Proceedings of the 2nd International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/74/20240458

Review of the data science in the field of healthcare

Weiye Hu

Faculty of Science, University of Waterloo, Waterloo, Ontario, Canada, N2L 3G1

W54hu@uwaterloo.ca

Abstract. The world is entering the digital age, where advancements in computer technology have resulted in the emergence of data-driven applications in the healthcare sector. Data science is a multidisciplinary domain that employs scientific methodologies such as data mining techniques, machine learning algorithms, and big data to derive knowledge and insights from many types of structured and unstructured data. The healthcare business produces extensive datasets containing valuable information regarding patient demographics, treatment regimens, and tumor sizes. This study will examine the procedures of data purification, data mining, data preparation, and data analysis employed in healthcare applications, using methods of literature review and analysis. Efficiently managing and analyzing data can greatly benefit the healthcare industry. By leveraging data science, the application of data-driven decision-making in healthcare holds the potential to enhance the quality of healthcare services.

Keywords: tumor, data science, big data, healthcare industry

1. Introduction

The advent of the digital era has brought to the emergence of data-driven applications in the healthcare sector, thanks to the advancements in computer technology [1]. Various types of data, such as Electronic Health Records (EHR) containing patient information, prescriptions, clinical reports, and tumor size statistics, can be overwhelming for doctors to remember. Even a minor oversight might result in a significant error [2]. The program, equipped with cutting-edge technology such as machine-learning algorithms and data analysis, has been developed to assist doctors in making informed decisions. The advent of data science has the potential to significantly enhance the quality of patient care [2]. Data science has been extensively applied in various areas within the healthcare sector. A concise overview of the entire data science process, including data cleansing, data mining, data preparation, and data analysis, will be provided.

The study will employ a literature review method to assess the feasibility of utilizing data science in the healthcare industry. It will provide a concise overview of the data science process within the healthcare profession, as well as outline the advantages of its applications. This review aims to enhance people's comprehension of the concept of data science and its various uses in the field of healthcare. Through enhancements in every facet, the program will progressively become more dependable, enabling doctors to make increasingly appropriate decisions.

© 2024 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/74/20240458

2. Data processing

2.1. Data Cleansing

Data is of utmost importance in the field of data science, owing to the increasing focus on data-driven decision-making in business. The significance of gathering accurate and exact information for decision-making has garnered significant attention. The decisions made in data science are heavily reliant on the accuracy of the data, making data correctness and conciseness the most crucial aspects. The data received from many sources is disorganized, containing a significant amount of irrelevant information that is not aligned with our desired forecast. Consequently, this will have a detrimental impact on the accuracy of the prediction outcome [4]. According to a survey done by "Price Waterhouse Coopers" in 2001, 75% of 599 organizations experienced financial losses as a result of data quality issues [5]. Data cleansing is an essential step for data scientists to get a high-quality dataset that is easily analyzable. Data quality refers to the degree to which data aligns with the specific needs and requirements of an organization. Attaining high-quality data necessitates the combined efforts of individuals, technology, and procedures; it cannot be attained through natural means alone. In the absence of effective quality management, even a minor error can result in significant financial loss, operational inefficiency, and failure to meet industry and government rules [6]. Data quality and data cleansing are closely intertwined. Data cleansing is a procedure that is performed on existing data to eliminate errors and enhance its suitability [7]. This involves identifying, detecting, and correcting errors through a prescribed process. By doing so, the data becomes prepared for analysis, ensuring its readiness for use by predictive models [4].

Data cleansing consists of five phases: data analysis, development of transformation procedure and mapping rule, verification, transformation, and backflow of cleaned data. The initial stage of data cleansing involves evaluating the data to detect errors and inconsistencies. This phase, also known as data auditing, aims to identify all forms of abnormalities. The transformation workflow involves a sequence of procedures aimed at identifying and removing anomalies in order to obtain information regarding the anomalies present [4]. Subsequently, it is imperative to conduct a verification process to ensure that all errors have been rectified. Following this, it is necessary to repeat the initial and secondary steps to confirm that all the corrupted data has been substituted with the cleansed data [4].

2.2. Data mining

Once the necessity of the acquired data has been confirmed, it is imperative to proceed with the subsequent data processing, namely data mining. Data mining is the systematic exploration and analysis of a substantial amount of unprocessed data with the purpose of discovering patterns and extracting valuable insights [8]. Data mining techniques encompass many methods such as data grouping, data clustering, data correlation, sequential pattern mining, regression, and data storage [2]. Data mining comprises six distinct phases. Prior to processing any data, it is essential to have a clear understanding of the goals. What ultimate accomplishment should be attained with data mining [5]. Next, it is necessary to comprehend and organize the gathered data, which involves understanding the data's source, ensuring its safety, and considering how these limitations will impact the data mining procedure. Subsequently, the data must be cleansed to attain the desired objectives. Utilizing machine learning technology, we may analyze the data to identify relationships, patterns, and correlations. This analysis can also enable us to make predictions about future outcomes based on the data. The findings will be assessed ultimately [8]. Change and monitoring might be implemented during the data mining process when the acquired data is deemed unrelated to the goals. By uncovering concealed data and patterns, it will enhance profitability and efficiency [8]. The process is typically characterized by strictness and organization, since it explicitly recognizes a problem, collects data pertaining to the problem, and aims to develop a solution.

Common technologies in data mining include K-Nearest Neighbor (KNN), Neural Networks, Decision Trees, and others. KNN is an algorithm that classifies data by considering its closeness to other data points. The non-parametric, supervised technique is employed to forecast the characteristics

DOI: 10.54254/2755-2721/74/20240458

of a group by analyzing individual data points [8]. Neural networks employ nodes to process data. These nodes consist of inputs, weights, and an output. Data is processed via supervised learning, which mimics the interconnectedness of the human brain. This model can be configured to assign threshold values for determining the accuracy of a model. Decision trees are utilized for the purpose of categorizing or forecasting a result by using a predefined list of criteria or decisions. A decision tree is a computational model that employs a sequential collection of inquiries to categorize a dataset according to the provided answers [8].

2.3. Applications of data analytics in healthcare fields

In recent years, there are six areas of analytics which developed well in healthcare field including disease surveillance, health care management and administration, privacy protection and fraud detection, mental health, public health, and pharmacovigilance. But in this paper, applications applied to mental health and public health will be heavily described.

2.3.1. Mental health. Globally, over 700 million individuals are afflicted with mental disease, while an additional 30 million people experience panic attacks and anxiety disorders [9-10]. Given the exponential increase in the prevalence of mental problems, it has become challenging to promptly diagnose mental illnesses. However, by identifying the emotional state of these individuals early on, it is possible to avoid the condition from deteriorating further. Wearable monitoring devices can be used to collect data, while electronic records can retain large amounts of data, including information on emotion recognition, symptom tracking, and disease management [9]. Panagiotakopoulos et al. created a therapy method that focuses on data analysis. The method is designed to assist doctors in the management of individuals with anxiety disorders. The system will receive and process data obtained from various devices in order to forecast the occurrence of mental disease. Additionally, a novel method is being developed to ensure the protection of data obtained from neurotechnological devices using artificial neural networks (ANN), a type of machine learning algorithm [9].

2.3.2. Public health. Data Science has made significant contributions to the field of public health. Rathore et al. established a response unit called body area networks (BANs) to address the Covid outbreaks. This model utilizes the data gathered from the sensors worn by patients to forecast their medical condition [11]. Multiple data points will be simultaneously collected to forecast the future trajectory of the Covid patient's condition [12].

Consultation online is becoming increasingly common and a possible solution to the scarcity of healthcare resources and inefficient delivery of resources. According to the research from Jiang et al., with the data collected from the largest online medical consultation platform in China, the system will provide a list of how reliable the doctor are based on the data including reputation as a physician, service-related features such as quality of service, the patient can find a reliable doctor for consulting [13].

3. Benefits and challenges

With the help of data science, the accuracy of diagnoses and the ability of identifying potential causes of illness will be improved which will lead to the rate of failure in treatments be reduced [14]. Because of the data science, the history of patient can be tracked through a system called electronic health records (EHR). Doctors can know all the information of the patient which prevent the patient for spending money on treatments they've already tried, and this will also avoid hospitalization due to the improper treatment which will reduce the burden on hospitals and lower the hospital expense [15].

The data collected are mostly unorganized or inaccurate, these data will make the system to predict inaccurate results. The privacy of patient is also important, it is hard to keep the correct balance between the patient’s privacy and the amount of data. Data standardization, effective storage and transfers will need a lot of manpower to maintain and update. The coding language is different for different people, this will also cause the language barrier [10].

DOI: 10.54254/2755-2721/74/20240458

Currently, the only approach to address these difficulties is to handle the collected data with greater caution. Addressing these difficulties comprehensively requires significant effort. One potential solution is for the government to develop a centralized facility for storing patient information from all hospitals and clinics. The necessary data can be collected under the supervision of a dedicated government department. Nevertheless, it still has a considerable distance to cover.

4. Conclusion

Data science has the potential to significantly impact healthcare services. By utilizing a dependable data source and implementing appropriate management and processing techniques, it is possible to obtain accurate information that can be used to anticipate outcomes in the field of healthcare. Despite encountering numerous hurdles, the advancement of technology and the establishment of a more dependable source will ultimately resolve all of these challenges.

In recent years, the healthcare industry has been actively searching for superior alternatives to enhance the quality of medical care. Data science is a transformative force in healthcare, with the objective of developing a system that maximizes the advantages for patients. Currently, only a small portion of the available data is being discussed, but as technology advances, a greater amount of data will be utilized for analysis, resulting in more tailored plans for individual patients.

References

[1] Sengupta PP (2013) Intelligent platforms for disease assess- ment: novel approaches in functional echocardiography. JACC: Cardiovascular Imaging 6(11):1206–1211. https://doi.org/10. 1016/j.jcmg.2013.09.003

[2] Muni Kumar N, Manjula R (2014) Role of big data analytics in rural health care-a step towards svasth bharath. Int J Comp Sci Inform Technol 5(6):7172–7178

[3] Subrahmanya SVG, Shetty DK, Patil V, Hameed BMZ, Paul R, Smriti K, Naik N, Somani BK. The role of data science in healthcare advancements: applications, benefits, and future prospects. Ir J Med Sci. 2022 Aug;191(4):1473-1483. doi: 10.1007/s11845-021-02730-z. Epub 2021 Aug 16. PMID: 34398394; PMCID: PMC9308575.

[4] Fakhitah Ridzuan and Wan Mohd Nazmee Wan Zainon. 2019. A Review on Data Cleansing Methods for Big Data. Procedia Comput. Sci. 161, C (2019), 731–738. https://doi.org/10.1016/j.procs.2019.11.177

[5] Li, Lin. (2012) “Data Quality and Data Cleaning in Database Applications.” [doctoral dissertation], School of Computing, Edinburgh Napier University.

[6] Saha, Barna, and Divesh Srivastava. (2014) “Data Quality: The other face of Big Data.” in 2014 IEEE 30th International Conference on Data Engineering. pp. 1294-1297

[7] Müller, Heiko, and Johann-Christoph Freytag. (2003) Problems, Methods, and Challenges in Comprehensive Data Cleansing, Humboldt University Berlin

[8] Twin, Alexandra. “What Is Data Mining? How It Works, Benefits, Techniques, and Examples.” Investopedia. Accessed January 9, 2024. https://www.investopedia.com/terms/d/datamining. asp.

[9] Walker ER, Druss BG (2017) Cumulative burden of comorbid mental disorders, substance use disorders, chronic medical conditions, and poverty on health among adults in the USA. Psychol Health Med 22(6):727–735

[10] Ayesha Kamran Ul haq, Amira Khattak, Noreen Jamil, M. Asif Naeem, Farhaan Mirza, "Data Analytics in Mental Healthcare", Scientific Programming, vol. 2020, Article ID 2024160, 9 pages, 2020. https://doi.org/10.1155/2020/2024160

[11] Rathore MM, Ahmad A, Paul A, Wan J, Zhang D (2016) Real- time medical emergency response system: exploiting IoT and big data for public health. J Med Syst 40(12):283

[12] P. Kellmeyer, “Big brain data: on the responsible use of brain data from clinical and consumer-directed neurotechnological devices,” Neuroethics, vol. 11, pp. 1–16, 2018.

DOI: 10.54254/2755-2721/74/20240458

[13] Jiang J, Cameron AF, Yang M (2020) Analysis of massive online medical consultation service data to understand physicians' eco- nomic return: observational data mining study. JMIR medical informatics 8(2):e16765. https://doi.org/10.2196/16765

[14] Cruz JA, Wishart DS. Applications of machine learning in cancer prediction and prognosis. Cancer Inform. 2007 Feb 11;2:59-77. PMID: 19458758; PMCID: PMC2675494.

[15] K. Shvachko, H. Kuang, S. Radia and R. Chansler, "The Hadoop Distributed File System," 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST), Incline Village, NV, USA, 2010, pp. 1-10, doi: 10.1109/MSST.2010.5496972.