IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 2, September 2024

The Use of Big Data Analytics in Health Care

Ms. Hemlata and Dr. Shweta
Assistant Professor in Computer Science,
Saraswati Mahila Mahavidyalaya, Palwal, India

Abstract: *Utilizing Big Data analytics has the potential to enhance patient outcomes, promote individualized care, strengthen provider-patient relationships, and decrease healthcare expenditures. This paper provides an introduction to healthcare data, specifically focusing on big data in healthcare systems. It also explores the various applications and advantages of utilizing big data analytics in the healthcare industry. In addition, we showcase the advancements in big data technology in the healthcare sector, including cloud computing and stream processing. The challenges associated with the analysis of large volumes of data in healthcare systems are also addressed.*

Keywords: Big data, Big Data analytics, Healthcare, Personalized medicine, Precision medicine, Cloud computing, Stream processing

I. INTRODUCTION

Accurately pairing patients with appropriate medical treatments for specific diseases can minimize unnecessary side effects, enhance the quality of treatment, and prevent inappropriate treatment or wastage in medical services. Additionally, it has the potential to introduce novel medical therapies by investigating new pharmaceuticals or repurposing existing drugs for innovative or more specific applications (Price and Nicholson, 2015). Systems biology is an effective approach that combines various data sources and investigations of biological processes. A significant amount of research focuses on network models to describe the development and immune responses of diseases, which aids in the identification of new biomarkers for early detection. However, it is crucial to prevent any bias in clinical data when utilizing these models (Ren and Krawetz, 2015).

Various categories of medical equipment, particularly wearable devices, consistently collect data; the rapid speed at which the data is generated often necessitates quick processing during an emergency. The potential value of a data source that is isolated may be limited, but by combining electronic medical records (EMRs) and electronic health records (EHRs) through data fusion, the profound value of healthcare data, such as public health warnings and personalized health guidance, can be maximized (Zhang et al., 2017). Structural MRI is a powerful technique for visualising the brain, providing detailed brain maps with high spatial resolution. This method yields a wealth of high-dimensional data, making it valuable for both research and clinical purposes in identifying structural characteristics of the brain (Ulfarsson et al., 2016).

Healthcare applications for mobile and web platforms have been created to enable patients to send queries about their symptoms to healthcare providers via a server. Mobile applications may contain first aid instructions and provide emergency assistance to patients for immediate treatment or guidance to appropriate departments (Panda et al., 2017). A mobile cloud computing (MCC) healthcare system was developed to gather and analyses real-time biomedical signals (such as blood pressure and ECG) from users located in different locations. A customized healthcare app is installed on the mobile device, and health data is synchronised with the healthcare system's cloud computing service for storage and analysis (Lo'ai et al., 2016). The utilization of advanced information technology enables the capture of large volumes of data in the healthcare sector, facilitating the analysis of information to enhance the process of policy-making. A life table is an appropriate tool for studying population ageing and medical expenses, as it offers valuable evidence for policy formulation (Wang et al., 2017). The expenses related to healthcare also rise in tandem with the advancing age of the population. Japan has commenced the utilization of Big Data technologies to enhance medical treatment and healthcare services for the elderly population. Big Data analytics enables the extraction of valuable insights from vast and complex datasets through the process of data mining (Tsuji, 2017). This paper was completed by conducting literature research on the Scopus and IEEE Xplore databases. The search was conducted using specific

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 2, September 2024

combinations of keywords, namely big data and healthcare, big data and health care and big data and medical. The search was limited to papers published between January 2015 and May 2018. The duplicated papers identified in both databases were eliminated, resulting in a selection of 316 papers for the literature review.

Healthcare Data

Various types of healthcare data sources encompass clinical text, biomedical images, electronic health records (EHRs), genomic data, biomedical signals, sensing data, and social media (Ta, 2016). Genomic data analysis enables individuals to gain a more comprehensive comprehension of the associations between various genetic markers, mutations, and disease conditions. Moreover, the process of translating genetic findings into personalized medical application is a complex endeavour that presents numerous unresolved obstacles. Clinical text mining converts unstructured clinical notes into valuable information. Information retrieval and natural language processing (NLP) are techniques used to extract valuable information from extensive amounts of clinical text. Social network analysis facilitates the identification of knowledge and novel patterns that can be utilized to model and forecast worldwide health trends, such as the occurrence of infectious epidemics. This analysis relies on diverse social media resources, including but not limited to web logs, Twitter, Facebook, social networking sites, and search engines (Ta et al., 2016). Prior to assessing the severity of diseases, it is imperative to employ appropriate diagnostic techniques. Table 1 (Verma and Sood, 2018) presents a diagnostic scheme utilised for the identification of the disease. Table 2 (Mendelson, 2017) presents a visual representation of five distinct levels that depict personal health-related data. Despite the lag in legal advancements, it is imperative to safeguard the fundamental rights of data subjects and privacy.

Table 1. A scheme for diseases diagnosis in a system

Disease	Diagnostic method	Health measures via IoT
Hypertension	Frequency based, scale based	Blood pressure
Obesity	Scale based	Body weight, blood pressure
Heart diseases	Frequency matching, pattern-matching	ECG pattern
Water borne or infectious disease	Frequency based, scale based	ECG, temperature sensor, Camera pill (gastro intestinal tract)
Stress index	Frequency based, pattern-matching based	Emotiv EPOC sensor, other stress measuring sensors
Respiration index	Frequency matching, pattern-matching	Respiration sensor

Table 2. Datafication layers of personal health data

Layers	Description
Layer 1	Internet search, Smartphone, purchase, HER, healthcare services, IoT sensors, medical and fitness device
Layer 2	Clinical and other data related to health in identified forms which are collected, stored, and distributed to third parties.
Layer 3	Mainly private companies collect raw data from Layer1, Layer 2, and other private and public sources. The processed results are distributed or sold either in a de-identifiable or identifiable form.
Layer 4	National government, international private or public entities re-process, re-distribute, or re-sell the data for various purposes.
Layer 5	International agreements and treaties governing the protection of privacy for personal data related to health

A wireless body sensor network (WBSN) consists of six wireless physiological sensors. The six sensors are employed to gather a patient's six essential indicators, which encompass body temperature, heart rate/pulse, blood glucose, blood pressure, ECG, and oximetry (You et al., 2018). The integration of healthcare data has been a significant concern, encompassing various types of information such as personal health records and epigenomics. Several integration methods have been created, such as data warehousing (consolidating data into a unified data structure), link integration (incorporating data into a webpage), service-oriented architectures (delivering data in a dynamic and user-friendly format on the web), view integration (combining multiple databases), and mash-ups (merging data from multiple web resources to create a new web application) (Murphy et al., 2017). The challenges in data fusion are outlined in Table 3 (Capobianco, 2017).

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 2, September 2024

Table 3. Critical aspects in data fusion

Aspect	Description
1	Missing data handling
2	Balancing data from different origins or sources
3	Dealing with inconsistent, contradicting and conflicting data
4	Establishing loss or objective functions and regularization/penalty terms
5	Differentiating between soft and hard data links, i.e. considering a random process from which the data is generated as subject to same parameters, or instead accounting just for dependencies, covariations, similarity/dissimilarity, etc.

Utilising advanced data analytics techniques in healthcare systems to analyses large volumes of data. Table 4 (De Silva et al., 2015) outlines that big data typically exhibits significant attributes in terms of volume, velocity, variety, variability, value, complexity, and sparseness. The field of healthcare can benefit greatly from the utilisation of big data, as it has the capacity to be applied in various areas such as disease surveillance, epidemic control, clinical decision support, and population health management (Sabharwal et al., 2016). The utilization of Big Data in the healthcare industry can yield substantial advantages, including the ability to identify diseases in their initial stages. Integrating Big Data analytics into smart healthcare systems introduces advanced electronic and mobile health (e/m-health) technologies that enhance efficiency and reduce medical expenses (Pramanik et al., 2017).

Table 4. Big data features

Aspects	Description	Examples in healthcare
Volume	Data size	Treatment plans, multiple conditions, and cohorts of patients
Velocity	Data generation rate (batches, streams, infrequent intervals)	Sensing and diagnostics transmitting patients' status and behaviors
Variety	Various formats and data types (numbers, text, images)	Clinical, medical, and omics data and images from various patients under diverse conditions
Variability	Data change with time	Health data from wearable sensors
Veracity	Imprecise or untruthful data	Clinician notes about patients' states, patients' feedback
Value	Inherent value (often achieved through data mining)	Analyzing numerous patients' feedback and identifying the side effects of a drug
Complexity	Hierarchies, linkages between items and recurrent structure of data	Multi-pharmacy, multi-morbidity
Sparseness	Low density of useful information (due to null values, missing data, etc.)	Many missing data of patient feedback on progress and symptoms

Predictive analytics can be utilised to forecast pharmaceutical outcomes, ascertain patients who derive the greatest advantage from pharmacist interventions, enhance pharmacists' comprehension of the risks associated with specific medication-related issues, and administer interventions customised to patients' requirements (Hernandez and Zhang, 2017). Precision medicine encompasses the entire spectrum of data, including its collection, management (including storage, sharing, and privacy), as well as analytics (such as data integration, data mining, and visualisation). Advancements in biotechnologies have led to the availability of large volumes of complex biomedical data. Utilising heterogeneous data requires the application of Big Data analytics. This field encompasses various application areas including health informatics, sensor informatics, bioinformatics, imaging informatics, and more (Wu et al., 2017). Accuracy is essential for the analysis of Big Data. PHRs may include abbreviations, typographical errors, and cryptic notes. Ambulatory measurements are potentially conducted in unregulated and less dependable conditions, in contrast to clinical data that is gathered by skilled professionals in a controlled clinical environment. Utilising uncontrolled, unsolicited data from social media platforms may lead to imprecise forecasts. Furthermore, it is worth noting that data sources can exhibit bias on occasion (Andreu-Perez et al., 2015). The issue of 'noise' data becomes increasingly problematic, particularly as it expands rapidly. Databases with varying levels of completeness and quality result in diverse outcomes, which heighten the likelihood of inaccurate findings and biassed investigations. The primary issues are inadequate data quality and biases resulting from the lack of randomisation. The value of big data is frequently enhanced by integrating disparate databases and conducting comprehensive analysis of all available and interconnected data (Sacristán and Dilla, 2015). Data pre-processing refers to the procedure of converting unprocessed data into a comprehensible format, which typically involves the following steps: 1) data cleansing, 2) data integration, 3) data transformation, 4) data reduction, and 5) data discretisation. Pre-processing is a crucial stage in the analysis of Big Data

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 2, September 2024

(Farid et al., 2016). Systems utilising large-scale data streams have been created, incorporating patient-specific hospital discharge records, electronic death certificates, and medical claims data that employ International Classification of Diseases (ICD) coding. The authors Simonsen et al. (2016) have suggested the utilisation of surveillance strategies that involve analysing large volumes of data obtained from crowdsourcing, social media, and Internet search queries. Big Data technologies, such as NoSQL databases, have been utilised for processing healthcare information. Certain features, such as local access and a rational relationship between logical and physical data distribution, are crucial for enhancing the performance of parallel processing in distributed databases (Salavati et al., 2017). A proposed approach and process, driven by Big Data, integrates both clinical and molecular information. The initial step involves the identification of candidate biomarkers and therapeutic targets/drugs. The cross-species analysis is used to complete the subsequent clinical or preclinical validation, resulting in a reduction of costs and time required for biomarker/therapeutic development (Wooden et al., 2017). A clinical data warehouse was established to store organised data, while a series of modules were developed to analyse unstructured content. The study aimed to develop an initial version of a framework within the context of big data. The framework executes the modules within a Hadoop cluster, utilising the distributed computing capability of Big Data (Istephan and Siadat, 2015). An architecture based on Hadoop was created to handle large amounts of health-related data from Twitter. Examining tweets in the healthcare field has the capacity to revolutionise the utilisation of cutting-edge technologies by individuals and healthcare professionals in order to attain novel clinical understandings (Cunha et al., 2015). Big Data analytics has utilised open source technologies like Hadoop, Kafka, Apache Storm, and NoSQL Cassandra. Apache Storm provides a collection of fundamental components for processing large-scale real-time data (Vanathi and Khadir, 2017). Table 5 presents a juxtaposition of Storm and Hadoop.

Table 5. A comparison of features between Storm and Hadoop (Vanathi and Khadir, 2017)

Features	Storm	Hadoop
Data handling	Handled as topology	Handled as jobs
Data Processing	Real-time oriented	Batch oriented
Database Compatibility	Cassandra, NoSQL	HBase, SQL
Performance	Low latency	High latency

Research has been conducted on attribute reduction using MapReduce, based on the Rough Set Theory (RST). The procedures involve utilising parallel large-scale rough set methods to acquire features and implementing them on MapReduce runtime systems such as Twister, Phoenix, and Hadoop to extract features from large datasets through data mining. Additionally, the framework structure of <key, value> pairs is utilised to expedite the computation of equivalence classes and attribute significance. The traditional attribute reduction process based on MapReduce is parallelised (Ding et al., 2018). Traditional high-performance computing (HPC) focusses on CPU-intensive computation using either supercomputing or high-performance networking (cluster or grid computing). On the other hand, Hadoop-enhanced computing is designed for large-scale distributed data processing, utilising both internal and external networking. The utilisation of Hadoop for processing large volumes of data offers three distinct benefits: enhanced operational effectiveness, increased dependability, and the ability to easily accommodate growth and expansion (Ni et al., 2015). Table 6 (Olaronke and Oluwaseun, 2016) presents a juxtaposition of tools employed for the analysis of large-scale data in the healthcare domain.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 2, September 2024

Table 6. A comparison of tools used for analyzing big data

Tools	Type of Databases	Platforms	Advantages	Limitations
Hadoop	Non-relational database	Open source and cloud-based platform	Stores data with any structure such as Web logs	Lacks technical support and security
MapReduce	Non-relational database	Open source and cloud-based platform	Works well with semi-structured and unstructured data such as visual and audio data.	Lacks indexing capabilities of modern database systems.
Google Big Query	Columnar database	Open source and cloud-based platform	Allows data to be replicated across diverse data centers.	Does not support indexes.
Microsoft Windows Azure	Relational database	Public cloud based platform	Allows users to make relational queries against structured, semi-structured and unstructured files.	The size of the database is limited; it cannot handle huge databases.
Jaql	It is a query language for JavaScript object notation.	It is a proprietary query language.	Supports both structured and semi-structured data.	No user defined types; schema information only for possible values of a domain

Industry 4.0 encompasses a strategic initiative that encompasses the manufacturing sector, with a specific focus on the production of medical devices and drugs. Precision medicine is a health application that utilises Big Data, incorporating multi-omics, IoT, Industry 4.0, and other technologies. Industry 5.0, as proposed by Özdemir and Hekim (2018), aims to utilise artificial intelligence, Internet of Things (IoT), and next-generation technology policy to effectively analyse and utilise Big Data. A sophisticated healthcare framework has been created utilising IoT technology to offer comprehensive healthcare services to individuals during their exercise sessions. A Bayesian belief network classifier was employed to forecast an individual's susceptibility to health-related risks using an artificial neural network model. Verma and Sood (2018) have identified data management, model development, visualisation, and business models as the four main domains of Big Data analytics. The authors Wu et al. (2017) have summarised various data mining methods for complex electronic health record (EHR) big data in Table 7.

Table 7. Some methods for EHR data mining

Methods	Advantages	Limitations
Hidden Markov models	Simultaneous segmentation, detection, and classification in a waveform	Sensitive to the design of trained Markov model
Logistic regression with LASSO regularization	Reduces feature space	Prone to over-fitting
Logistic regression, local regression, cox regression	Direct estimates of relevant hazards for Cox regression; simple to interpret and implement	Sensitive to an outlier
Rule mining, directed acyclic graph, Allen's interval algebra	Temporal modeling/mining capabilities	Requires specific design of experiment
Conditional random fields	Resistant to differences in class prevalence; supporting temporal analysis	Sensitive to feature space size and regularization
Windowing, episode rule mining, relational subgroup discovery	Valid sequential methods for some clinical applications	Tradeoffs between complexity, simplicity, and temporal resolution

II. CHALLENGES OF BIG DATA IN HEALTHCARE SYSTEMS

The challenges of healthcare big data lie in the processes of capturing, storing, sharing, searching, and analyzing health data. Organizing data after extracting it from various layers and integrating the data poses another challenge (Reddy and Kumar, 2016). Integrating physiological data with high-throughput "omics" techniques for clinical recommendations poses a significant challenge. The growing abundance of genomic data, along with the impact of gene annotation and errors in analytical practices and experiments, has made the analysis of functional effects using high-throughput sequencing methods a complex task (Belle et al., 2015). The matter of obtaining consent for the utilization of healthcare data, including genetic data, has been a significant concern. The creation of databases using extensive and nationwide

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 2, September 2024

population data for future research, with the necessary ethical approval and governance, has sparked academic discussions regarding its legality. There is ongoing debate regarding the usefulness of Big Data in enhancing healthcare systems (Knoppers and Thorogood, 2017). The subsequent are overarching obstacles encountered in the realm of healthcare when dealing with Big Data (Mathew and Pillai, 2015): Ensuring the protection and confidentiality of information: Conventional privacy and security measures are effective for small datasets. However, applying these measures to large and continuously flowing datasets, especially when dealing with sensitive patient health data, can pose challenges. Data quality has a direct impact on the accuracy and dependability of the insights derived from the data, as well as the decision-making process in the context of patients' healthcare. Inadequate real-time processing: A delay in processing intricate data models can lead to a decrease in the quality of patient care. The integration of heterogeneous data sources, such as data fragmentation across hospitals, labs, electronic health records (EHRs), and financial IT systems, poses a significant challenge in combining data into a unified database system. There are no established criteria for healthcare data. Diverse types of healthcare data are produced and gathered by different entities, including practitioners' records, medical images, and data from wearable sensors. The absence of standardized criteria for this data poses challenges for subsequent processing.

III. CONCLUSION

Conventional data processing methods are insufficient for managing large volumes of data in healthcare systems. Big Data analytics surpasses the constraints of conventional data analytics and has the potential to revolutionise healthcare. Big Data analytics has the potential to revolutionise disease surveillance, epidemic control, clinical decision support, and population health management. Hadoop-enhanced computing refers to the use of Hadoop-based Big Data for performing intensive computations on large-scale distributed data. This approach offers several advantages, including improved efficiency, reliability, and scalability. The healthcare systems face various challenges when it comes to the analytics of Big Data. The challenges of Big Data encompass the tasks of capturing, storing, sharing, searching, and analyzing data across various domains. Furthermore, challenges in Big Data analytics within healthcare systems encompass data security and privacy, data quality, real-time processing, integration of diverse or dissimilar data, and establishment of standards for healthcare data.

REFERENCES

[1]. Istephan, S., & Siadat, M. R. (2015, November). Extensible query framework for unstructured medical data-a big data approach. In Data Mining Workshop (ICDMW), 2015 IEEE International Conference on (pp. 455-462). IEEE.

[2]. Andreu-Perez, J., Poon, C. C., Merrifield, R. D., Wong, S. T., & Yang, G. Z. (2015). Big data for health. IEEE J Biomed Health Inform, 19(4), 1193-1208.

[3]. Big data analytics in healthcare. BioMed Research International, 2015. Capobianco, E. (2017). Systems and precision medicine approaches to diabetes heterogeneity: a Big Data perspective. Clinical and Translational Medicine, 6(1), 23.

[4]. Cunha, J., Silva, C., & Antunes, M. (2015). Health twitter big bata management with hadoop framework. Procedia Computer Science, 64, 425-431.

[5]. De Silva, D., Burstein, F., Jelinek, H. F., & Stranieri, A. (2015). Addressing the complexities of big data analytics in healthcare: the diabetes screening case. Australasian Journal of Information Systems, 19, S99-S115.

[6]. Ding, W., Lin, C. T., Chen, S., Zhang, X., & Hu, B. (2018). Multiagent-consensus-MapReduce-based attribute reduction using co-evolutionary quantum PSO for big data applications. Neurocomputing, 272, 136-153.

[7]. Farid, D. M., Nowe, A., & Manderick, B. (2016, December). A feature grouping method for ensemble clustering of high-dimensional genomic big data. In Future Technologies Conference (FTC) (pp. 260- 268). IEEE.

[8]. Belle, A., Thiagarajan, R., Soroushmehr, S. M., Navidi, F., Beard, D. A., & Najarian, K. (2015).

[9]. Hernandez, I., & Zhang, Y. (2017). Using predictive analytics and big data to optimize pharmaceutical outcomes. American Journal of Health-System Pharmacy, 74(18), 1494-1500.