A Systematic Review of Healthcare Big Data

Rakesh Raja, Indrajit Mukherjee, and Bikash Kanti Sarkar

Department of Computer Science and Engineering, Birla Institute of Technology, Mesra, India

Correspondence should be addressed to Rakesh Raja; rajarakeshchauhan@gmail.com

Received 24 December 2019; Revised 14 March 2020; Accepted 20 June 2020; Published 13 July 2020

Copyright © 2020 Rakesh Raja et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Over the past decade, data recorded (due to digitization) in healthcare sectors have continued to increase, intriguing the thought about big data in healthcare. There already exists plenty of information, ready for analysis. Researchers are always putting their best effort to find valuable insight from the healthcare big data for quality medical services. This article provides a systematic review study on healthcare big data based on the systematic literature review (SLR) protocol. In particular, the present study highlights some valuable research aspects on healthcare big data, evaluating 34 journal articles (between 2015 and 2019) according to the defined inclusion-exclusion criteria. More specifically, the present study focuses to determine the extent of healthcare big data analytics together with its applications and challenges in healthcare adoption. Besides, the article discusses big data produced by these healthcare systems, big data characteristics, and various issues in dealing with big data, as well as how big data analytics contributes to achieve a meaningful insight on these data set. In short, the article summarizes the existing literature based on healthcare big data, and it also helps the researchers with a foundation for future study in healthcare contexts.

1. Introduction

The era of big data has opened the door in the healthcare industry as a response to the digitization of healthcare data. Over the past decade, the exponential growth in data [1] has introduced a new domain called big data within the field of information technology (IT) and data science. The term big data is commonly used to describe a large amount of data which are too big and not easy to handle using traditional techniques of the database management system. The idea of big data is not very new, but the manner in which it is characterized is continuously changing. In 1997, Michael Cox and David Ellsworth introduced the term “big data” for the first time in the world during a paper conferred at an IEEE conference to explain the visual representation of data and the difficulties it exhibits to computer systems [2]. The data that go beyond the processing capacity of traditional database management systems are termed as big data. These data are so large that they do not fit the structure of typical database management systems.

The notion of big data given by Doug Laney was characterized by volume, velocity, and variety known as 3Vs [3]. Generally, big data can be defined as a collection of very large amount of data with a wide range of types, making it very hard to process using conventional database management systems. As per the author in [4], big data is a data set with large volume, high speed, and high diversity that requires a new style of processing to facilitate decision-making and exploring knowledge and optimization of techniques. Typically, a massive volume of data may be referred to as big data when capturing, analysing, and visualizing of data with current technologies are overwhelming. Big data plays an important role in the current digital era due to the significant advancement of healthcare technologies [5]. As the sources of big data concerned in healthcare industries and various sectors are well known for their volume and diversity, hence, the healthcare domain gained its effect through the impact of big data. The healthcare industries have generated enormous amount of healthcare data over the past couple of years. These healthcare data are similar to the big data in terms of their characteristics, therefore named as healthcare big data. Healthcare data generally incorporate electronic medical records (EMRs) such as patient’s medical history, physician notes, clinical reports, biometric data, and other medical data related to health. All these data together result in healthcare big data. The evolution of healthcare big data is...
advance and cost-effective for both public and private healthcare. The success of healthcare applications with regard to big data entirely relies upon the underlying architecture and use of suitable tools as proven in pioneering research efforts. It also gives an idea of the analytics of big data in healthcare systems. More specifically, big data analytical tools and techniques have the potential to improve the quality of medical services and reduce the medical cost of patients by exploring the association and understanding the nature of healthcare data. In 2016, Kohli et al. discuss how electronic health records (EHR) facilitate integration of patient health history for planning safe and proper treatment [6]. More about big data and healthcare big data definition are presented in Table 1.

2. Systematic Literature Review (SLR) Method

The purpose of the research process for conducting a systematic literature review (SLR) (based on the relevant articles and studies published in academic journals) focuses on the following objectives:

- Analysing different perspectives about the concept of big data in healthcare
- Exploring the origins of healthcare big data
- Identifying tools and techniques for healthcare big data analytics
- Highlighting the potential advantages and applications of big data in healthcare
- Drawing attention to overcome the big data challenges in healthcare

By discussing these goals in depth, the systematic review aims to assist in understanding the overall context of big data and its applications in the healthcare sector.

2.1. Research Questions. The following are the key research questions that are to be addressed for conducting the SLR of the proposed study:

RQ1. What are the characteristics of big data in the healthcare domain?
RQ2. What are the challenges and opportunities of healthcare big data?
RQ3. What are the features of big data analytics in healthcare?
RQ4. What techniques are used for big data analytics in healthcare?
RQ5. What are the applications of big data analytics in healthcare?
RQ6. What research has been pursued in healthcare big data since 2015?

2.2. SLR Protocol. Based on the SLR protocol designed in [12], this literature review follows the below mentioned guidelines.

2.2.1. Search Strategy. The two main electronic research databases: ScienceDirect and IEEE Xplore, were used to search for the collection of relevant articles related to the proposed research. However, some good and relevant works published by Springer publ. are also included in the present study.

2.2.2. Search String. The keywords defined by the authors for search process were “Big data,” “Healthcare,” and “Big data analytics” in context to the research domain. To conduct an SLR, the search process was carried out to identify the relevant articles for addressing the research questions based on predefined keywords using Boolean operators.

2.2.3. Selection Criteria. The authors agreed to select articles based on the following inclusion-exclusion criteria:

(1) Inclusion Criteria

- The articles relevant to healthcare big data and big data analytics
- The articles published during year 2015 to 2019
- The articles from journals publications only
- The articles written in the English language

(2) Exclusion Criteria

- The articles not in the range of 2015 to 2019
- The articles other than journal publications

2.2.4. Study Selection Process. The methodology for the literature review process was performed in different stages. The details of the study selection process of SLR are shown in Figure 1. Initially, all the articles relevant to big data, healthcare big data, and big data analytics were selected in the preliminary stage of screening as per the searching keywords. Based on inclusion-exclusion criteria, these articles were screened in the first stage, and irrelevant articles which were not published between 2015 and 2019 were excluded. During the second stage of screening, the selected articles were further screened on the basis of title, abstract, and keywords. The articles which were not associated with the proposed study were excluded. Finally, in the last stage of screening, these articles were further screened on the basis of abstract using the Boolean AND operator applied to all the three authors’ defined searching keywords. As a result, 34 articles relevant to the research domain were selected from 8355 articles, for further study by the authors.

2.2.5. Quality Assessment. During the review, quality assessment plays a significant role in the SLR protocol. The quality assessment of articles was done by all authors after the analysis and evaluation of abstracts of selected articles. These articles were selected with respect to each defined key research question based on inclusion-exclusion criteria.
Table 1: Big data and healthcare big data definition.

<table>
<thead>
<tr>
<th>Sources</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>[7]</td>
<td>Healthcare big data can be defined as digitalized version of health information which is so vast and complex that they are not easy to manage using traditional software and/or hardware, nor can they be easily handled using conventional data management tools and methods</td>
</tr>
<tr>
<td>[8]</td>
<td>Big data means enormous amount of digital data that organizations and governments collect about individuals and their general surrounding environments where the generated data are about 2500 petabytes or even more</td>
</tr>
<tr>
<td>[9]</td>
<td>Big data in healthcare refers to the data sets with ( \log(n \times p) \geq 7 ), and that they have high variety and high-speed characteristics</td>
</tr>
<tr>
<td>[10]</td>
<td>Big data can be defined as wealth of information described by massive volume, high velocity, and wide variety in order to have specific technology and analytical techniques to transform it into worth</td>
</tr>
<tr>
<td>[11]</td>
<td>Healthcare data generally incorporates electronic health records (EHRs) such as patient’s medical history, physician notes, clinical reports, biometric data, and other medical data related to health, as well as social media posts such as blog posts, tweets, Facebook post notifications, and publications in medical journals</td>
</tr>
</tbody>
</table>

Proposed research questions (RQ)

RQ1. Characteristics of big data in healthcare domain
RQ2. Challenges and opportunities of healthcare big data
RQ3. Features of big data analytics in healthcare
RQ4. Techniques used for big data analytics in healthcare
RQ5. Applications of big data analytics in healthcare
RQ6. Research pursued in healthcare big data since 2015

Figure 1: SLR process.
3. Big Data: A Comprehensive Overview

3.1. Big Data in General. Big data refers to a collection of extensive and complicated data sets that are hard to handle using conventional database systems. As per the zdnet.com, big data pertains to the tools and techniques that allow an organization to generate, exploit, and maintain vast amounts of data with storage facilities. Each one of us is continuously producing enormous amount of data. And, big data is being generated by every computerized system as well as social networking sites. It is transmitted by the digital system, sensor devices, cameras, handheld devices, smartphones, and their applications [13]. Big data arrives at an unprecedented rate, large data size, and greater diversity from various sources. To extract significant worth from such large amount of data, we need high computational power, analytical capabilities, and expertise. This explosion of data attempts to change the opinion of people to think about everything in terms of big data. In recent times, transactional data, web-based data, sensor data, and electronic medical data keep developing with rapid speed. These data can be classified into web-based data, sensor-based data, demographic data, transactional data, and machine-generated data [14] (as stated below):

- Web-based data are acquired from social networking sites such as Facebook, Twitter, and Blogs
- Machine-generated data are extracted from sensor-based devices and other gadgets
- Transactional data are retrieved from biometrics, vital sign, radiology, and other medical images
- Human-generated data comprise E-mails, doctor’s prescriptions, and digitalized version of medical reports

This remarkable development in data growth has led to this new concept known as big data. In article [15], it is stated that big data is a complex set of data that has a significant impact on the ability of conventional data warehouses to store, maintain, perform, and analyse data. A formal definition of big data has been provided in [10]. It is stated there as follows: big data is a wealth of information described by huge quantity, high velocity, and wide variety in order to have specific technology and analytical techniques to transform it into worth. Looking at it another way, the McKinsey Global Institute defines big data as data sets whose size exceeds the capability of conventional database systems to collect, store, maintain, and analyse data. According to the authors in [16], big data is the assemblage of data collected from different sources such as corporate databases, websites, maps, movies, and public databases.

3.2. Characteristics of Big Data. The common characteristics of big data are illustrated in the following:

- **Volume:** this implies data size usually measured in terabytes (TB = 10^12 bytes), petabytes (PB = 10^15 bytes), and zettabytes (ZB = 10^21 bytes), and so forth
- **Velocity:** this indicates the rate of generation of data
- **Variety:** this refers to the nature of data which big data can include such as structured, semistructured, and unstructured data
- **Veracity:** this refers to the trustworthiness of the data
- **Value:** the term itself is related to the worth of data being extracted

Apart from the abovementioned features of big data, several researchers and scientists have introduced new features to big data due to various applications available; i.e., the big data definition keeps changing according to the advancement of technology, data storage, and data transmission rate, as well as other system capabilities. The different explanations for the definition of big data are from 3Vs to 4Vs [17, 18], 5Vs [19], and 10Vs [20]. In particular, these dimensions are expanding as time goes by; and we currently have 42 distinct dimensions for big data till 2017 as per [21], and also the dimensions will keep on expanding as the big data evolves further. Figure 3 describes the generic notion of big data.
4. Big Data in Healthcare Domain

4.1. Healthcare Big Data. A pioneering renovation is taking place in the healthcare industry. The healthcare industry is generating a large volume of healthcare data due to the advancement in technology and digitization of medical records. In recent years, health information technology (HIT) has developed the power to generate, store, and transmit data electronically worldwide within seconds and also has the potential to deliver tremendously better productivity and service quality to healthcare. It allows each stakeholder in the healthcare sector to possess his/her own database of patients’ medical records in a digital form. The healthcare sectors have produced huge amounts of healthcare data by keeping records, consent and regulatory requirements, and patient care [11]. All these data together form healthcare big data. To be more specific, healthcare big data can be defined as electronic medical records (EMRs) which incorporates patient’s medical history, physician notes, clinical reports, biometric data, and other medical data related to health, as well as social media posts such as blog posts, tweets, Facebook post notifications, and publications in medical journals [11]. Importantly, the exponential growth of healthcare data is another major issue in the current healthcare information systems (HISs). This transformation is not only about the large volume of healthcare data; however, we are also experiencing an exponential rise in the velocity at which these data are generated, as well as large diversity of medical data.

The evolution of advancement in technologies like sensor systems, cameras, and smartphones is a significant source of healthcare data. Every day, new sources of data are introduced. This makes it much more difficult to process or analyse big data in healthcare using common database management tools. Typically, when massive volume of healthcare data are captured, stored, and analysed properly in order to gain insight, it will enhance the healthcare service...
outcome through smart decisions and also reduce healthcare costs. However, effective data analytical tools and techniques as well as powerful computing systems are required for this purpose. Healthcare big data analytics (BDA) in particular has started to emerge as a promising tool for taking care of issues in numerous healthcare disciplines. In addition, the role of a data analyst is to mine the big data, exploring the association and understanding trends and patterns of healthcare data. This enhances the health and improves the quality of life of an individual, as well as provides appropriate early-stage treatment at low cost.

The amount of data stored in healthcare sectors continued to increase curiosity about healthcare big data. There is an enormous amount of data ready to be analysed. One of the principle motivations behind big data is to focus on healthcare. The basic motive of nations around the world is to improve the healthcare facilities and decrease the medical costs. However, the revolution of massive volume of data in healthcare remains a barrier for achieving this goal. Electronic healthcare data from all around the world were estimated at 500 petabytes in 2012, reaching 25 petabytes by 2020 [22]. Thus, healthcare can be described as a wide variety of services offered by medical professionals to people, families, or societies to encourage, maintain, or restore better health. The quality of the healthcare system is significant because it determines hospital sustainable growth and helps people to maintain the optimal state of health. In certain cases, the quality of healthcare services is too high, and it ends up costly for patients. Consequently, it is essential to address the key healthcare procedures and related quality parameters that act in collaboration to ensure the best possible outcomes for patients and reduce the healthcare costs.

4.2. Sources of Healthcare Big Data. This section deals with several important sources of healthcare data. Big data in healthcare can revolutionize the medical field through early-stage disease detection using adequate analytical tools and techniques by incorporating and analysing health-related information in a comprehensive manner. Currently, the evolution of advancement in technologies like sensor systems, cameras, wearable devices, and mobile applications is widely used in the domain of the medical field [23, 24]. As a result, more medical information is being explored in a consistent manner. Data in medicinal services are fragmented and dispersed, originating from disparate sources with multiple formats [25]. The facts confirm that information on health is large and heterogeneous. The reason is on the ground that they originate from various internal and external sources accessible at multiple locations. External sources include web data, social media data, and machine-generated data, and internal sources include transactional data, biometric data, and human-generated data. Various healthcare data and their sources are summarized in Table 4.

4.3. The 5Vs of Healthcare Big Data Characteristics. In this section, the important Vs about healthcare data are briefly stated. The five key characteristics that have been found in most literature [12, 35] to define healthcare big data are as follows:

**Volume.** Based on the general discussion of big data, healthcare data are a perfect case of big data. The volume refers to the data size that grows exponentially day to day, and by 2020, the volume of big data may reach to 44 zettabytes [36]. Compared to most of the industries, the healthcare sector generates massive amounts of data in the form of electronic medical records (EMRs), biometric data, clinical data, radiology images, genomics, etc. All these data collectively form healthcare big data [37–39]. Obviously, the utilization of several tools such as Hadoop, MapReduce, and MongoDB is getting more popular among healthcare organizations due to their ability to store and measure massive volume of data [40, 41].

**Velocity.** Velocity refers to the speed at which data are generated, as well as data acquired from various healthcare systems [42].

**Variety.** Variety refers to the heterogeneity and diversity of data. The healthcare industry generates and collects data at a staggering rate from different sources such as social networking sites, sensor devices, cameras, and smartphones. However, these healthcare data may be in any one of the forms, structured, unstructured, or semistructured. Example of structured data is clinical data, whereas data such as physician notes, images, social media data, mobile data, and radiograph films are unstructured or semistructured. Figure 4 depicts the types of healthcare data, along with examples.

**Veracity.** The veracity characteristics of healthcare data refer to the trustworthiness of the data, which in this context is equivalent to quality assurance of data. It gives the degree of authenticity about healthcare knowledge.

**Value.** Value is the most important and distinctive characteristics of all the 5Vs of healthcare big data, as it has the ability to transform healthcare data into worth of information. Its concept is exactly in line with that of healthcare data.

5. Big Data Challenges in Healthcare

The evolution of big data introduces several challenges, constraints, and problems due to exponential growth of healthcare data. Big data is constantly changing, and this change of data presents a lot of challenges in storing, analysing, and retrieving the massive volume of data. Certainly, the conventional database systems could not be used to store, process, and extract the information due to its massive size and diversity of data.

The main challenges encountered by healthcare BDA are as follows:

- Quality and storage of data
- Data analysis of good quality
- Expertise in data analytics
Data security and confidentiality

Multiple sources of data

Healthcare big data challenges encountered are no different. Big data characteristics are the main issues that need to be addressed. It is vital to move towards big data technology in order to provide better medical facilities. Big data technology, however, introduces a potential risk to certain categories.

5.1. Issues in Healthcare Big Data. Big data issues that generally occur in the healthcare organizations are covered by four main categories [35, 43]:

Data Governance. Data management and regulation is the governance of data. As the healthcare sector moves towards data analytics, data governance is a major challenge. Healthcare data generated are diversified in nature, requiring standardization and governance.

Economic Challenges. The facilities in the medical field between patients and healthcare professionals throughout clinical visits depend on the paid service. Subsequently, advancement in technologies associated with this process places a burden on the medical community and generates an unnecessary impact for the personnel against such unpaid services.

Big Data Technology Challenges. Big data in healthcare is enormous and highly fragmented which causes problems in quality of information, as well as technology-wise, big data creates a barrier to accomplish the healthcare vision [44].

Security and Privacy Issues. In the era of big data, the privacy of healthcare data must be seriously considered due to the potentially sensitive information about individual healthcare stakeholders. Healthcare data are highly sensitive data which must be secured from unauthorized access so that it cannot be made publicly available, as well as healthcare fraud can also be prevented from attackers. Therefore, data security is one of the most important challenging tasks in the healthcare domain.

While studying and analysing several published research papers with reference to the SLR protocol, this research focuses on how recent developments in ICT (information and communication technologies) together with big data techniques can be effectively incorporated to address these challenges of healthcare big data and make a significant contribution towards healthcare services [45–51]. Based on [17, 52], we the authors classify healthcare BDA into three categories, namely, descriptive analytics, predictive analytics, and prescriptive analytics. Among these different BDA techniques, this literature review reflects that there are various tools, for example, Hadoop [53] and MapReduce [54] that have been developed for healthcare big data management. These are described in Section 6. A few of the well-known BDA techniques used in the areas of healthcare are described in Table 5. The categories provided in Table 5 are drawn from the literature [12, 66–68].

6. Big Data Analytics in Healthcare

Healthcare BDA has a potential to improve the quality of care and reduce the medical cost of patients by finding the associations from massive volume of healthcare data, thereby offering a wider perspective of clinical expertise based on medical evidence and various tests. Advanced analytical tools and techniques used in healthcare systems provide services that satisfy a growing need and enable healthcare agencies to process massive volume of data, analyse it in real time, and extract knowledge from medical records of all patients. In 2017, Palanisamy and Thirunavukarasu have presented various analytical avenues that exist in the patient centric healthcare system from the perspective of various stakeholders [69]. The main goal of the article is to assist researchers and data scientists to make
6.1. Types of Healthcare Big Data Analytics. BDA mainly perform three types of analytics, namely, descriptive analytics, predictive analytics, and prescriptive analytics. The descriptive analytics facilitates to explore insights and allows healthcare practitioners to understand what is happening in a given situation [73, 74]. In the context of healthcare data, the descriptive analytics analyses the data gathered in order to interpret, understand, summarize, and visualize significant health-related information. On the other hand, predictive analytics assist healthcare stakeholders to identify the healthcare services and responding appropriately according to the requirements of patients. It also enables clinicians to be capable of making patient-related decisions on the basis of system predictions [73, 74]. Predictive analytics involves various statistical techniques used to analyse and extract valuable insights from big data [17]. Hadoop/MapReduce is one of the most widely used techniques to develop a predictive model for healthcare systems. Prescriptive analytics is comparatively a modern type of analytics that combines descriptive and predictive analytics [75]. Though predictive analytics recommends what will happen in the future, prescriptive analytics provides the best course of action to be taken by healthcare providers in the future [73, 74]. By incorporating clinical and genomic data, prescriptive analytics continuously repredicts the healthcare services and improves the predictive accuracy in order to provide more appropriate diagnoses and treatments for healthcare providers [76, 77].

The medical industry is flooded with enormous volumes of data that require validation and analysis. BDA has a power and capability to perform essential computing and analytical ability to process large volumes of healthcare data. It facilitates medical professionals, clinical researchers, and healthcare stakeholders to improve their results through the use of their internal and external sources of big data [78, 79]. As per the healthcare providers, the assessment of patient data, which incorporates patient medical history (EHR), doctors’ prescriptions, diagnostic reports, biometric data, clinical tests, and other medical data related to health, assists them to follow the advancement of a recommended course of treatment and interrupt the course so that changes can be made if necessary. Thus, it helps to eliminate unnecessary visits and reduce readmission rates. Furthermore, the drug company and other medical organizations take benefit of analytical advantages in designing marketing strategies. Indeed, pharmaceutical industries can study their current market status by capturing and analysing the healthcare data such as sales record and interpretation of drug information prescribed by healthcare professionals for each patient and disease to develop the strategic goals. Therefore, the health insurance company can develop an appropriate health plan for every patient by analysing their demographic data, clinical trials, and statistical data related to health factors [69].

An enormous amount of data are accumulated in the healthcare sector from patients’ medical histories, clinical trials, and diagnostic reports. Like healthcare big data, data analytics can be characterized by volume, velocity, and variety known as 3Vs [3, 17]. BDA is the use of advanced analytical techniques to analyse, extract, and discover meaningful patterns and insight from large data sets [80, 81]. BDA plays a crucial role in enhancing healthcare facilities and increases patients’ clinical outcomes. It therefore has the ability to improve the quality of care and life styles and reduce medical costs. Based on the systematic review on the current state of big data research by Wang and Hajli, BDA in the context of healthcare can be characterized as the capability to acquire, store, process, and analyse large amounts of health data in different forms and provide meaningful information to users, which enables them to explore business values and insights in a timely manner [82].

6.2. Necessity of Healthcare Big Data Analytics. BDA in healthcare is needed to enhance the healthcare quality by taking the associated healthcare services into account:

**Provision of Personalized Healthcare.** Big data in healthcare can revolutionize the medical field through early-stage disease detection and reduce medical cost for the patients using appropriate analytical tools in a

<table>
<thead>
<tr>
<th>BDA techniques</th>
<th>Healthcare application areas</th>
<th>Examples</th>
<th>Sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>Machine learning</td>
<td>Early detection of diseases</td>
<td>Predicting epidemics, disease monitoring</td>
<td>[12, 55]</td>
</tr>
<tr>
<td>Data mining</td>
<td>Prediction of heart disease at early stages</td>
<td>Health analytics, determination of epidemics</td>
<td>[56–59]</td>
</tr>
<tr>
<td>Neural network</td>
<td>Diagnosis of chronic diseases</td>
<td>Predicting of patients’ future disease, patient safety</td>
<td>[60–62]</td>
</tr>
<tr>
<td>Pattern recognition</td>
<td>Improvement of public health disease surveillance</td>
<td>Empowering public health, health literacy, improving quality of care</td>
<td>[63, 64]</td>
</tr>
<tr>
<td>NLP</td>
<td>Improve the quality of care and accuracy of clinical decisions</td>
<td>Cost reduction, high-risk factor identification</td>
<td>[57, 65]</td>
</tr>
</tbody>
</table>
comprehensive manner. This helps to develop a personalized healthcare system for healthcare stakeholders [83, 84].

**Early Detection of Spread of Diseases.** This concentrates on early prediction of viral (infectious) diseases (i.e., before spreading) on the basis of social network analysis. More and more social media of the patients suffering from a disease in a specific geographical area are monitored to identify the development and spread of viral disease. This assists the healthcare experts to counsel the sufferers to take the necessary preventive action.

**Monitoring the Clinical Performance.** There is a lot of enthusiasm to evaluate clinical performance in order to screen and enhance the quality of healthcare services. The reform of the hospital is of major concern in the strategic plan of the healthcare sectors. This can be achieved by monitoring and setting up the hospital in accordance with medical council’s standards.

### 6.3. Big Data Analytical Techniques in Healthcare

In the past, traditional technologies and data warehouses were used by the data analyst to store, process, and manage data. However, the revolution of massive volume of data in healthcare cannot be handled using conventional database systems, tools, and techniques. Nowadays, many advanced technologies with high computing power and storage capacity have been developed in order to address the low performance and difficulty of traditional systems. Accordingly, in [85], “big data technologies can be referred to as advanced technologies that have a high computing power and analytical ability to process large volumes of data collected from various sources to extract insight from it.” As per the authors in [86, 87], big data techniques cover a wide range of fields such as machine learning, statistical analysis, and image analysis. A few of the well-known BDA techniques used in the areas of healthcare are shown in Table 5. The categories that are generated in Table 5 are taken from the literature [12, 66–68, 86]. Big data plays a significant role across all domains such as government organizations, trade associations, healthcare industries, education, and research and development. BDA also empowers the secondary use of clinical data in the healthcare sector [88]. Big data acceptance has shown enormous growth from 17 percent in 2015 to 53 percent in 2017 according to Forbes [89].

In the current digital era, healthcare is one of the sectors that generates a large volume of healthcare data, and these healthcare big data can be characterized by its volume, velocity, and variety known as 3Vs [3]. Data mining techniques can be applied on this massive amount of healthcare data so as to identify new interesting patterns and valuable insights for quality medical services. The Hadoop is an open source software framework for BDA in healthcare as well as the most popular implementation of the MapReduce programming model [90]. It allows distributed storage and processing of large variety of healthcare big data whether it is structured, semistructured, or unstructured such as patient’s EHR, physician’s notes, laboratory data, clinical trials reports, and insurance data as compared to conventional database systems. Figure 5 shows a general conceptual architecture of big data analytics [7].

### 6.4. Platform and Tools for Healthcare Big Data Analytics

There are currently several techniques available for performing BDA. The few tools and techniques that support the Hadoop distributed platform are being discussed below [91, 92]:

**Hadoop Common.** It refers to the set of common utilities that assist other modules of the Hadoop framework. Hadoop Common is a fundamental part of the Apache platform in addition to the HDFS, YARN, and MapReduce. Hadoop Common is generally called as Hadoop Core.

**Apache HDFS.** HDFS refers to the Hadoop Distributed File System that can be used to process unstructured data on commodity hardware predominantly. HDFS is the primary data storage where each file is divided into blocks of fixed size and distributed across numerous servers (nodes). HDFS employs the master/slave architecture using NameNode (master node) and DataNode (slave node) [93, 94].

**Hadoop MapReduce.** MapReduce is a programming framework that enables us to process massive amount of data in parallel in a distributed computing environment. This framework consists of two main functions, namely, Map and Reduce that can effectively manage structured as well as unstructured data [95, 96]. As the name MapReduce indicates, reducer function occurs after the completion of the mapper function.

**Apache Hive.** Hive is a data warehouse framework designed to query and analyse huge amount of data stored in Hadoop HDFS. It is an ETL (extract, transform, and load) tool for the Hadoop ecosystem. Hive is built on top of the Hadoop platform and provides a declarative language similar to SQL known as the Hive query language (HiveQL) that enables SQL programmers to perform data analysis conveniently [97].

**Pig.** Apache Pig is a parallel computing framework that runs on the Apache Hadoop platform. Pig Latin is the language for this platform which is used for analysing large volumes of data due to its distributed architecture. In fact, Pig Latin is like the SQL language and is easy to learn. The main distinction is that Pig Latin can process semistructured and unstructured data [93, 98].

**HBase.** Apache HBase is an open source, multidimensional distributed database system in a Hadoop ecosystem. It runs on the top of the Hadoop Distributed File System (HDFS). HBase can store large volumes of data usually measured in terabytes (TB) to petabytes (PB) and does not support a structured query language like SQL; indeed, HBase employs a NoSQL approach.

**Mahout.** Apache Mahout is an open source distributed framework that supports BDA on the Hadoop platform and is designed for machine learning using the MapReduce program. The Apache Mahout enables us to
develop collaborative filtering, classification, clustering, association mining, and statistical algorithms related to machine learning with the help of data science techniques [93, 99].

For more about the tools and techniques, one may refer to [7, 53].

7. Big Data Benefits in Healthcare Sector

Healthcare sectors extending from a single physician’s office to a large set of networks of healthcare service providers have a potential to acknowledge significant benefits by digitizing, integrating, and effectively using big data analytical tools and techniques in healthcare.

Based on the recently published studies [65, 66, 100], following are some of the major benefits:

- **Clinical operations**: the information on healthcare helps to determine methods of diagnosing and treating patients that are more clinically important and cost-effective.
- **Patients**: healthcare information can help patients to make the right decision at the right time and improve patients’ health while reducing the healthcare cost.
- **Healthcare providers**: the data acquired from medical organizations assist the stakeholders to develop new healthcare strategies for patients to minimize the unnecessary hospitalizations.
- **Research and development**: healthcare data support researchers and scientists to enhance healthcare services through more precise and appropriate treatments.
- **Public health**: healthcare data also assist to assess the health risks as well as analyses trends of diseases to enhance public health surveillance.

8. Application of Big Data Analytics in Healthcare

The buzzword big data in the digital world is highly in demand in every sector especially in the field of healthcare. This has laid a foundation for various applications in the healthcare sector. Healthcare BDA has a potential to improve the quality of care and reduce the medical cost of patients by discovering the associations from massive volume of healthcare data, thereby offering a wider perspective of clinical expertise based on medical evidences and various tests [101]. Healthcare BDA also helps the clinicians and policy makers to develop public policy and service delivery based on open health prescribed data, disease prevalence data, and economic deprivation data [102]. As per the authors in [100, 101, 103, 104], the major areas for the applications of BDA in healthcare are as follows:

- **Healthcare Monitoring**: Healthcare data analytics can be used to continuously monitor the health status of the users (patients) in order to enhance their lifestyle [93].

- **Healthcare Risk Prediction**: A deep analysis of healthcare data helps healthcare stakeholders and medical practitioners to develop solutions for risk prediction. It also enables clinicians to be capable of making patient-related decisions on the basis of system predictions [73, 74]. Data analytics in healthcare can also be used to identify and manage high-risk and high-cost patients [105].

- **Behavioural Monitoring**: Another prospective implementation of BDA in healthcare is monitoring of patients with abnormal behaviour [66]. In 2005, Nambu et al. proposed the home healthcare system to capture the behavioural data of patients for diagnosing their health conditions [106].
Fraud Detection and Prevention. One of the major and important application of data analytics in the healthcare sector is fraud detection and prevention. As per the authors in [107], data mining and machine learning techniques are mainly used for fraud detection in healthcare.

Clinical Decision Support Systems. In the medical field, clinical decision support systems are designed to facilitate healthcare professionals in making clinical decisions to diagnose diseases based on patient’s health condition [108, 109].

Personalized Healthcare Recommendation System. Big data plays a significant role in the healthcare domain to develop a personalized recommendation system to give precise and relevant medical recommendation (advice) to an individual (patient) based on their current health status and medical history [110]. The authors in [111] proposed an intelligence-based health recommendation system using BDA to study and research health records of patients, assess risk and the severity of different diseases, and then provide recommendations based on outcomes of prediction. The authors in [112] suggested a clinical recommendation system that is beneficial for patients to access accurate recommendations based on their own health status.

Drug Discovery and Clinical Trials. Healthcare BDA is widely used by the pharmaceutical industry for drug discoveries so that it can help physicians, pharmaceutical developers, and other healthcare professionals for getting the right drug to the right patient at the right time [107, 113, 114].

Image Informatics and Telediagnosis. Imaging informatics is the study of methods for generating, managing, and representing imaging information in various biomedical applications. It is concerned with how medical images are exchanged and analysed throughout complex healthcare systems [115, 116]. The authors of the study [117] introduce a novel telemammography system for early detection of breast cancer with the help of image processing and machine learning techniques. Computer-aided diagnosis plays a significant role in medical imaging [118].

Healthcare Knowledge System. According to [119], a knowledge management system is developed based on healthcare big data in order to support clinical decision-making and disease diagnosis. The healthcare knowledge system is based on a variety of databases such as electronic health record (EHR), medical imaging data, and unstructured clinical notes and genetic data.

Public Health Information. As per [115, 120, 121], BDA in healthcare can also be used to track and monitor public health status for decision-making and policy development.

Based on the studies of different authors, it is revealed that the BDA in healthcare has a potential to improve the quality of healthcare, decreasing the readmission rates and reducing the medical cost of patients by exploring the association and understanding the nature of healthcare data [7, 93, 122]. Furthermore, image processing, signal processing, and genomics are presently the three main areas for the application of data analytics in the healthcare domain [123].

9. Conclusion

This systematic review focuses on the existing literature to study healthcare big data based upon defined keywords and research aspects in the healthcare domain. The proposed research uses an SLR protocol and guidelines to review the systematic study of the past and the cutting-edge articles of the big data in healthcare. The purpose of an SLR protocol is based on the following objectives:

- Analysing different perspectives about the concept of big data in healthcare
- Exploring the origins of healthcare big data
- Identifying tools and techniques for healthcare big data analytics
- Highlighting the potential advantages and applications of big data in healthcare
- Drawing attention to overcome the big data challenges in healthcare

The present study will help the researchers with a useful base for future work to understand the overall context of healthcare big data and its applications. The limitation of the proposed research is that the electronic search process was performed in only two journal databases from 2015 to 2019, and the rest of the databases were skipped while accessing the quality of journal articles which can be addressed in future research.

Data Availability

Data sharing is not applicable to this article as no data sets were generated or analysed during the current study.

Conflicts of Interest

The authors declare that no conflicts of interest exist regarding this publication.

References


