IDC 6145 Big Data for Data Science

EARLY DISEASE PREDICTION BASED ON PATIENT DATA

Heather Anderson 04 August 2024

Abstract

Diagnosing diseases early is a worldwide need that would save money on healthcare and improve patient outcomes. One solution to this problem would be to utilize big data in the form of medical records and other patient data to be able to better recognize diseases and illnesses before they occur or while they are in the early stages. The objective would be to predict diseases early based on this big data. Through data science tools, it would be possible to gather the relevant data and combine it in a way that is useful, and then process and analyze it effectively.

INTRODUCTION

Utilizing big data in the healthcare field would help with the issue of disease and illness in the general population. Through data such as medical records and other patient information, it is possible to create a system that could recognize diseases and illnesses early. This problem is important to study because it could drastically change the way healthcare is viewed and could save money for the healthcare facilities and the government, as well as individuals. It would also increase the quality of life for many individuals who may be able to catch issues early and avoid further complications.

The question to ask in this situation is, “How is this possible to achieve?” The objective is to develop a realistic plan of approach to this topic. The first thing that should be done is to show that this data meets the qualifications for “big data”. This specific topic is a big data problem because the data that would need to be acquired and processed is extremely big in volume. Medical records and other patient data take up a lot of storage space and would need to be constantly updated with new records being added as new data becomes available. This would also be challenging for traditional machine learning; However, with data that is large in volume, processing tools can be used to make it small enough for analysis.

BACKGROUND REVIEW

In examining the background for this topic, it is important to first revisit the basics and further prove that the topic of early disease prevention based on patient data is in fact a big data problem. The most straightforward is to use the five V's of big data to ensure that it meets the qualifications. The five V's are as follows: Value, Volume, Velocity, Variety, and Veracity (3).

Value, which refers to the insights and benefit that can be derived from the data (3), is again highlighting the importance of the applications. Research shows that using data to drive healthcare decisions can make the system more efficient, cut costs, and provide better patient outcomes, emphasizing how valuable big data is for early disease detection (4). Analyzing large sets of past medical records helps spot early signs of diseases, predict potential outbreaks, and tailor treatments for individuals. This approach improves how accurately we diagnose conditions, refine treatment methods, and enhance patient care.

Volume, which refers to the large amount of data that is generated and collected every day (3), also needs to be considered. Medical records and patient histories create a huge amount of data, often reaching terabytes or petabytes in size. This shows just how big the dataset can be. For effective early disease detection, these large volumes of data need to be collected, stored, and processed. An example dataset is the publicly available MIMIC-III dataset, which stands for The Medical Information Mart for Intensive Care III. The dataset consists of 112,000 clinical reports records, and data includes vital signs, medications, laboratory measurements, observations and notes charted by care providers, fluid balance, procedure codes, diagnostic codes, imaging reports, hospital length of stay, survival data, and more (7).

Velocity refers to the speed at which data is generated and processed (3). In the context of healthcare, data like patient monitoring, lab results, and electronic health records are updated constantly. To spot diseases early, it's important to analyze this data quickly. Fast processing helps detect issues sooner and adjust treatments based on the most current information, improving early detection and patient care.

Variety, which pertains to the different types of data that are generated and collected, can include structured, semi-structured, and unstructured data (3). While a good portion of the data collected would likely be more structured data (such as numerical values in lab tests or vital signs collected) there are other kinds of data that would fit the semi-structured or unstructured categories as well. For example, semi-structured data might include patient notes in health records, and unstructured data might include things like medical imaging. Being able to manage different types of data helps to analyze more thoroughly and predict diseases more accurately.

Veracity involves the accuracy and reliability of data and is really only relevant when the data is messy, broken and missing (3), which is common in big data and especially in the context of healthcare and early disease prediction. Healthcare data can be prone to errors or inconsistencies, or can be incomplete, which can impact disease predictions. Good data management, including cleaning and checking the data, is necessary to overcome these issues.

A company called HETT Insights is one of the many working to embrace data driven healthcare. They hold events and invite speakers to join in on the cause, and enthusiastically dive into the subject. They believe that with advanced analytics, real-time monitoring, and detailed electronic health records, healthcare professionals can make better decisions that improve patient outcomes and streamline operations (4). In addition, a company called Holon Solutions is dedicated to automating complex processes within healthcare and thus allowing providers to focus more on patient care, which they say is what matters most (5). They also have a motto that "Healthcare should feel human," and they say that technology should serve people, not the other way around (5).

In researching, it is clear that there are several gaps or shortcomings in existing research. One of the gaps is in data integration such as integrating various data types (health records, imaging data, lab results, genetic information) which limits the ability to perform comprehensive analyses (8). Another shortcoming is the limitation on real-time data processing and analysis capabilities which would also have a significant impact on the effectiveness of early disease detection or prediction. Data privacy and security is another potential gap because research may not fully address the implications of this topic. Biased data is also a concern that is always present and can lead to distrust if not addressed properly, as well as avoiding disparities in healthcare outcomes by ensuring the research and data sufficiently covers the undeserved or minority populations fairly. And finally, there seems to be a lack of interpretability of the models generated from research on big data models and algorithms (1). Overall, addressing these gaps and shortcomings will help strengthen the effectiveness of applications such as the early prediction of diseases or illnesses.

METHOD

In addressing the methodology used for data driven disease prediction, it is important to understand the steps included in the data life cycle and the aspects that are involved in each. The steps in the data life cycle include data acquisition, data processing, and data analysis. The aspects, which are not necessarily steps but more of points of consideration during each of the steps in the life cycle, include data storage, data security and privacy, and data governance and management. It is also important to discuss potential limitations or challenges.

For the first step of the data life cycle, the data acquisition phase, data could be collected such as electronic health records (EHRs), including patient demographics, medical histories, surveys, medical bills, reports, and diagnostic tests (6). It could even be taken a step further with integrating real-time health sensor data for continuous patient monitoring. The MIMIC-III data set, for example, already contains so many different types of information and records so that would definitely be a great place to start with collecting relevant data.

In the second step, data processing, Hadoop ecosystem tools such as HDFS and MapReduce would be excellent tools for processing and storing large volumes of medical data. The data quality and integrity could be ensured through preprocessing techniques. As mentioned above, the large volume of data can be processed and made small enough to use for analysis if necessary.

For the third and final step in the data life cycle, data analysis, machine learning and/or deep learning algorithms would need to be applied in order to analyze patterns and extract predictive features or insights. One study by Aakash Chotrani, looks at various diseases such as cardiovascular conditions, cancer, infectious diseases, and metabolic disorders. It compares different machine learning models—like decision trees, support vector machines, neural networks, and ensemble methods—by evaluating not just their prediction accuracy but also their efficiency and scalability for use in real-world healthcare settings (2).

The aspect of data storage should be addressed at each of the steps. For data storage in the acquisition phase, using cloud storage and NoSQL databases works well due to the need to collect large amounts of data from electronic health records (EHRs) or real-time sensors. Cloud storage can easily handle large and varied datasets, while NoSQL databases like MongoDB or Cassandra are great for managing different types of data and handling lots of information at once. In the processing phase, Hadoop Distributed File System (HDFS) is great for storing and managing large amounts of data across multiple servers (8). Cloud storage options are also effective, as they easily scale and work well with tools for processing big datasets. And finally, in the analysis phase, NoSQL databases or in-memory databases can provide quick access to data, while cloud-based analytics services offer powerful data warehousing solutions as well. This is especially important to consider since it is likely to involve working with deep learning and machine learning models.

Security and privacy throughout the process is another aspect that needs to be considered each step of the way. For instance, in the data acquisition phase, it is important to ensure that the patients' privacy is met and that they comply with regulations like HIPAA. In the data processing step, it might be wise to restrict access to the data to authorized users only and use encryption to be safe. And in the analysis phase, it might be a good idea to consider adding additional layers of security such as helping to protect the patient's identities through making the names anonymous to further ensure that the insights derived do not compromise patient confidentiality.

It is important to also address the aspect of data governance and management throughout the process. In the data acquisition step, establishing clear standards for quality and accuracy is crucial, especially with varied datasets like health records and sensors. During data processing, maintaining data quality involves cleaning the data and keeping accuracy intact. And in data analysis, making sure results can be repeated and ethical standards are followed requires keeping detailed records and tracking changes carefully.

The chosen methodology fits the research goals because it uses advanced data science tools to tackle the challenges of early disease detection with big data. Cloud storage and HDFS are used for their efficiency in handling large amounts of data, NoSQL databases are selected for their ability to manage various types of data, and machine learning and deep learning algorithms are applied to analyze complex patterns and make predictions.

Some potential limitations or challenges of the proposed methodology include constantly staying up to date on the latest security measures and meeting all of the legal and ethical standards (especially due to dealing with sensitive medical information), the complexities of the models, costs of implementing and processing infrastructure, and the computational constraints and need for high-speed data access (8).

References

(1) Batko, Kornelia, and Andrzej Ślęzak. “The use of Big Data Analytics in healthcare.” *Journal of big data* vol. 9,1 (2022): 3. doi:10.1186/s40537-021-00553-4

(2) Chotrani, A. “Comparative Analysis of Machine Learning Models for Disease Prediction”. *Journal of Science & Technology*, vol. 3, no. 2, Apr. 2022, pp. 10-20, https://www.thesciencebrigade.com/jst/article/view/65.

(3) Fang, Ian. *Introduction to Big Data*. Course Note Hosting, https://uwf-fang.github.io/big_data/big_data_intro.html. Accessed 04 Aug. 2024.

(4) HETTShow *Data-driven decision making in healthcare: Hett Insights*, *Data-Driven Decision Making in Healthcare - HETT Insights*. Available at: https://blog.hettshow.co.uk/transforming-healthcare-the-power-of-data-driven-decision-making (Accessed: 04 August 2024).

(5) Holon Solutions (2024) *The impact of big data on Disease Prediction and Prevention*, *Holon Solutions*. Available at: https://www.holonsolutions.com/the-impact-of-big-data-on-disease-prediction-and-prevention/ (Accessed: 04 August 2024).

(6) Kanchanamala, P. & Das, Smritilekha & Neelima, G.. (2022). *Symptoms-Based Disease Prediction Using Big Data Analytics.* 10.1007/978-981-16-8987-1_36.

(7) *MIMIC-III dataset MIMIC-III Dataset - Papers with Code*. https://paperswithcode.com/dataset/mimic-iii Accessed: 04 August 2024.

(8) P, Kauser & Singh, Shrishty & Pathak, Chavi & Singh, Simran & Student,. (2022). *Big Data Analytics for Chronic Disease Prediction.* 4. 10.56726/IRJMETS30011.