Data Analytics in Healthcare Systems – Principles, Challenges, and Applications

Chapter · May 2021

DOI: 10.1201/9781003185246-1

4 authors:

s. Suganthi
Galgotias University
7 PUBLICATIONS 2 CITATIONS
SEE PROFILE

Vaishali Gupta
Galgotias University
13 PUBLICATIONS 3 CITATIONS
SEE PROFILE

Varsha Sisaudia
Delhi Technological University
11 PUBLICATIONS 41 CITATIONS
SEE PROFILE

T. Poongodi
Galgotias University
52 PUBLICATIONS 180 CITATIONS
SEE PROFILE

1 Data Analytics in Healthcare Systems – Principles, Challenges, and Applications

S. Suganthi, Vaishali Gupta, Varsha Sisaudia, and T. Poongodi

CONTENTS

1.1 INTRODUCTION

The healthcare industry is multidimensional, with multiple data sources involving healthcare systems, health insurers, clinical researchers, social media, and government [1], generating different types and massive amounts of data. It is impossible to handle this big data with traditional software and hardware and the existing storage methods and tools. Data analytics is the process of the analysis of data to identify trends and patterns to gain valuable insights. The data generated in the health industry are characterized by the four Vs of big data, namely volume, velocity, variety, and veracity, which play crucial roles in health data analytics. Also, evidence-based decision making has gained importance, which involves the sharing of data among various data repositories. According to Deloitte Global Healthcare Outlook, it is expected that global healthcare expenditure will continue to increase at an annual rate of 5.4% between 2017 and 2022. This is due to the increased importance of personalized medicine, the use of advanced technologies, the demand for new payment models, improvement and expansion of care delivery sites, and competition. Various research attempts, based on big data, have provided strong evidence that the efficiency of healthcare applications is dependent upon the basic architecture, techniques, and tools used. Statistical data and reports can be generated with the use of patient records, aiding in knowledge discovery, and thereby influencing value-added services to the patients, improving healthcare quality, the making of timely decisions, and minimizing the costs incurred. Hence, there is a need to incorporate and integrate big data analytics into existing healthcare systems. Despite healthcare analytics having massive potential for value-added change, there are many technological, social, organizational, economic, and policy barriers associated with its application [2].

1.1.1 Data Analytics in Healthcare

Health industries employing data analytics can use big data for the early detection of diseases and their treatment, clinical operations, genomic analysis, patient profile analytics, and prevention of fraud. Data processing involves analytics being applied to the transformed data to obtain meaningful insights from the healthcare data for evidence-based decision making. There exist three levels of analytics with increasing complexity and value, which are given below.

- **Descriptive Analytics** describes the current events or summarizes the past events by generating reports with the help of statistical tools such as tables and graphs. Thus, it helps medical practitioners to study and understand the patient's behavioral patterns in the past with the help of the patient's operational data, which can be used to solve problems in current situations.
- **Predictive Analytics** enables the user to predict the future with the use of descriptive data, using empirical methods, such as machine learning, modeling, and data mining, to analyze the data.
- **Prescriptive Analytics** enables the user to choose the best solution from several possible alternatives to the issues in question. The real-time data are analyzed with the historical data by the use of artificial intelligence, data mining, and machine learning. The practitioners can determine the possible effects of their decisions by using simulation and optimization techniques and can prepare themselves to handle any event in case of failure or success.

1.1.2 Characteristics of Big Data

The main characteristics of big data are the four Vs, which are as follows.

- **Volume:** The healthcare industry generates an enormous amount of data coming from various sources, such as EHRs (Electronic Health Records), LIMS (Laboratory Information Management System), diagnostic or monitoring instruments, supply chains, insurance claims/billing, pharmacy, real-time locating systems [3], and social media. The data that are collected are used in continuous learning by employing various technologies and processes to derive insights from the information, which will help improve the quality of healthcare. In addition, the reduction in storage costs and the development of advanced architectures have led to large volumes of data being stored, processed, and managed, using the existing traditional systems adopted by the healthcare industry.
- **Velocity:** The speed with which medical data are generated is high and requires specific processing requirements. The healthcare industry has not adapted to technological advances, and hence various processing methods have been adopted to cope with the speed with which the data are produced. Batch processing, stream processing, near-real-time processing and real-time processing methods are used to handle data in the healthcare industry [4].
- **Variety:** Data in healthcare come from several sources and are in various forms. About 80% of the total healthcare data is unstructured (e.g., images, signals, text), which does not fit into any predefined data format, data type, or structure. The remaining 20% of the data is structured (e.g., temperature, blood pressure, patient demographics, etc.), possessing a predefined data format. The structured data are easy to handle, store, process, and manipulate. Hence, the unstructured data must be converted to a structured form by efficient automatic transformation methods. When combined with structured data, the unstructured data provide valuable information, which can be harnessed to improve value-added services in healthcare by adopting more-efficient methodologies.
- **Veracity:** This refers to the accuracy of the data collected, which is directly proportional to the value of the insights which can be obtained from them. The results derived from data analytics are error free if the data obtained are trustworthy, accountable, authenticate, and available.

1.2 ARCHITECTURAL FRAMEWORK

Hadoop/MapReduce is an open-source platform for big data analytics used in healthcare, which performs parallel processing in a distributed environment, involving multiple nodes in the network. The use of Hadoop and MapReduce technologies has been found to be fruitful in many healthcare applications, by improving the performance of, for example, image processing, neural signal processing, protein

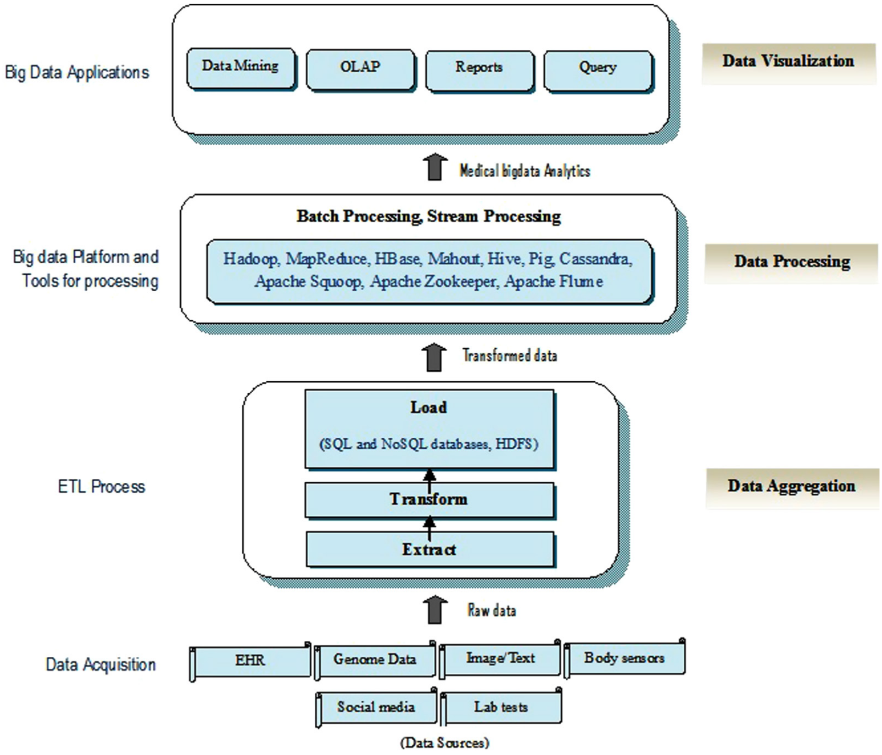

FIGURE 1.1 Conceptual framework of big data architecture in healthcare.

structure alignments, signal detection algorithms, and lung texture classification [5]. The architectural framework of big data in healthcare is composed of three major components, namely data aggregation, data processing, and data visualization [6]. Figure 1.1 illustrates the conceptual framework of big data architecture in healthcare.

1.2.1 Data Aggregation

Data aggregation in healthcare involves the process of collecting and integrating raw data from various modalities and multiple systems and converting them into a single standard format suitable for analysis, processing, and storage in a data warehouse. The functionalities involved in the process include data extraction, data transformation, and data loading.

- **Data Extraction**
  Healthcare data occupy large volumes and come from heterogeneous sources. The primary sources of medical data include medical records, health surveys, claims data, disease registries, vital records, surveillance data, peer-reviewed literature, clinical trial data, and administrative

records. The data from these sources can be structured or unstructured. The structured data, which contain numeric, categorical, and nominal data types, have a predefined format and are easy to handle. Unstructured medical data, such as image data from imaging devices, text data (e.g., doctor notes, shorthand notations etc.), and signal data (e.g., biosignals from wearable devices), do not have a predefined format. They must be converted into a standard format for further processing. The medical data can be in any format, which includes EHRs, images, biomedical data, genomic data, sensor data, and clinical text data. These data are in the form of text/ASCII, XML, JSON, or images in DICOM formats. Usually, medical image data from local workstations are stored in PACS (Picture Archiving and Communication Systems) and are transferred and communicated to other workstations following DICOM standards.

- **Data Transformation**
  The acquired raw data are transformed in order to apply particular business logic by passing through the phases of *data filtering*, which is the process of removing unwanted data ("data cleaning"), which includes the processes of normalization, noise reduction, and management of missing values, and *data manipulation*. The data converted in this way are in a standard format that is consistent and suitable for further analysis.
- **Data Loading**
  Customized relational databases are usually designed for each particular healthcare system, with its own defined data models and schemes for the storage of medical data. The transformed medical data, which are suitable for further analytics, are loaded into the target database or a data warehouse, such as HDFS (Hadoop Distributed File System), SQL relational databases, or NoSQL databases, or combinations of these.

1.2.2 Data Processing

The data processing used in healthcare includes *batch processing* and *stream processing* methods [7]. Batch processing is the method of analyzing data in batches, which are collected and stored over a period and in which response time is not considered. On the other hand, stream processing is the method of analyzing huge volumes of data, to which a real-time response is required. Some applications in healthcare require real-time processing of data and they are characterized by noisy data with missing or redundant values, continuous changes in data, and the need for a rapid response. Stream processing overcomes these difficulties with simple and rapid information extraction by using data-mining methods, such as clustering, classification, and frequent pattern mining [7]. Apache Hadoop MapReduce is a popular framework used for batch processing, whereas Storm and S4 are frameworks used for stream processing, with Apache Spark and Apache Flink being frameworks used for both batch and stream processing.

The Hadoop platform is most widely used for batch processing in which parallel processing of huge volumes of data are carried out in a distributed manner. It is a framework in which the process of big data analytics is conducted through a collection of various tools, methodologies, and libraries. It consists of two main components, the HDFS and Hadoop MapReduce.

- **Hadoop Distributed File System (HDFS)**
  The HDFS is a distributed file system used for storing and retrieving extremely large volumes of data at great speed. The data are split into several blocks with uniform block sizes of 64 MB or 128 MB, which are distributed across many server nodes, enhancing parallel processing. The HDFS consists of two nodes, namely a *NameNode* and a multiple number of *DataNodes*. The DataNodes contain the application data and the NameNodes contain metadata related to the storage and retrieval of data from the DataNodes. The data are replicated and the copies are distributed to many nodes, making the system fault tolerant in the case of any node failure.
- **MapReduce**
  MapReduce is a framework and programming model for distributed processing of huge datasets, involving multiple nodes. The processing is broken down across several individual nodes and carried out in parallel because the data are vast. The process consists of two phases, the map phase and the reduce phase, conducted with the mapper function and the reducer function, respectively.
  1. **Map Phase:** the data stored in the HDFS are split into smaller, fixed-sized elements and passed to a mapper function. The mapper function produces structured output data tuples of each component in the form of key/value pairs, which are written into an intermediate file.
  2. **Reduce Phase:** the output from the map phase is passed to the reduce phase, before which it is shuffled to consolidate and combine the relevant data tuples into smaller ones. The reducer aggregates the output from shuffling by merging the same keys to form a smaller set of tuples, which is written to a single output file.

Data storage can be carried out with tools other than HDFS, such as HBase, Hive, Cassandra, Pig, Apache Flume, Apache Squoop, and other relational databases. Whereas Apache Oozie is used in the case of large numbers of interconnected systems, Apache Zookeeper is used to maintain application reliability, and Mahout is used for machine- learning purposes [7].

1.2.3 Data Visualization

Visualization is the graphical representation of data, which helps the practitioner gain more insights from the data. The analytical tool cleanses and evaluates the data with the help of data-mining algorithms, evaluation, and software tools before the data are visualized. The main applications of data analytics to healthcare are queries, reports, online analytical processing (OLAP), and data mining. They are used for displaying predictive reports, proactive messages, visualization of patient health records, real-time alerts, and dashboards for monitoring the daily health status of patients.

1.3 DATA ANALYTICS TOOLS IN HEALTHCARE

One of the critical challenges in healthcare systems is utilizing the massive amount of data generated daily in an efficient and cost-effective manner. Hence, healthcare systems require effective and efficient tools to ensure the appropriate use of the data. Figure 1.2 depicts the data analytics tools used in healthcare.

1.3.1 Data Integration Tools

Data integration in healthcare applications refers to the act of combining health data from a myriad of sources into a unified set of accumulated data that provide actionable business intelligence. Integration of multiple medical databases can be useful in identifying different methods of disease prevention, providing more sophisticated and personalized care, and reducing costs by avoiding overuse of resources.

As health data come from multiple, disparate sources like medical devices, wearable devices, etc., healthcare professionals face a major challenge in dealing with these unstructured data. Collecting and consolidating health data from various sources is extremely beneficial but still poses a major challenge to the health sector. The process of integrating huge volumes of medical data is complicated and presents significant challenges. Some of these challenges include:

- Lack of standard data formats
- Data privacy and confidentiality regulations
- Data format inconsistency among various healthcare applications
- Need for greater integration processing power
- Low end-user adoption

However, data integration tools can be used to integrate health data from multiple sources to generate meaningful insights from the data. Data integration tools include software and platforms that can aggregate data from disparate sources. The

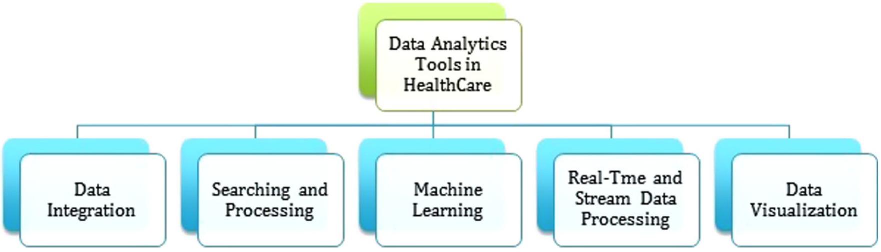

FIGURE 1.2 Data analytics tools in healthcare.

following are some data integration solutions that can be considered in healthcare organizations to make better and more efficient use of healthcare data.

- **Attunity** is an integration tool that can aggregate disparate data and files across all major databases, including cloud platforms, data warehouses, and Hadoop. It also supports the Health Level 7 (HL7) messaging standard, which is a healthcare standard. It can integrate and connect with web applications in real time.
- **Informatica** is an advanced, multi-cloud and hybrid data integration tool that can integrate data from multiple, disparate datasets, such as data warehouses, Hadoop, enterprise applications, message applications, and mid-range systems. It also provides data management tools for companies in the healthcare field to facilitate patient services with improved outcomes and reduced costs. Informatica's cloud integration allows administrators to integrate data with in-house applications, claims processing, etc., in health organization environments.
- **Information Builder** is a data integration and business intelligence tool that can measure and aggregate very large healthcare data collected throughout the patient lifecycle. These tools ensure the availability of data in real time across the healthcare environment and enhance the quality of health services.
- **Jitterbit** [8] is a single, secure, cloud-based data integration platform that aggregates structured and unstructured health data or clinical data retrieved from sources such as EHR. It enables more-efficient operations and provides complete access to health data in a format that can be used with other systems.
- **Magic** is a data integration tool that connects disparate systems for healthcare organizations. It ensures the best possible care of patients by keeping all health-related records up to date and available to all healthcare providers. Magic's integration platform combines diverse systems into a single interface *via* the graphical user interface.

1.3.2 Searching and Processing Tools

Since healthcare organizations deal each day with a large volume of unstructured data, there is a great need for data indexing, searching, and processing tools to optimize the efficient use of clinical data . These tools are employed for effective management of data that is stored in HDFS, which allows multiple files to be stored and retrieved at the same time in a big data environment. The following are some of the searching and processing tools used in healthcare.

- **Lucene** [8] is a scalable tool for indexing large blocks of unstructured text that provides advanced, full-text search capabilities. It can integrate easily with Hadoop to facilitate distributed text management.

- **Google Dremel** [9] is a distributed system that uses multi-level execution trees for interactive query processing of large datasets.
- **Cloudera Impala** is an MPP (Massively Parallel Processing) SQL query tool that supports the accessing of a massive volume of data stored in Hadoop clusters. It is a scalable and flexible parallel database technology that enables users to directly query data stored in HDFS file formats without requiring data format transformation.
- **Apache Hive** is a database query tool which facilitates querying, writing, and managing large datasets stored in distributed storage using a HiveQL language, which is a SQL-like query language. It allows structure to be projected onto data already in storage.

1.3.3 Machine Learning Tools

In a healthcare environment, machine learning tools are used to convert comprehensive health data into actionable knowledge that supports effective decision making to perform informed clinical activities [8].

- **Apache Mahout** [10] is an open-source, powerful, scalable machine learning library that runs on top of Hadoop MapReduce. The Mahout library facilitates the execution of distributed or scalable machine learning algorithms. The Mahout algorithms mainly focus on classification, clustering, and collaborative filtering techniques.
- **Skytree** is a big data machine learning tool that offers more accurate and faster predictive analytics models that are easy to use. These models enhance the processing of massive datasets in a precise manner without down sampling.
- **Apache SAMOA** (Scalable Advanced Massive Online Analysis) is an open-source-distributed, streaming machine learning framework that enables users to create distributed, streaming machine learning algorithms and to execute them on multiple DSPEs (Distributed Stream Processing Engines) [11].
- **BigML** [8] is a scalable, open-source machine learning tool that provides a framework to perform sophisticated machine learning workflows such as classification, regression, cluster analysis, anomaly detection, and association discovery. It includes a cloud infrastructure that can readily be integrated with machine learning features to build cost-effective, scalable, flexible, and reliable applications.

1.3.4 Real-Time and Streaming Data Processing Tools

With the rapid growth of massive Internet of Things (IoT), including wearables and medical devices in smart healthcare systems, there is a need for real-time and streaming data processing tools to carry out real-time processing of health data.

- **Apache Storm** [12] is an open-source, real-time data processing platform to process streaming data in real time. It also provides a fault tolerant, real-time computational framework. The applications of Apache Storm include real-time analytics, log processing, ETL jobs, continuous computation, distributed real-time processing, and machine learning.
- **SQL Stream Blaze** [12] is a streaming analytics platform that supports a variety of available streaming sources in all formats and at all speeds. Applications of SQL streaming include high throughput, data discovery, data wrangling, real-time threat detection, and analytics.
- **Apache Flink** [12] is a distributed processing engine, and it provides a framework for stateful computations over bounded and unbounded data streams. It is capable of both batch and stream processing and offers efficient, fast, accurate, and fault-tolerant handling of massive streams of data.
- **Apache Kafka** is an open-source, distributed event-streaming platform that provides a framework for building high-performance real-time streaming data pipelines, streaming analytics, and applications. There are five core application programming interfaces (APIs) in Kafka, namely Producer API, Consumer API, Streams API, Connector API, and Admin API, to facilitate message passing, storage, and stream processing services.

1.3.5 Visual Data Analytical Tools

Data visualization is a pictorial or graphical representation of data that enables users in any organization to analyze and understand the trends or patterns of data. In healthcare, data visualization tools help in creating and monitoring the dynamic data of a patient, presenting clinical records, identifying patterns and trends, and carrying out time-series analysis to improve healthcare services and public health policy.

- **SAS Visual Analytics** [13] is a web-based analytical tool that allows multiple users to access a massive amount of real-time data simultaneously from a LASR analytical server. It allows parallel networking by transferring data from one machine to another machine to access secure data quickly.
- **Tableau** [13] is a business intelligence visualization tool that transforms raw and large datasets into a defined format to provide real-time structured data to support decision making. One of the uses of Tableau has been to quickly diagnose genetic diseases and to help health practitioners in providing rapid treatment to the patients.
- **QlikView** [13] is a visualization tool that transmits related data from different sources into electronic medical records. It provides in-memory analysis and reduces the risk of medical error by tracking safety metrics and lowers the cost of delivering services to the patients. It ensures that all regulatory compliance in the healthcare system is delivered and maintained in a timely manner.

1.4 DATA ANALYTICS TECHNIQUES IN HEALTHCARE

Healthcare big data refers to multidimensional health data amassed from various sources, including medical imaging (X-ray, magnetic resonance imaging (MRI), and computed tomography (CT) scan images), structured data EHRs, biomedical signals (ECG, EEG, etc.), handwritten prescriptions, and data from wearables and medical devices. Since health data are dynamic and complex, they are difficult to manage and analyze using traditional techniques and technologies. There is a great demand for effective data analytics techniques to study these diverse data and to facilitate decision making. Figure 1.3 depicts the various big data analytics techniques used in healthcare, some of which are described below [5].

- **Data Mining** is useful for discovering patterns and for extracting meaningful information from large databases. With the rapid growth in massive health data, data-mining techniques have helped to search for new and valuable information (knowledge) from large complex databases in

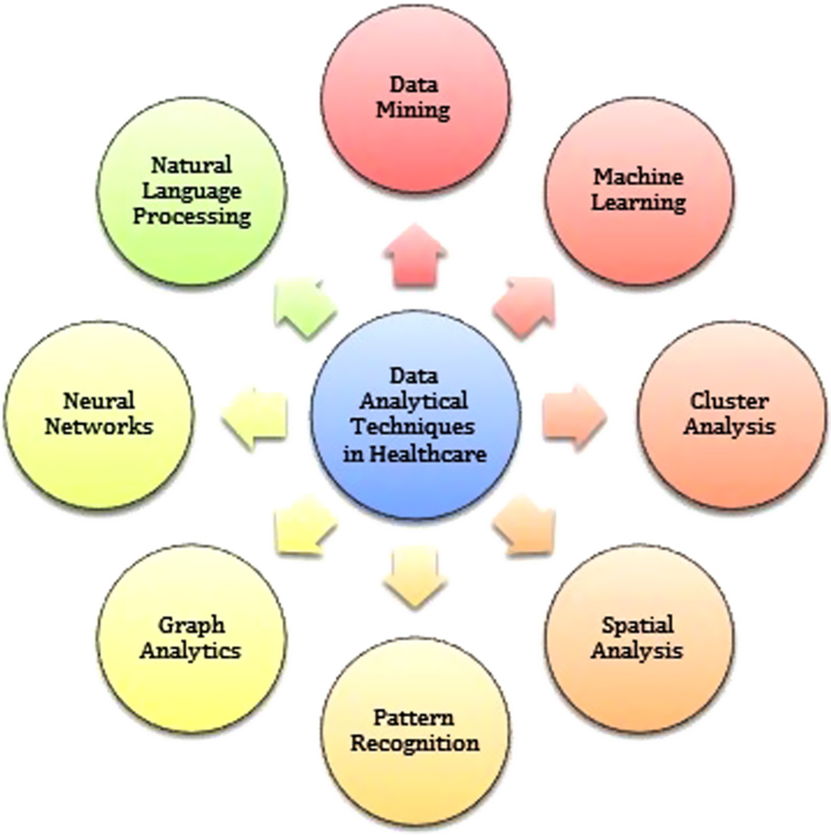

FIGURE 1.3 Data analytics techniques in healthcare.

healthcare systems, that facilitate the decision-making process. The main application areas of data mining include prediction and determination of various diseases, biosignal monitoring of patients, assistance in diagnosis and treatments, and exploratory data analysis in healthcare.

- **Machine Learning (ML)** in healthcare helps to analyze data and suggest outcomes. Applications of ML include prediction of diseases and diagnosis, drug discovery and manufacture, hospital performance assessment, smart health records, personalized patient care, etc.
- **Cluster Analysis** helps to discover hidden structures and clusters found in massive databases. In healthcare, cluster analysis helps to identify subgroups in the patient population defined by the patient's characteristics, disease severity, and treatment responses. Cluster analysis can be used, for example, to determine obesity clusters, to identify high-risk patient groups, and to identify population clusters with specific disease determinants in order to optimize treatment.
- **Graph Analytics** techniques analyze data in the form of a graph, where individual data entities are represented as nodes and the relationships that connect those entities are represented as edges. Graph analytics are used in healthcare to estimate healthcare fraud risk and hospital performance analysis based on quality measures [14].
- **Natural Language Processing (NLP)** is the technique used to make computers understand human speech and text. In healthcare, NLP has great potential to search, analyze, and interpret large volumes of patient-related datasets. Applications of NLP include the provision of efficient patient care, control of health service cost, extraction of meaningful health-related data from clinical notes [15], provision of training, consultations, and treatments, etc.
- **Neural Networks** are highly useful in analyzing, modeling, and interpreting healthcare data. Applications of artificial neural networks (ANN) to healthcare include clinical diagnosis, prediction of cancer, medical analysis, and drug development.
- **Pattern Recognition** methods help to enhance the clinical decision support system by focusing mainly on the patient's condition, based on symptoms and demographic information. They also help to improve public health screening in healthcare systems [16].
- **Spatial Analysis** plays a vital role in the effective use of geographic information systems (GIS) in healthcare, to facilitate health data exploration, data modeling, and hypothesis testing. Spatial analysis applications include direct patient care, epidemic disease prevention and intervention, assistance with strategic planning, etc.

1.5 APPLICATIONS OF DATA ANALYTICS IN HEALTHCARE

Big data and analytics have various applications in healthcare. With the digitization of patient records, medical history, X-rays, etc., there is a tremendous opportunity for data analytics to uncover useful information. Along with the volume of data is the sheer variety of data generated from different sources at different time points. The volume of and variation in data generated by this sector are what makes it a topic of great interest for data analysts. Conventional computing mechanisms and systems fail to provide real-time monitoring and preventive plans for patients as well as for the doctor. Hence, there is a need for smart strategies that decipher the incoming data to uncover trends and anomalies and which give recommendations for patients, helping doctors in their practice.

Some applications of data analytics in healthcare are as follows.

- **Image-Based Analytics**
  Image-based datasets are a common source of information and are primarily used by doctors for internal imaging. X-rays, mammography, CT, positron emission tomography–CT (PET–CT), ultrasound, and MRI are some of the imaging technologies commonly used for diagnostics [5, 17]. Many organizations and medical institutions release open datasets in a hope to foster research activities. Neuro-images and MRI of the brain are widely used for detecting tumors, an anomaly where the cells are enlarged and form solid neoplasms. Early-stage detection is necessary for effective treatment and recovery, and for decreasing the risk of mortality [18–20]. X-rays are carried out for detecting fractures, pneumonia, cancer, etc. A large amount of research has also been carried out for analyzing X-rays and CT scans to promote early detection of the novel coronavirus COVID-19, with minimal human intervention and interaction [21, 22].

  With the growing number of medical records, the reliance on computer-aided diagnosis and analysis is increasing [8]. High-performance computing and advanced analytical methods, like ML and optimization techniques, are aiming to minimize predictive errors. Countries with low doctor-to-patient ratios can benefit greatly from such machine-aided diagnostics.

  A significant challenge associated with image-based analytics is the amount of data generated. Images are space intensive. A single X-ray can take up several megabytes (MBs) of storage space. The quality of an image plays a crucial role in correct diagnosis and hence must not be compromised. What adds to the processing challenges is that the data are highly unstructured. Image processing techniques like segmentation, denoising (noise reduction), and enhancement form the pre-requisites before useful features, like color, contour, shape, pixel intensity, edges, etc., can be extracted to train models for classification and diagnosis.
- **Signal-Based Analytics**
  In a world full of wearable sensors, time-based signals are being generated at a frequency greater than that at which they can be processed. Wearable devices, like smart watches, smart rings, and fitness trackers, continuously track heart rate, blood pressure, sleep patterns, calories burned, etc. Apart from personal devices and gadgets, time-stream data are being generated by electrocardiograms (ECG), ventilators, electroneurograms (ENG), electroencephalograms (EEG), phonocardiograms (PCG), etc. [23]. Analysis of

these signals plays a significant role in deciding and prescribing medication, care regime, routine check-ups, etc. Readings of such signals can provide useful information concerning the current status of the patient.

With lifestyle changes over the past decade, heart diseases have become more common [24]. ECG signals, providing first-hand information on the well-being of the heart, can be sensed through carbon nanotubes and sent directly to doctors for real-time analysis, while the patient may still be at home [25]. Such a set-up reduces the need for patients to physically be present at the hospital or clinic to obtain a consultation. Furthermore, a live feed can also be sent to intelligent devices which have been trained to use ML models, to identify any anomaly in the signals [26]. On detection of any abnormal signs, the doctor can be contacted.

The mental state of a driver can be monitored using physiological signal analysis. This can prevent accidents which occur due to negligence or lack of attention by the driver [27]. In stressful times like today, when people are restricted by lockdown to their homes because of a global pandemic, mental health monitoring becomes more crucial than ever. Thus, detection and monitoring of neurological and mental health markers in patients can be a game changer [28].

Signals from a wide range of sensors can be generated continuously. These signals may be analyzed periodically for a routine check-up, or continuously, if the condition of a patient is critical. Nevertheless, all data still need to be saved in the patient's history, for future reference. Imagine the amount of data being generated by an individual wearing a smart watch 24/7. As with image-based analytics, such data are enormous in volume and rich in variety. Every sensor generates a different class of data. Such time-stream data needs to be placed in context in order to derive meaningful results regarding the current status of the patient. Furthermore, it adds another layer of difficulty to the analysis of signal-based data. The complexity of the problem is increased by the need for real-time stream diagnostics and analytics.

- **Clinical Diagnosis and Research**

  The study of signs, symptoms, and medical history of a patient to assess underlying conditions or disease is known as a clinical diagnosis. It usually does not involve any laboratory testing but uses the prior records of the patient, along with current symptoms and characteristics, to reach a preliminary level of diagnosis. Subsequently, additional tests can be ordered by the doctor to pinpoint the exact extent of the disease or to uncover anomalies and false symptoms. While a doctor uses years of experience to reach this first level of diagnosis, it is a time-consuming process to go through the medical history of the patient and past treatments to which the patient has been subjected.

  With the digitization of data, the entire medical history of every patient is stored as an EHR. Information like clinical notes, prescriptions, administrative data, laboratory test results, medical imaging data, the patient's personal data, etc., can all be easily retrieved and stored permanently because of ever-decreasing hardware cost [17, 29].

Data analytics can be helpful not only in the detection of disease at the earliest stage, but can also attain high levels of accuracy, predicting the timeline and disease development trajectory. Furthermore, data analytics can bring to the doctor's notice whether there is any change in the vital signs, indicating a deviation from a healthy state. It can provide transparency to patients and provide them with a more personalized experience of the entire healthcare management system [8, 30]. Healthcare organizations also benefit from data analytics as they help to provide cost-effective care and personalized predictions for each patient. Clinical data can also be useful for research purposes, such as the demographics of patients with a particular condition, predicting the sales of drugs and their profitability, identifying drug competitors, usage patterns of drugs, and effective drug design, uncovering inter-drug associations, etc. [8].

Clinical decision support systems (CDSS) [5, 23, 31] may therefore be an all-round solution for automated clinical diagnostics and research. But achieving systems with high levels of accuracy and efficiency is a big challenge. The data gathered from different sources in multiple formats over time adds volume, variety, and velocity to the data. Systems with high computational power to deal with structured, semi-structured, and unstructured data need to be implemented. Furthermore, handwritten clinical notes, prescriptions, and medical journals need advanced ML algorithms with concepts of NLP. Thus, the heterogeneity of clinical data remains currently the most significant challenge and an open research area.

- **Disease Transmission and Prevention**

  Some diseases are infectious and can be spread *via* direct or indirect contact with the infected person or carrier. To prevent the outbreak of such conditions, it becomes critical to study the means of transmission and to predict the spread of the disease to develop better mitigation plans and improved disease management strategies [32]. With the help of data analytics, mathematical and stochastic models can be generated to predict the outreach of the disease and to estimate its impact.

  Many researchers have studied the transmission of the novel coronavirus that emerged in Wuhan in late 2019 and which has spread throughout the world. It has been declared by the World Health Organization (WHO) to be the worst epidemic in the past two decades [33]. Based on early available data, symptoms, numbers of positive cases, and international travel history, a predictive model was developed by researchers to identify the extent of transmission and the risk it posed to human life [34]. Many government organizations have also funded projects to research the prevention and preparedness of individual countries. Massive amounts of data gathered globally have been used to develop preventive measures to stop or minimize further spread of the virus. Many countries opted for complete lockdown, halting businesses, closing schools and universities, banning travel, etc. Medical equipment, ventilators, masks, personal protective equipment (PPE), and sanitizers have been mass produced to deal with the prevailing situation. Predictive analyses for other diseases, like HIV/AIDS, have previously been carried out to enable early detection and treatment [35]. Transmission of all chronic illnesses could be prevented if the necessary measures are undertaken, although this would be costly [36].

Data play an essential role in understanding the transmission model of communicable diseases. The success of any disease prevention model depends largely on the timely identification of the pathogen. Hence, accurate classification, clustering, and associative models need to be developed, which can help physicians to prescribe medication and other treatments [36].

- **Precision Medicine**
  With an increase in the volume of data being generated on a daily basis from sensors, implants, EHR, clinical practices, etc., the potential to foster healthcare and medical functionalities has increased dramatically. Population health management, prevention of disease transmission, and CDSS, have all become possible. Ever-increasing sizes of EHR, and their dimensionality and variety, along with the incorporation of behavioral, social, and omics data, have given data analysts the power to propose and develop models for personalized patient care and precision medicine. The layered architecture of such models, working on different aspects of healthcare, has facilitated the development of such comprehensive and detailed healthcare solutions. By incorporating data from varied sources, and identifying relationships and patterns of interest, the entire healthcare solution framework has migrated from a disease-centric view to a patient-centric view. Diverse formats of data, the generation of bulk volumes of data, and the inherent uncertainty associated with sources of big data complicate the task of data curation. Transformation of raw data into useful facts and information is vital for these healthcare framework solutions to achieve their intended goals [8].

  Precision medicine promises better healthcare delivery by improving prognosis, diagnosis, and treatments being given to patients. Improving the quality of clinical practices translates directly to personalized healthcare routines for each individual, optimized for their direct benefit [37].

  Some challenges identified for precision medicine involve upgrading of ongoing clinical practices to incorporate newer, disruptive technologies in the field of big data, the ability to handle large volumes of data generated from different data sources and their analysis to achieve meaningful results [38]. High-scale cognitive computing, using advanced ML models, can drive data-driven analysis of biomedical big data [39].
- **Health Insurance**
  Insurance agents and companies today enjoy an extensive database of customers belonging to different demographics. The application of data analytics to this large dataset can help the agents to identify patterns, clusters, and typical human behavior. With insurance models and schemes moving

online, consumers expect transparency and an accurate breakdown of the money they are spending on health insurance. Thus, the agencies cannot add some hidden costs to make additional profits. Competition among agencies is stiff and, hence, attracting customers with lucrative benefits is how these business agencies function [40].

Insurers today need to focus on individual customizable plans for individuals, instead of targeting groups of people. Predictive analysis is an essential branch of analytics, which is widely used in such scenarios. It identifies the appropriate coverage for the needs of people and further customizes plans based on individual inputs. Smart models deployed online first take the necessary information from users before developing proposals for each individual. This personal experience and the tailor-made environment are valuable for customers.

The insurance costs can be intelligently calculated based on the current health and medical history of an applicant. For instance, with the growing obesity rate, the insurers may calculate the potential risk associated with an individual applicant and modify the premium for health insurance accordingly. Furthermore, as obesity increases the risk of heart failure, preliminary tests may be advised before the policy can be issued [41].

Thus, health insurance companies use big data analytics to deliver a personalized experience to each individual, to predict the occurrence of fraud before it is realized, and to fast-track claims by predictive analytics.

- **Service Delivery Systems**

Today, healthcare providers across the globe are facing competition and understand the need to deliver the highest quality of services to their customers. Companies work on low profit margins and hence need to optimize their service delivery systems to minimize the cost of providing the service.

The easiest way to do this is by automating the healthcare services and reducing human intervention with respect to suggesting remedial actions, preventive measures, prescriptions, diagnostics, etc. The majority of the costs incurred by such organizations are on high-quality medical equipment that needs to be upgraded over time, to keep pace with global technological advances. The research and development cost for new medicines is also subject to extensive clinical trials before they can be rolled out to the public for consumption, leading to high costs [42].

Furthermore, medical devices, equipment, and sensors continuously generate large volumes of data which can be used for forecasting and decision making. Hence, data analytics can serve as a powerful tool for improving healthcare delivery systems.

A service delivery network usually consists of multiple organizations working together collaboratively to provide an overall healthcare service delivery system. These organizations are closely connected with one another and share data among themselves to make smart, informed decisions. Players involved in such networks may include sales representatives, doctors, physicians, insurance personnel, laboratory technicians, hospital staff, equipment vendors, etc. They continuously explore ways to understand data being generated from different sources and to understand the impact such data have on their policies and services. This co-development and deployment environment aims to reduce the cost of healthcare services [43, 44].

1.6 CHALLENGES ASSOCIATED WITH HEALTHCARE DATA

With the use of big data analytics, the healthcare industry has improved in many aspects, including operational efficiency in healthcare management, reduction in healthcare costs, improved drug discovery, higher-quality healthcare services, personalized patient care, effective treatments, and improved clinical outcomes. Despite the benefits, however, big data introduces several challenges.

- One of the foremost challenges facing the healthcare industry nowadays involves capturing health-related data. Due to the extensive use of IoT devices and wearables in the smart healthcare system, which constantly generate massive volumes of streaming data, the capture and processing of data has become a major challenge. Lack of efficient governance practices for data sources is another challenge for the capture of data from heterogeneous data sources, a problem which can lead to inaccurate data.
- Due to the exponential volume growth of streaming health data, data storage has become a primary challenge for the healthcare industry. Most health organizations prefer in-house data storage so that they can have control over data security and data access, but, with the rapid growth of health data, in-house data storage infrastructures become difficult to scale up as maintenance and scale-up costs are high.
- The privacy and confidentiality of medical data, in the form of a patient's health data, are of utmost importance in healthcare. Data sharing among the various practitioners in a healthcare system escalates the need for privacy, while informed consent issues are another challenge faced by data analytics in healthcare.
- Since healthcare data are generated from multiple sources, such as medical images, wearable sensors, and EHRs, there are no fixed unified standards for these data, leading to difficulties in consolidating and processing of the data. Therefore, the lack of data protocols and standards are one of the governance challenges facing healthcare data analytics.
- As digital healthcare has no geopolitical boundaries, health services are available across international and national borders. It is difficult to form uniform legislation because medical licenses, privacy of patient data, and the advertisement and marketing of healthcare services may vary between countries, representing major challenges for the healthcare industry.
- Data analytics in healthcare also pose new ethical and legal challenges, including personal autonomy, risk of compromising the patient's privacy, need for informed consent, trust and transparency when using biomedical big data.

1.7 CONCLUSION

Data analytics with big data in healthcare is still at the developing stage and advances in tools and techniques will improve and their applications will expand. In addition, establishing proper standards and governance of data, ensuring data privacy and security, and updating the healthcare systems continuously are some of the challenges faced by the healthcare industry. Improving communication and data sharing among related sectors in healthcare would increase the overall efficiency by providing value-added services, with minimal additional costs incurred.

REFERENCES

1. A. Belle, R. Thiagarajan, S. M. R. Soroushmehr, F. Navidi, D. A. Beard, K. Najarian. Big data analytics in healthcare. *Biomedical Research International*, 2015, 16 pages, 2015. http://dx.doi.org/10.1155/2015/370194.
2. A. Kankanhalli, J. Hahn, S. Tan, G. Gao. Big data and analytics in healthcare: Introduction to the special section. *Information Systems Frontiers*, 18, 233–235, 2016.
3. M. J. Ward, K. A. Marsolo, C. M. Froehle. Applications of business analytics in healthcare. *Business Horizons*, 57, 571–582, 2014.
4. S. Kumar, M. Singh. Big data analytics for healthcare industry: Impact, applications, and tools. *IEEE*, 2 (1), 48–57, 2019.
5. N. Mehta, A. Pandit. Concurrence of big data analytics and healthcare: A systematic review. *International Journal of Medical Informatics*, 114, 57–65, 2018.
6. Y. Wang, N. Hajli. Exploring the path to big data analytics success in healthcare. *Journal of Business Research*, 70, 287–299, 2017.
7. N. El aboudi, L. Benhlima. Big data management for healthcare systems: Architecture, requirements, and implementation, Hindwai. *Advances in Bioinformatics*, 2018. https://doi.org/10.1155/2018/4059018.
8. V. Palanisamy, R. Thirunavukarasu. Implications of big data analytics in developing healthcare frameworks–A review. *Journal of King Saud University-Computer and Information Sciences*, 31(4), 415–425, 2019.
9. S. Melnik, A. Gubarev, J. J. Long, G. Romer, S. Shivakumar, M. Tolton, T. Vassilakis. Dremel: Interactive analysis of web-scale datasets. In: *36th International Conference*, 2010.
10. C. E. Seminario, D. C. Wilson. Case study evaluation of Mahout as a recommender platform. In: *6th ACM Conference on Recommender Engines (RecSys 2012)*, pp. 45–50, 2012.
11. https://www.softwaretestinghelp.com/big-data-tools/.
12. https://www.predictiveanalyticstoday.com/top-open-source-commercial-stream-analytics-platforms/.
13. https://www.yourtechdiet.com/blogs/impact-data-visualization-healthcare/.
14. N. Downing, A. Cloninger, A. Venkatesh, A. Hsieh, E. Drye, R. Coifman, et al. Describing the performance of U.S. hospitals by applying big data analytics. *PLoS One* 12(6), e0179603, 2017.
15. A. Khalifa, S. Meystre. Adapting existing natural language processing resources for cardiovascular risk factors identification in clinical notes. *Journal of Biomedical Informatics*, 58, S128–S132, 2015.
16. D. D. Luxton, J. D. June, A. Sano, T. Bickmore. Intelligent mobile, wearable, and ambient technologies for behavioral health care. *Artificial Intelligence in Behavioral and Mental Health Care*, Elsevier, 137, 2015.

17. B. Ristevski, M. Chen. Big data analytics in medicine and healthcare. *Journal of Integrative Bioinformatics*, 15 (3), 1–5, 2018.
18. A. R. Kavitha, C. Chellamuthu. Brain tumour detection using self-adaptive learning PSO-based feature selection algorithm in MRI images. *International Journal of Business Intelligence and Data Mining*, 15 (1), 2019.
19. S. Tchoketch Kebir, S. Mekaoui, M. Bouhedda. A fully automatic methodology for MRI brain tumour detection and segmentation. *The Imaging Science Journal*, 67(1), 42–62, 2019.
20. T. V. N. Rao, H. Katukam, D. Guvva. Early brain tumour detection in MRI using enhanced segmentation approach. image, 8, 9, 2019.
21. L. Brunese, F. Mercaldo, A. Reginelli, A. Santone. Explainable deep learning for pulmonary disease and coronavirus COVID-19 detection from X-rays. *Computer Methods and Programs in Biomedicine*, 196, 105608, 2020.
22. A. Jacobi, M. Chung, A. Bernheim, C. Eber. Portable chest X-ray in coronavirus disease-19 (COVID-19): A pictorial review. *Clinical Imaging*, 64 (April), 35–42, 2020.
23. C. K. Reddy, C. C. Aggarwal. An introduction to healthcare data analytics. In *Healthcare Data Analytics*, 2015, pp. 1–18.
24. S. Dalal, V. P. Vishwakarma. GA-based KELM optimization for ECG classification. *Procedia Computer Science*, 167, (2019), 580–588, 2020.
25. M. Bansal, B. Gandhi. IoT & Big Data in Smart Healthcare (ECG Monitoring). In *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, 390–396, IEEE, 2019, February.
26. S. Dalal, V. P. Vishwakarma, V. Sisaudia. ECG classification using Kernel extreme learning machine. In *2nd IEEE International Conference on Power Electronics, Intelligent Control and Energy systems (ICPEICES-2018)*, pp. 988–992, 2018.
27. Barua, S., Ahmed, M. U., & Begum, S. Distributed multivariate physiological signal analytics for drivers' mental state monitoring. In *International Conference on IoT Technologies for HealthCare*, pp. 26–33, Springer, Cham, 2017, October.
28. M. Neumann, O. Roesler, D. Suendermann-oeft, V. Ramanarayanan. On the utility of audiovisual dialog technologies and signal analytics for real-time remote monitoring of depression biomarkers. In *Proceedings of First Workshop on Natural Language Processing for Medical Conversations*, pp. 47–52, 2020.
29. Belle, A., Thiagarajan, R., Soroushmehr, S. M., Navidi, F., Beard, D. A., & Najarian, K. Big data analytics in healthcare. *BioMed Research International*, 2015, 2015.
30. Wang, Y., & Hajli, N. Exploring the path to big data analytics success in healthcare. *Journal of Business Research*, 70, 287–299, 2017.
31. Shafqat, S., Kishwer, S., Rasool, R. U., Qadir, J., Amjad, T., & Ahmad, H. F. Big data analytics enhanced healthcare systems: A review. *The Journal of Supercomputing*, 76(3), 1754–1799, 2020.
32. Wong, Z. S., Zhou, J., & Zhang, Q. Artificial intelligence for infectious disease big data analytics. *Infection, Disease & Health*, 24(1), 44–48, 2019.
33. Koubâa, A. Understanding the covid19 outbreak: A comparative data analytics and study. arXiv preprint arXiv:2003.14150, 2020.
34. Kucharski, A. J., Russell, T. W., Diamond, C., Liu, Y., Edmunds, J., Funk, S., ... & Flasche, S. Early dynamics of transmission and control of COVID-19: A mathematical modelling study. *The Lancet Infectious Diseases*, 20(5), 553–558, 2020.
35. Das, N., Das, L., Rautaray, S. S., & Pandey, M. Detection and prevention of HIV aids using big data tool. In *2018 3rd International Conference for Convergence in Technology (I2CT)*, pp. 1–5. IEEE, 2018, April.
36. Razzak, M. I., Imran, M., & Xu, G. Big data analytics for preventive medicine. *Neural Computing and Applications*, 32(9), 4417–4451, 2020.

37. Panayides, A. S., Pattichis, M. S., Leandrou, S., Pitris, C., Constantinidou, A., & Pattichis, C. S. Radiogenomics for precision medicine with a big data analytics perspective. *IEEE Journal of Biomedical and Health Informatics*, 23(5), 2063–2079, 2018.
38. Hulsen, T., Jamuar, S. S., Moody, A. R., Karnes, J. H., Varga, O., Hedensted, S., ... & McKinney, E. F. From big data to precision medicine. *Frontiers in Medicine*, 6, 34, 2019.
39. D. Cirillo, A. Valencia. Big data analytics for personalized medicine. *Current Opinion in Biotechnology*, 58, 161–167, 2019.
40. Gupta, S., & Tripathi, P. An emerging trend of big data analytics with health insurance in India. In *2016 International Conference on Innovation and Challenges in Cyber Security (ICICCS-INBUSH)*, pp. 64–69, IEEE, 2016, February.
41. Revels, S., Kumar, S. A., & Ben-Assuli, O. Predicting obesity rate and obesity-related healthcare costs using data analytics. *Health Policy and Technology*, 6(2), 198–207, 2017.
42. Alotaibi, S., Mehmood, R., & Katib, I. The role of big data and twitter data analytics in healthcare supply chain management. In *Smart Infrastructure and Applications*, pp. 267–279, Springer, Cham, 2020.
43. M. A. Pikkarainen. Data as a driver for shaping the practices of a preventive healthcare service delivery network. *Journal of Innovation Management*, 1, 55–79, 2018.
44. M. Usak, M. Kubiatko, M. Salman. Health care service delivery based on the Internet of things: A systematic and comprehensive study. *International Journal of Communication Systems*, 33, 1–17, 2019.