficient more convenient. Its expression is as follows:

$$\lambda_i(t) = \frac{\min_i \min_t |y(t) - g_i(t)| + \delta \cdot \max_i \max_t |y(t) - g_i(t)|}{|y(t) - g_i(t)| + \delta \min_i \min_t |y(t) - g_i(t)|} \tag{5.2}$$

$|y(t) - g_i(t)|$ is the distance between the reference sequence and the corresponding t data in the i contrast sequence, where max is the most significant distance and min is the shortest distance. δ is also known as the distinguishing factor, and the value range of δ is usually (o,1).

4). The correlation factor reflects the degree of correlation between each comparison sequence and the baseline sequence [17]. When measuring correlation degree, the correlation coefficient should be used to calculate the average value of different time points. Relevance r_i is calculated as follows:

$$SSE = \sum_{i=1}^{t} \sum_{q \in Z_i} |q - n_i|^2 \tag{5.3}$$

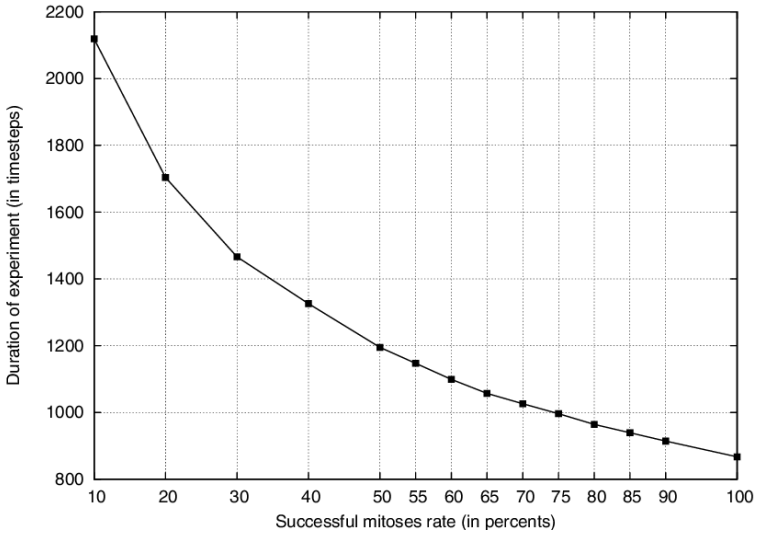

Fig. 7.1: Elbow chart.

When the desired correlation degree is closer to the same time, the higher the correlation degree between the sequences, the higher the correlation degree between the identifier property and the sensitive data.

6. Grey Correlation analysis of medical big data (PCA-GRA-BK algorithm). This project studies the PCA-GRA-BK method based on privacy and validity. The GRA method was used to evaluate each quality index's correlation degree to determine the comprehensive level applicable to each index [18]. The quality information with maximum value must be selected for processing in data collection, thus extending the universality of quality information. First, it must be divided into categories to prevent the K-type anonymous system from being too general. An adaptive method is used to select the number of categories and K-anonymize them. This improves the efficiency of classification and reduces the packet loss rate. Records satisfying K anonymizers are selected from set T and added to K hidden tables [19]. The value of the optimal class n represents the number of clusters of a class. The sample set K is divided into n samples, and the maximum quasi-recognition item of the sample set is found.

7. Experimental analysis. This project intends to use the elbow method to determine the optimal number of classes. By calculating the sum of error squares (SSE) between classes, class families and SSE values are taken as coordinates by points [20]. The number of optimal clusters is determined based on the value of the inflection point closest to the shape of the elbow.

$$SSE = \sum_{i=1}^{t} \sum_{q \in Z_i} |q - n_i|^2 \tag{7.1}$$

Z_i is the ith cluster, q is the sampling point in Z_i, and n_i is the average of all samples on Z_i. According to point A's coordinates, the elbow stroke method is used to compare the test results and get a reasonable number of clusters. It can be seen from Figure 7.1 that when the number of groups is 6 , the judgment criterion of the elbow method is satisfied, so the optimal number of classes in this data set is set to 6. Then this paper carries out cluster analysis based on the optimal class size. The performance of Datafly, PCA-GRA Datafly, PCA-GRA-KK and PCA-GRA-KK algorithms is compared and analyzed. The amount of information is an important index to evaluate the algorithm's performance, and the loss degree of this index is small, indicating that a lot of original data is lost in the system [21]. It has a high use efficiency. The calculation method of

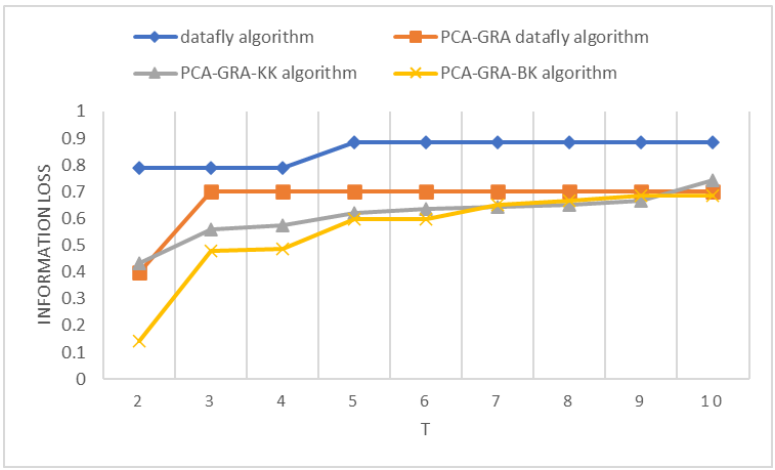

Fig. 7.2: Graph of changes in information loss rate.

information loss is given:

$$IL(RT) = \frac{\sum_{i=1}^{n_q} \sum_{j=1}^{n} \frac{L}{|DHG_A|} + \sum_{t=1}^{n} |k[Q]|}{|K| \times n_q} \tag{7.2}$$

K is the initial data table, RT is the anonymized data table, n is the number of tuples included in the data table, n_q is the number of quasi-identifier features, L is the number of times that the class j identifier of the i record is promoted in the generalization tree, $|DHG_A|$ is the height of the generalization tree of feature A, $|k[Q]|$ is the value of the class identifier Q in the record. The PCA-GRA-KK method is combined with the PCA-GRA-BK method to cluster groups with high similarity. Then, local generalization and K-anonymization are performed to reduce the data loss caused by the overall generalization. The PCA-GRA-BK method is used to optimize the clustering, thus reducing the packet loss rate of the grouping. The critical problem in anonymizing medical data is preserving the original data's information as much as possible. The results showed the best PCAGRA-BK method (Figure 7.2).

8. Conclusion. This study successfully constructed an extensive data-driven precision medical service system based on the PCA-GRA-BK attracted. By integrating principal component analysis (PCA), grey association analysis (GRA) and Bayesian classifier (BK), the system realized rapid processing and accurate analysis of large-scale medical data. The system has shown good performance and accuracy in disease risk prediction and personalized treatment plan recommendation through simulation modeling. The experimental results show that the PCA-GRA-BK algorithm can effectively extract critical information from complex and changeable medical data, identify the core factors affecting health, and constantly improve the accuracy of disease prediction through the dynamic updating characteristics of the Bayes classifier. In addition, the personalized treatment recommendation function of the system can provide more suitable treatment plans based on considering the specific situation of patients, thus improving the pertinence and effectiveness of medical services. However, although this research has achieved positive results, it still needs to face challenges in practical applications such as data privacy protection, algorithm transparency, and system stability. Future research efforts will further optimize the algorithm, strengthen data security and privacy protection measures, and explore more clinical application scenarios to ensure the system's sustainable development and broad application.

REFERENCES

[1] Dash, S., Shakyawar, S. K., Sharma, M., & Kaushik, S. (2019). Big data in healthcare: management, analysis and future prospects. Journal of big data, 6(1), 1-25.

[2] Anshari, M., Almunawar, M. N., Lim, S. A., & Al-Mudimigh, A. (2019). Customer relationship management and big data enabled: Personalization & customization of services. Applied Computing and Informatics, 15(2), 94-101.
[3] Singh, R. P., Javaid, M., Haleem, A., & Suman, R. (2020). Internet of things (IoT) applications to fight against COVID-19 pandemic. Diabetes & Metabolic Syndrome: Clinical Research & Reviews, 14(4), 521-524.
[4] Vaishya, R., Javaid, M., Khan, I. H., & Haleem, A. (2020). Artificial Intelligence (AI) applications for COVID-19 pandemic. Diabetes & Metabolic Syndrome: Clinical Research & Reviews, 14(4), 337-339.
[5] Roh, Y., Heo, G., & Whang, S. E. (2019). A survey on data collection for machine learning: a big data-ai integration perspective. IEEE Transactions on Knowledge and Data Engineering, 33(4), 1328-1347.
[6] Sreenu, G. S. D. M. A., & Durai, S. (2019). Intelligent video surveillance: a review through deep learning techniques for crowd analysis. Journal of Big Data, 6(1), 1-27.
[7] Vellido, A. (2020). The importance of interpretability and visualization in machine learning for applications in medicine and health care. Neural computing and applications, 32(24), 18069-18083.
[8] Liu, H., Ong, Y. S., Shen, X., & Cai, J. (2020). When Gaussian process meets big data: A review of scalable GPs. IEEE transactions on neural networks and learning systems, 31(11), 4405-4423.
[9] Mikalef, P., Boura, M., Lekakos, G., & Krogstie, J. (2019). Big data analytics capabilities and innovation: the mediating role of dynamic capabilities and moderating effect of the environment. British journal of management, 30(2), 272-298.
[10] Kumar, S., Tiwari, P., & Zymbler, M. (2019). Internet of Things is a revolutionary approach for future technology enhancement: a review. Journal of Big data, 6(1), 1-21.
[11] Yadav, S. S., & Jadhav, S. M. (2019). Deep convolutional neural network based medical image classification for disease diagnosis. Journal of Big data, 6(1), 1-18.
[12] Tian, S., Yang, W., Le Grange, J. M., Wang, P., Huang, W., & Ye, Z. (2019). Smart healthcare: making medical care more intelligent. Global Health Journal, 3(3), 62-65.
[13] Hariri, R. H., Fredericks, E. M., & Bowers, K. M. (2019). Uncertainty in big data analytics: survey, opportunities, and challenges. Journal of Big data, 6(1), 1-16.
[14] Javaid, M., & Khan, I. H. (2021). Internet of Things (IoT) enabled healthcare helps to take the challenges of COVID-19 Pandemic. Journal of oral biology and craniofacial research, 11(2), 209-214.
[15] Price, W. N., & Cohen, I. G. (2019). Privacy in the age of medical big data. Nature medicine, 25(1), 37-43.
[16] Qadri, Y. A., Nauman, A., Zikria, Y. B., Vasilakos, A. V., & Kim, S. W. (2020). The future of healthcare internet of things: a survey of emerging technologies. IEEE Communications Surveys & Tutorials, 22(2), 1121-1167.
[17] Niebel, T., Rasel, F., & Viete, S. (2019). BIG data–BIG gains? Understanding the link between big data analytics and innovation. Economics of Innovation and New Technology, 28(3), 296-316.
[18] Shen, M., Deng, Y., Zhu, L., Du, X., & Guizani, N. (2019). Privacy-preserving image retrieval for medical IoT systems: A blockchain-based approach. Ieee Network, 33(5), 27-33.
[19] Alaei, A. R., Becken, S., & Stantic, B. (2019). Sentiment analysis in tourism: capitalizing on big data. Journal of travel research, 58(2), 175-191.
[20] Lundervold, A. S., & Lundervold, A. (2019). An overview of deep learning in medical imaging focusing on MRI. Zeitschrift für Medizinische Physik, 29(2), 102-127.
[21] Santos, M. K., Ferreira, J. R., Wada, D. T., Tenório, A. P. M., Nogueira-Barbosa, M. H., & Marques, P. M. D. A. (2019). Artificial intelligence, machine learning, computer-aided diagnosis, and radiomics: advances in imaging towards to precision medicine. Radiologia brasileira, 52(06), 387-396.

Edited by: Hailong Li
Special issue on: Deep Learning in Healthcare
Received: Jun 21, 2024
Accepted: Jul 25, 2024