IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 24 October 2023, accepted 8 November 2023, date of publication 9 November 2023, date of current version 17 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3332030

TOPICAL REVIEW

Leveraging Big Data Analytics for Enhanced Clinical Decision-Making in Healthcare

FATIMA HUSSAIN[1], MUHAMMAD NAUMAN[1], ABDULLAH ALGHURIED[2], ADI ALHUDHAIF[3], AND NADEEM AKHTAR[1]

[1]Department of Software Engineering, Faculty of Computing, The Islamia University of Bahawalpur, Bahawalpur, Punjab 63100, Pakistan
[2]Department of Industrial Engineering, Faculty of Engineering, University of Tabuk, Tabuk 47512, Saudi Arabia
[3]Department of Computer Science, College of Computer Engineering and Sciences, Prince Sattam bin Abdulaziz University, Al-Kharj 11942, Saudi Arabia

Corresponding author: Muhammad Nauman (nauman@iub.edu.pk)

This work was supported by Prince Sattam bin Abdulaziz University under Project PSAU/2023/R/1444.

ABSTRACT Recently, the rate of data generation has reached unprecedented levels, leading to a huge amount of data volume. In addition, modern-day computing systems generate data in diverse formats, ranging from unstructured to structured and semi-structured. As technological advancements experience exponential growth, novel trends, and strategies are emerging in the field of Big Data to enhance data quality and derive valuable insights, particularly in industries like healthcare. The primary objective of this study is to investigate the challenges and applications of Big Data in healthcare, with a specific focus on improving clinical decision-making. By analyzing 185 papers published between 2012 and 2023, this review article aims to provide a comprehensive overview of the techniques and methods employed in utilizing Big Data Analytics in the healthcare domain. Furthermore, the article aspires to assist the research community in identifying suitable approaches and methodologies for their healthcare-related studies.

INDEX TERMS Big data analytics, healthcare industry, medical big data, big data management, review study.

I. INTRODUCTION

The proliferation of data has experienced an unprecedented surge across diverse sources, owing to the widespread adoption of state-of-the-art technological innovations [1], [2]. These advancements encompass smartphone devices, social networks, wearable devices, major corporate platforms, virtual networking graphs, and the Internet of Things (IoT) [3]. Consequently, an unparalleled accumulation of data has ensued, profoundly influencing various domains and unveiling novel avenues for research and analytical exploration [4]. The remarkable hike in data volume, accompanied by the enticing prospects and potential inherent in data analysis, along with the corresponding challenges in storage, processing, and analysis, has given rise to the conceptualization of "Big Data". This overarching term encompasses the phenomenon of managing vast and intricate datasets that conventional data management approaches are inadequate to address [1].

Researchers across diverse domains have leveraged Big Data to corroborate their research findings and draw robust conclusions. The prominent applications of Big Data Analytics technology include transport [5], customer care [6], smart grids [7], education [8], [9], aviation route optimization [10] etc. In particular, the healthcare industry has emerged as a prominent adopter of Big Data strategies, driving the development of adaptive innovations. However, the digitization of medical data has undergone significant expansion, resulting in data generation at an exponential rate and in diverse formats, including unstructured, structured, and semi-structured datasets. According to the Institute for Health Technology Transformation (iHT2), Big Data in the healthcare sector has reached the scale of zettabytes, with many approaches anticipated to push it further into the yottabyte dimension [11], [12].

In the past decade, the healthcare industry has witnessed remarkable technological advancements, primarily driven

The associate editor coordinating the review of this manuscript and approving it for publication was Mansoor Ahmed.

by significant progress in digital, disruptive, interactive, and omnipresent medical data technologies and interfaces [13], [14], [15], [16]. These developments have led to the generation of diverse types of data within healthcare applications. As a consequence, the healthcare sector now produces substantial volumes of heterogeneous datasets that offer opportunities for enhancing service delivery, healthcare administration, treatment safety, and treatment quality improvement [17].

In the recent past, funding organizations have begun to spend significant money on creating interventions and exploring the potential for both technical and nontechnical societies. For instance, the National Institute of Health (NIH) established healthcare data initiatives in 2017 that collect patient records such as Electronic Health Records (EHRs), imaging, genetic information, environmental parameters, and socio-behavioral relevant information. Medical scholars have indeed promoted more studies in this area of research to provide novel significance and importance for enhancing quality services [18], [19]. Moreover, healthcare data are comparable in size and type to Big Data from other sectors. The application system of medical Big Data assists in developing solutions to enhance clinical outcomes while also generating revenue with innovative techniques for dealing with contemporary issues for healthcare systems. This seems to be due to medical Big Data potential allowing for the detection of extracted features, which results in meaningful information for targeted therapy and many other clinical decisions [20].

In literature, the focus of research is now shifting towards discovering, and providing novel solutions for continuing issues and developing difficulties in certain disciplines [5], [21]. This review article discussed Big Data applications in the healthcare sector. First, define Big Data and its properties. Second, several important features of the Big Data process and technologies are then explained. Following that, appropriate medical Big Data applications are discovered. Consequently, Big Data Analytics is covered as a whole, in particular for the healthcare industry. The overall structure of this paper is depicted in Fig. 1. The main contributions of this research work are:

1) We briefly discuss the unique attributes of Big Data in the healthcare industry, shedding light on the distinct challenges and opportunities posed by its characteristics.
2) We summarize the challenges faced while implementing Big Data Analytics in healthcare. This exploration will contribute to understanding the complexities of integrating such advanced techniques into the healthcare domain.
3) We also briefly explain enhance decision-making through Big Data Analytics by examining real-world applications, insights will be gained into how these techniques can drive informed and efficient decisions.
4) We also explain various tools and techniques commonly employed to implement Big Data Analytics in healthcare are crucial. This information will provide insights into the technological landscape of the field and aid in comprehending its advancements.

This review article is organized as follows: Section II describes the article selection process. Section III defines Big Data Analytics and the use of Big Data Analytics in industries, particularly in the healthcare domain, and presents a literature review. Subsequently, Section IV presents Big Data Processes and Section V discusses Big Data Technologies used for analysis. The role of Big Data in the healthcare sector is demonstrated in Section VI while Section VII presents its challenges. Section VIII describes Big Data attributes and their sources. We present discussions in Section IX and finally, Section X concludes this review article.

II. ARTICLE SELECTION

For this review, extensive databases and publishers such as ACM Digital Libraries, IEEE Explore Digital Library, Science Direct, Google Scholar, and Springer were thoroughly investigated. The search utilized keywords related to Big Data Analytics, Healthcare, and Medical Big Data. A total of 19,156 research data entries were initially identified. Subsequently, based on the latest literature review, comprising Journal Papers, Conference Papers, and Review Papers, a refined selection process was carried out. The search was limited to Big Data publications in healthcare spanning the past 11 years, from 2012 to 2023. As a result, 150 articles were retrieved, and 29 articles were carefully chosen for detailed examination, as depicted in Table 1 and Fig. 2-Fig. 3. Each selected article underwent meticulous analysis to uncover gaps and motivations for conducting this research.

III. BIG DATA ANALYTICS

Big Data Analytics (BDA) is the convergence of two key elements, namely Big Data and Analytics [1], [2]. It refers to the systematic process of extracting valuable insights and meaningful information from large and complex datasets. By employing various methodologies and techniques, BDA enables businesses to make informed decisions and provides support for effective decision-making processes [22]. The BDA utilizes data from diverse sources, such as social media, sensors, and transferable data from Customer Relationship Management (CRM) and Enterprise Resources Planning (ERP) systems. The primary objective of BDA is to reveal concealed patterns, trends, and insights embedded within vast and intricate datasets. By harnessing this data, organizations gain a deeper understanding of their market and make well-informed decisions to enhance their business operations [23], [24], [25], [26], [27]. The process involves the use of various techniques and technologies, including data visualization, Machine Learning, and Artificial Intelligence. The BDA also requires a robust and scalable infrastructure, such as a distributed computing platform, to handle the large volume, velocity, and variety of data. In general, Big Data Analytics is the process of insights and knowledge from large and complex datasets by using advanced techniques and

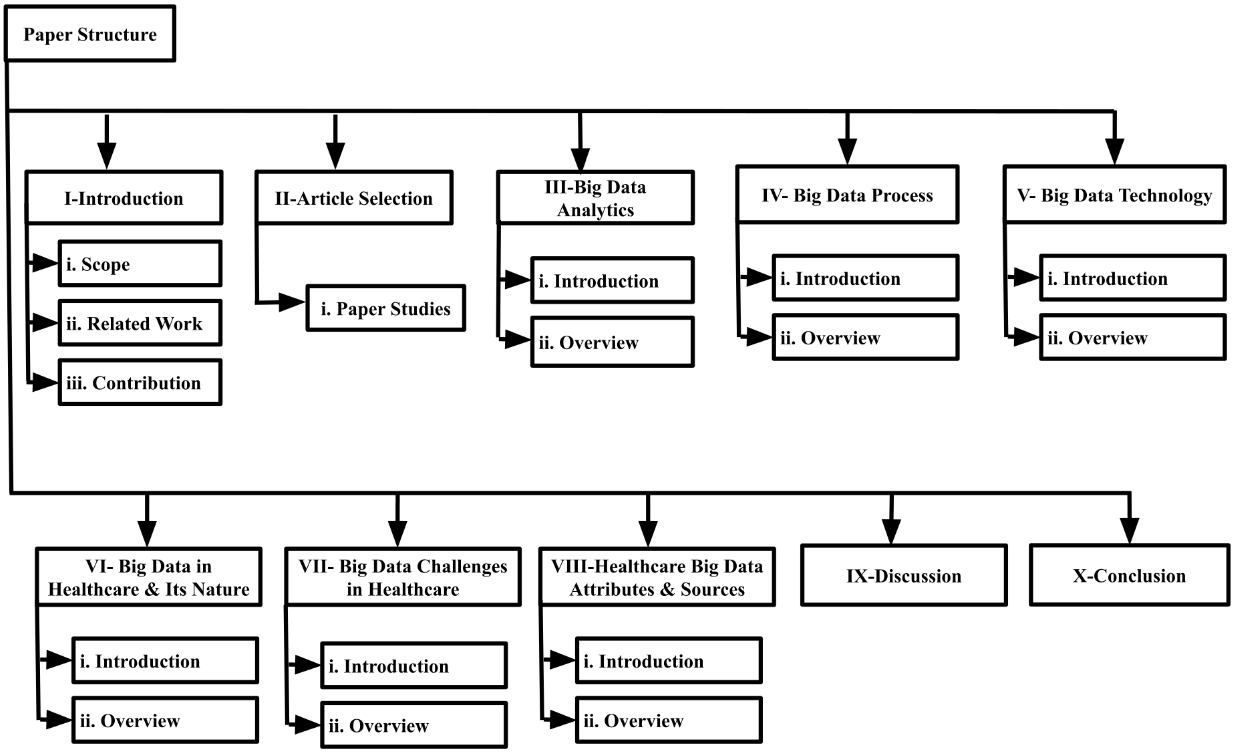

FIGURE 1. Topic covered in this review paper.

technologies, with the goal of informing business decisions and supporting decision-making.

Big Healthcare Data Analytics (BHDA) is a technique to analyze datasets in healthcare to uncover trends, patterns, and insights that can inform healthcare management, medical research, and clinical decision-making. It also describes statistical, contextual, predictive, cognitive, and quantitative models for effective and timely decision-making. Medical Practitioners, Healthcare Stakeholders, Pharmaceutical and Clinical Researchers, Hospital Operators, and Health Insurance may enhance their findings [28], [29], [30]. In medical BDA, data is generated from a variety of sources for example medical imaging, genomics, clinical trials, and Electronic Health Records (EHRs). The purpose of medical BDA is to increase patient care and findings, reduce healthcare costs, and accelerate medical research. By making a plan for an individual patient, the data that need to be analyzed are age, gender, clinical findings, and medical history [31], which outcomes as a result save lines and cost savings.

The BDA in the medical sector refers to the method for analyzing large datasets about the health and well-being of patients. Healthcare data can take several different forms, for example, literature from medical journals, blogs regarding healthcare, social medical, human body monitoring sensors, financial data, machine equipment, and laboratories which may be provided within medical services (like HER and LIMS) [1], [32] or may originate from outside sources (such as Pharmacies, Government, and Insurance Providers), and they may be in an organized or structured manner [1], [33].

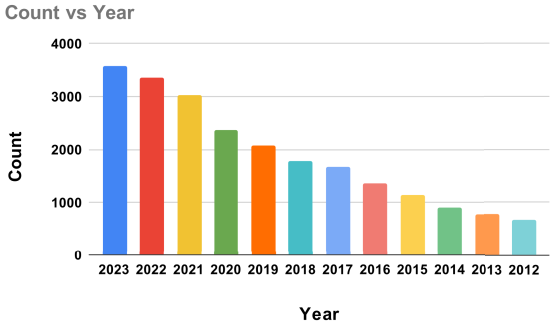

FIGURE 2. Summary of per year publication.

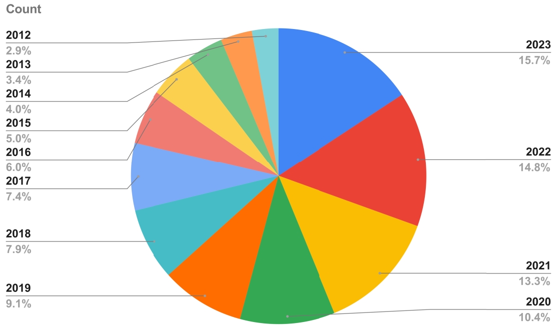

FIGURE 3. Percentage wise paper selection.

TABLE 1. Year wise research article.

Year	IEEE	Springer	Science Direct	ACM
2023	38	721	1966	848
2022	166	140	2324	731
2021	237	182	1966	635
2020	233	178	1484	467
2019	231	164	1196	487
2018	195	143	1041	412
2017	235	121	899	426
2016	151	77	794	335
2015	125	62	640	319
2014	76	44	462	332
2013	44	13	416	309
2012	9	7	360	285

To demonstrate the complexity of data size, consider in 2012 health data explosion, which began with 500 petabytes [23] and will reach 165 zettabytes in 2025 [23]. The BDA approaches relate to methods such as optimization, prediction, simulation, and others that help managers and policymakers make decisions and gain insight. Computer professionals are continually creating new applications to assist healthcare stakeholders in increasing their prospects for better value. For good decision-making on Big Data, many sectors want to build their own infrastructure to assist managers in making better decisions [33].

In BDA, different methods examine the patient record for improvement in a result which is conducted from large datasets to improve the healthcare industry [14], [15]. Firstly, in BDA the main problem is managing and storing unstructured data in a structured format [16]. The role of BDA in the healthcare industry will assist every physician with the medical histories of patients, allowing for optimal decision-making about the treatment of specific patients [14]. The human body's main part is the heart, cardiac attacks are one of the many heart problems that can occur [34], [35]. The BDA is the process of using advanced analytical and computational technologies to extract insights from huge and diverse datasets. The BDA is being used to enhance patient care outcomes, as well as to support research and new innovation in treatments and therapies in the healthcare sector [13].

The healthcare sector generates and collects data from different resources such as Electronic Health Records (EHRs), claims and billing systems, clinical trials, genomics, and wearable devices. The data is leveraged to uncover patterns and insights that can increase the efficiency and quality of healthcare delivery. By analyzing a large population of patients data in healthcare organizations can identify risk factors and patterns that can help to predict and prevent chronic conditions such as diabetes, heart disease, and cancer. This can help to enhance the health of the population and decrease healthcare costs [36].

By analyzing data from genomics and clinical and Electronic Medical Records (EMRs), healthcare organizations can identify genetic markers and other factors that can help to predict a patient's response to different treatments. This can help to improve treatment outcomes and reduce the risk of side effects.

For example, by analyzing medical Big Data from EHRs and claims systems, healthcare organizations can identify inefficiencies in their operations and improve the coordination of care. Additionally, it may also be used to identify patients who are at high risk of readmission, which can assist in cutting costs and enhance treatment quality [4], [37]. As a result, BDA is transforming the healthcare sector by providing new opportunities to improve patient care and outcomes.

IV. BIG DATA PROCESSES

Due to different types of data being considered to leverage value, a four-step process must be adopted. In this subsection, we try to explain how can we analyze and process low-density data for decision-makers in their decisions and projects. The following four steps of Big Data must be followed in Fig. 4. The large data chain value refers to this procedure [21], [38].

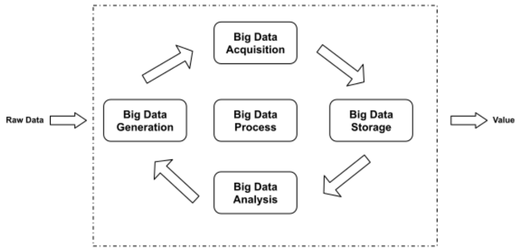

FIGURE 4. Chain value of big data process.

A. BIG DATA GENERATION

Data generating is intended to generate enormous amounts of data for analysis. Data is produced from numerous platforms, including inner data from enterprise information technologies, IoT, Bio-Medical, and the Internet. An organization's internal data includes supply chain data which includes quality, manufacturing, sale, and administrative, as well as inventory data such as human resource data. Internet data contains information from comments and likes, data from click streams, messages, and file logs. Genes, clinical information, and medication information are all included in bio-medical data [39], [40].

B. BIG DATA ACQUISITION

Big Data acquisition is explained in three steps step 1: data collecting, step 2: transmitting data, and step 3: preprocessing data.

1) BIG DATA COLLECTING

Raw data is produced by different sources, that is unstructured data, structured data, and semi-structured data, by utilizing computational methods and technology. According to [39] and [41], most researchers describe Big Data sources into four categories: Open data, IoT, Information systems, and Smartphones. In particular, Tablets, PCs, and Smartphones create large amounts of smartphones from the installed application. Information systems are taken into as a centralized data center carrying all data on an organization's strategy. Embedded sensors refer to various devices in the network that may deliver streamed and manage the data updating through the internet network. Open Data is the tremendous quantity of data that can be retrieved from sources including journal papers, research articles, web pages, and forums [39].

2) TRANSMITTING OF BIG DATA

Data processed and analyzed from data sources are referred to as Big Data transmission [39].

3) BIG DATA PRE-PROCESSING

Pre-processing of data steps assures data analysis and storage in an efficient and improved way. Furthermore, data collection for pre-processing is to remove noisy, unrelated data, inadequate, and, unnecessary data, resulting in reduced storage needs and improved analytic accuracy. Also, low-density data must be combined with additional data to get additional value [39].

4) STORAGE OF BIG DATA

The databases managed enormous amounts of data of many forms and types for further processing and analysis while also ensuring data availability, data security, and data dependability. Earlier, datasets were somewhat confined; as a result, the Variety, Volume, and Velocity, of the data were significantly small, justifying the adaptation of an RDMS. Nowadays, the adoption of the World Wide Web is essential to have easily accessible and efficient data centers for the processing of data. Furthermore, storage of data technology is becoming more important and is a significant investment by many corporations [39].

5) BIG DATA ANALYSIS

Big Data Analysis steps are most significant and crucial in Big Data Chain Value, which is produced as an output. For mining and extracting useful and hidden information for huge volumes for processing and storing data with the help of techniques and tools [39], [42].

V. BIG DATA TECHNOLOGIES

Analyzing and organizing structured data on a modest scale, analysts have used data center-based rational databases. In accordance with widely recognized Big Data characteristics, traditional technologies are ineffective and incapable of handling massive amounts of data and extracting significant insights from them. Many platforms and technologies on novel distributed architectures having substantial memory storage and processing capabilities have been created to address poor performance and complexity experienced whiles the use of traditional technologies. Technologies in Big Data include corporate as well as open-source services and software for informative storage, querying, analyzing, management, processing, and access [43], [44].

A. HADOOP ECOSYSTEM AND BIG DATA

1) ABILITIES OF APACHE HADOOP

Apache Hadoop [18], [45] refers prominent Big Data platform with a significant supportive association. Existing technologies found complexity and low performance while analyzing and computing Big Data. Many prominent Information Technology organizations, including Facebook, Twitter, LinkedIn, IBM, Adobe, Amazon, and, Yahoo are now using Hadoop for Big Data Analytics [46], [47], [48].

Hadoop can take a few seconds to search Terabytes (TB) of data while others take time to search data. Hadoop also executes programs while retaining fault tolerance, which is common in distributed situations. To ensure this, it maintains data on servers to prevent data loss. The Hadoop platform's strength is based on two major fundamental percepts: in the HDFS and MapReduce framework, the Hadoop platform can establish elements as needed in accordance with consumers' requirements and objectives. Indeed, the Hadoop community has contributed various open-source modules to its ecosystem [46], [49].

Early disease detection using a Hadoop-based system data from spectrography, Magnetic Resonance Spectroscopy (MRS), Magnetic Resource Imaging (MRI), and findings from Neurophysical tests are integrated into such some online system [50]. Some online systems [51] and healthcare information systems [31] use Hadoop approaches and IoT to offer recommendations regarding cardiac disorders. Fig. 5 demonstrates the Hadoop Ecosystem.

2) BIG DATA STORAGE LAYER

HBase is used as a storage medium where massive data is stored on the Hadoop ecosystem which uses HDFS and non-relational databases.

a: HDFS (HADOOP DISTRIBUTED FILE SYSTEM)

Information is stored by HDFS which can accommodate a large number of clusters and delivers worthwhile and reliable storage [52]. In HDFS, unstructured data and structured data are stored in huge volumes (file size in Terabytes). Nonetheless, users need to be aware that HDFS is not a generic file system. Furthermore, HDFS assistances in improving system work and decreasing network congestion. It also supports data duplication in fault tolerance [46], [53], [54].

The master-slave architecture underlies HDFS. It distributes massive amounts of data around the cluster. The cluster, in essence, masters the supervises all operations in the file system, slave nodes organize and achieve data storage on single nodes. Master information systems can be broken down and stored data in a secondary name node. Hadoop relies on data replication to offer data availability [55], [56].

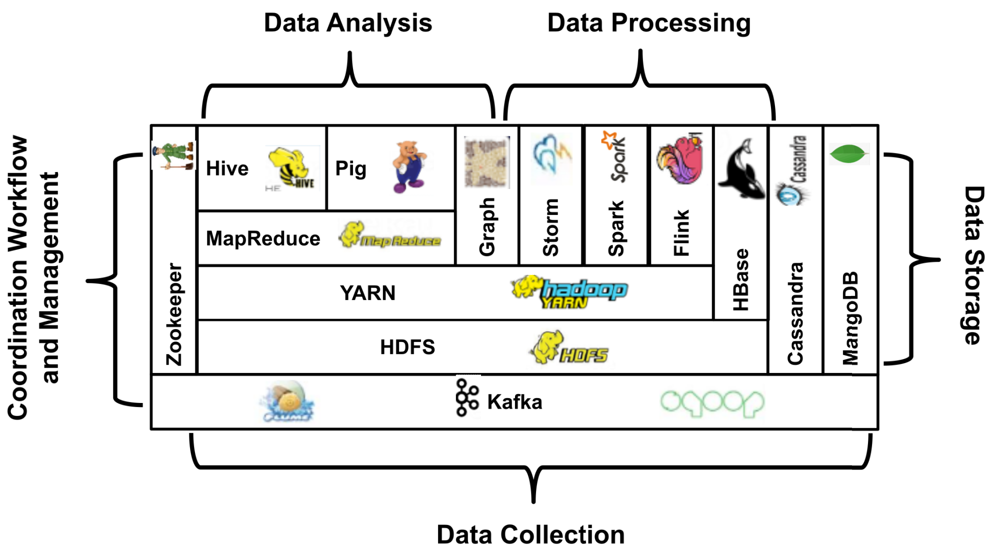

FIGURE 5. Hadoop ecosystem [21].

b: HBASE

HBase is a distributed, column-oriented NoSQL database that runs on Hadoop and allows for real-time random access to large datasets. It is designed to handle a high volume, high velocity, and high-variety data of structured and semi-structured data, including EMR healthcare data [57]. Low latency operations are performed on HBase. It is designed as a value/key data architecture that is column-oriented. It has the ability to scale out and update database tables of parallel support in distributed clusters. It is adaptable in terms of updating table rates and scaling out parallel support in distributed clusters. In the Bigtable like structure, HBase provides flexible organized hosting for every table. The table logically holds data in rows and columns. Such tables have the benefit of handling millions of columns and billions of rows [46], [58], [59].

HBase offered several capabilities e.g. customizable table sharing, natural language search, real-time queries, linear and modular scaling, and consistent access to the sources of Big Data [60]. It provides solutions related to Big Data and data-driven websites i.e. Facebook and Messaging [41], [46], [61].

3) DATA PROCESSING LAYER

a: MAPREDUCE

In 2004, MapReduce was another Apache Hadoop cornerstone that was created when Google introduced a new term in study [62]. It is a Java-based framework that is used for parallel processing to handle large volumes of distributed data on a cluster [63]. MapReduce is made by two terms one is Mapper and the second is Reducer. In Mapping data is equally split and then assigned key/value to the data. Then Reducer phase gets data from a mapper and gives the required output of the data. Iterative processing is not intended for MapReduce [46]. Hadoop does not support the processing of real-time streaming or in-memory computing, nor it is always simple to apply the MapReduce paradigm to all issues. In contrast, stream computing stresses data velocity and involves continuous data input and output. Real-time computing, low latency processing, and high throughput are all features of Big Data Streaming Computing (BDSC). The necessity of Big Data Analytics in healthcare is the ability to bring out information from large amounts of data, which makes BDSC a promising option [64], [65]. The use of BDSC in healthcare using MapReduce is very important to get hidden values in real-time in healthcare [66], [67], [68].

b: YARN

Yet Another Resource Negotiator (YARN) provides a more general-purpose alternative to MapReduce [54]. It is unconventional resource management runs and allows many applications that run in parallel using HDFS. Furthermore, it supports both batch and stream processing. Scalability and security are also features of this system. Besides, YARN employs dynamic allocation of system resources, allowing it to expand its exploitation resources. YARN, like the MapReduce framework, has a master-salve design [46], [69].

c: CASCADING: A MAPREDUCE PLATFORM FOR LARGE SCALE FLOWS

Cascading framework [69] is a data flow component like Pig. Cascading's only goal is to make it possible for development to make business Big Data applications instead of having to understand its intricate working behind Hadoop or write application code accessing API of Distributed Processing Engine such as MapReduce. Cascading used high-level logical structures to design, develop, and deploy to construction as Java and Scala classes. Along with MapReduce, Cascading support Flink and Tex as the Distributed Processing Components. Cascading includes supporting Spark and Storm as future ideas [68]. It offers many intriguing advantages that enable advanced query management and complicated workflow management on Hadoop clusters. Portability, scalability, test-driven development, and integration are all supported.

4) DATA QUERYING LAYER

a: APACHE PIG

Another open-source framework called Apache Pig is used to produce high scripting languages [69]. Apache Pig has a feature to reduce the complexity of the MapReduce algorithm. By enabling MapReduce jobs and processes to execute in parallel over Hadoop. Owing to its user platform environment, it makes it easier to browse and analyze large datasets in parallel using HDFS, much like Hive. A pig can also connect with programs written in other programming languages, binaries, and shell scripts. Map Data is Pig's unique framework [70].

b: JAQL

In addition to Hadoop, JAQL is a domain-specific language designed to offer querying language that facilitates Big Data processing [71]. High-level queries are converted into tasks MapReduce in JAQL. It was developed so that it could query semi-structured data using the JavaScript Object Notation standard (JSON). Moreover, a broad range of other information types and data schema, including flat files, XML, and CSV data. Consequently, a data structure is not required for JAQL and Pig. Numerous built-in features, primary operators, and I/O connectors were available with JAQL. These features allow for the processing, storage translation, and conversion of data into JSON format.

c: APACHE HIVE

To make Apache Hadoop use simpler, Apache Hive, a data warehousing solution, was developed [40], [72] unlike MapReduce, which manages data within files using HDFS, Hive enables the representation of data in a structured database that is more familiar to consumers. In actuality, tables are from the bulk of Hive's data model. These tables are partitioned and reflect HDFS directories. Then, buckets are created from each division. Using a descriptive dataset, diabetes is analyzed using 'Hive'. Efficient prediction models are developed to provide data related to diabetes investigation [73], [74].

Furthermore, Hive provides a language similar to SQL called HiveQL that enables users to approach and modify Hadoop-based data stored in DHFS or HBase [75]. Consequently, a variety of corporate applications are acceptable for it. Real-time transactions are not appropriate for Hive [76]. It actually has a low-latency operation as its foundation. Hive, like Hadoop, is built for processing on a large scale, so even straightforward jobs might take a while. Hive transparently converts some sorts of queries into batch-processed MapReduce operations, such as summaries and ad-hoc search joins [77].

5) DATA ACCESS LAYER: DATA INGESTION

a: APACHE FLUME

Apache Flume is a distributed, reliable, and available service for efficiently collecting, aggregating and moving large amounts of log data [78], [79]. It can manage streaming data flows and has a straightforward adaptable design. To manage vast dispersed data sources, Flume is built around a straightforward extensible data model. Flume offers a number of features, including failure recovery service, configurable reliability mechanisms, and fault tolerance. Flume is a standalone component that can operate on other platforms even though it works well with Hadoop.

b: APACHE SQOOP

An open-source application called Apache Sqoop is a Command Line Interface (CLI) for transmuting large volumes of data between structured data repositories and Apache Hadoop (such as NoSQL databases, enterprise data warehouses, and relation databases) [80]. Sqoop provides optimal system utilization, fast performance, and fault tolerance to lessen the processing demands placed on external systems. Using similar high-level language (Pig, Hive, or JAQL) or MapReduce the imported data is transmitted [81].

c: CHUKWA

Chukwa is a Hadoop-based data-collection system [82]. Its objective keep track of massive dispersed systems. Data is gathered from different data providers using HDFS, and it is analyzed using MapReduce. It takes Hadoop's scalability and robustness. For outputs, it provides a UI (user interface) for displaying, following, and analyzing. Chukwa provides a robust and adaptable framework for Big Data. It gives analysis the ability to gather, examine, monitor, and show Big Data collection. The reliability issue in Chukwa moves the retry logic as near as feasible to the data source. Either local disc log files or HDFS is used to store data. By default, no further copies are produced. Data transmission is only considered successful when data from one source to another source is copied. The technological issue comes in two forms. The monitoring system may be invisible to older applications, thus reliability must be connected with them, and during runtime, this evaluation must be carried out effectively and continually [83], [84].

6) DATA STREAMING

a: STORM

An open-source, Storm is a system for distributed real-time processing [56]. Its structure is a DAG with ''spout and bolts,'' where spouts create tuples from input streams and bolts instantly handle those tuples. Storm can be crucial for apps that use streaming analytics. Storm's ISpout interface could be able to handle any incoming data. Users may really consume data from a variety of real-time synchronous and asynchronous systems using Storm (like JMS, Kafka, Shell, and Twitter). Based on bolts, Storm makes it possible to write data to any output system. Storm has the IBolt interface, which supports all types of output systems, including Hive, HBase, HDFS, and other messaging systems.

It will redirect the process to a different computer and restart it there if it crashes repeatedly. It may be applied in a variety of situations, including distributed RPC, continuous computing, online ML, and real-time analytics [46], [69]. In order for other Hadoop tools to asses the findings, Storm must first prepare them. A million tuples may be processed in a single second. Storm offers a condensed programming style, similar to MapReduce, that masks the difficulty of creating distributed software [49].

b: APACHE SPARK

In-memory computing is supported by Spark [76], a distributed computing framework that is extremely fast. Spark supports streaming data as well as HDFS and MapReduce interfaces, RDD (Resilient Distributed Dataset) is the concept on Spark's data model that is built. Such objects are accessible without disk access necessary. In the event that a partition is lost, it may also be reconstructed. The storm cluster and the Hadoop cluster appear to be comparable. However, with Storm, multiple topologies may be used for distinct storm tasks. Instead, the sole alternative under the Hadoop platform is to build MapReduce tasks for the related applications. The following is a key distinction between MapReduce tasks and topologies. The MapReduce task terminates, but the topology continues to process messages indefinitely or until the user terminates [76].

7) STORAGE MANAGEMENT

a: HCATELOG

Apache HCatelog provides storage management and table administration for Hadoop users [85]. It allows different data processing technologies to work together (like MapReduce, Hive, and Pig). Data type techniques and common schema make this possible. For any data format (such as Sequence Files formats, CSV, JSON, RCFile) for which a Hive SerDe (serializer-deserializer) can be developed, it provides an interface to simplify read and write data operations. In order to do that the system administrator makes accommodations for SerDe, Input Format, and Output Format. The HCatalog abstracted table provides a relational representation of the data in DHFS and allows for tabular viewing of multiple data types. Users are not required to be aware of the place or process used for data storage. HCatalog allows users to write and read data on the grid in Hadoop storage management and table with data processing tools [46], [58].

8) DATA ANALYTICS

a: MAHOUT

Apache Mahout is a library of scalable ML algorithms that can be used in Big Data Analytics [86]. Hadoop can be extended with Mahout to use MapReduce to run algorithms. It is made to function on various platforms. In essence, Mahout [87] is a collection of Java libraries. It benefits ML systems and algorithms across Big Datasets by assuring their effective implementation and scalability [46], [58]. It provides a collection of algorithms that can be used to accomplish common data mining and ML tasks such as classifications, clustering, and recommendation systems. Mahout top of Apache Hadoop and Apache Spark. One of Mahout's main features is that it implements popular ML algorithms like k-mean, Canopy, and Singular Value Decomposition that are enhanced for large-scale data processing on a Hadoop cluster. These algorithms can be used to perform tasks such as amounts of data that are typical in BDA. Apache Mahout is a potent tool for Big Data Analytics, with integration with Hadoop and Spark, support for a variety of data formats, evaluation and validation libraries, and development capabilities for data scientists and analysts. It offers scalable machine-learning algorithms that can be applied to large datasets. Google, IBM, and Amazon are among the firms that have used scalable ML algorithms.

9) MANAGEMENT LAYER: PROCESS AND COLLABORATION

a: AVRO

Apache Avro is a data serialization and data exchange format that is often used in Big Data systems, particularly in the management layer of Big Data systems [88]. Avro [86] is an Apache Hadoop open-source framework that provides developers with two services: Data exchange and Data serialization [46], [58]. Avro also provides support for Remote Procedure Calls (RPC) which allows for efficient communication between different systems and programming languages [89].

b: OOZIE

Apache Oozie is used to manage Hadoop tasks and is a server-based workflow scheduler in a Java web application of Big Data systems [90]. It is an effective management system that is scalable, extendable, and capable of handling a high volume of workflow. Directed Acyclical Graphs (DAGs) represented to process workflow of jobs. Various Hadoop job types, including MapReduce, Hive, Pig, Distcp, and Sqoop tasks are supported by Oozie [90]. Oozie server is the main component of Apache Hadoop [91].

c: ZOOKEEPER

Zookeeper is an open-source coordination service created to manage applications and clusters in a Hadoop context [91]. It offers Java and C-cloud software APIs and is applied in Java. Apache Zookeeper is a distributed coordination service that is often used in Big Data systems to help manage the coordination and workflow of distributed applications and services. Zookeeper provides a simple and consistent interface for distributed systems to coordinate and communicate with each other, which is especially beneficial in Big Data systems that often involve many machines working together to process large amounts of data. The main advantage of Zookeeper is the ability to maintain a centralized, highly available configuration repository for distributed systems. Zookeeper also provides a number of primitives for distributed coordination, such as leader election, barriers, locks, and queues. These allow applications to coordinate their work, for example by ensuring that only one machine is running a particular task at a time, or that a group of machines is all waiting for a signal to proceed. These features can be used to coordinate the execution of Big Data processing tasks, besides managing data distribution and computation tasks across a cluster of machines [91].

10) SYSTEM DEPLOYMENT

a: HUE

Apache Hue is a web-based interface for Big Data systems, it provides a platform for data access, data discovery, data collection, and data visualization [92]. Hue is built on top of several popular Big Data technologies including Apache Impala, Apache Hadoop, and Apache Hive and it provides a simple and user-friendly interface to interact with these systems. By offering a user-friendly interface that enables users, including non-technical ones, to explore and analyze massive datasets through features like sharing dashboards, stored queries, and workflows, Apache Hue makes it easier to access and study Big Data in Hadoop clusters. Hue's APIs can be integrated with visualization software, for example, to provide a comprehensive Big Data Analytics solution. Hue is built on top of popular Big Data technologies like Hadoop and Hive, it's easy to integrate into an existing Big Data stack [93].

b: BIGTOP

Apache Bigtop is an open-source project for the packing and distribution of Big Data software stacks [94]. It is focused on providing a consistent, high-quality distribution of popular Big Data technologies such as Apache Hadoop, Apache Hive, Apache Spark, and Apache HBase. Apache Bigtop makes it easier to build and maintain Big Data systems by ensuring the compatibility and proper operation of all components by providing a single distribution of several technologies.

Bigtop is the best option for managing and testing Big Data clusters on-premises or in the cloud, especially for businesses using open-source Big Data software and seeking cost-effective deployment. Bigtop offers tools and integration with configuration management and orchestration tools like Ansible, Chef, and Puppet.

c: APACHE AMBARI AND APACHE WHIRR

Apache Ambari and Apache Whirr are both open-source projects for the development of Big Data systems. Apache Ambari is a management and monitoring platform for Big Data clusters that are primarily used for Hadoop-based systems. It provides an easy-to-use web-based interface for managing and monitoring Hadoop clusters, as well as tools for configuring and managing other Big Data technologies such as Hive, Pig, and HBase.

Apache Whirr, on the other hand, is a library for launching and managing cloud services, including Hadoop clusters using a simple API. Whirr can be launched on a variety of cloud providers, such as Amazon Web Services (AWS), OpenStack, and Microsoft Azure, and it can be used to automate the process of provisioning and configuring Hadoop clusters on these cloud platforms [95]. Many Hadoop components are supported by Ambari e.g. HDFS, MapReduce, HCatelog, Hive, ZooKeeper, HBase, Oozie, Pig, and Sqoop. Ambari is useful for managing and monitoring Big Data clusters, while Whirr is useful for automating the process of launching and configuring Big Data clusters on Cloud platforms. These tools can be integrated with each other for better automation of the Big Data development process. Fig. 6 illustrates the conceptual architecture of Big Data Analytics.

B. CLOUD COMPUTING

Cloud computing refers to developing applications to store, process, and manage data on cloud-renting third-party services [98], [99], [100]. The use of BDA in the healthcare sector is very helpful in predicting the early detection of disease. Using Big Data's potential in many sectors, including, minor and major organizations i.e. healthcare, education, and others, is endeavoring. In healthcare, Big Data analysis is used to decrease treatment expenses, prevent diseases, anticipate pandemic breakouts, etc. [13].

The cloud-based system provides accurate results during experiments on diabetes patients' data collected from body sensors referred to as HaaS (Healthcare as a Service) [101], [102]. As previously mentioned, organizations across a variety of industries are using Big Data in critical decision-making.

Pakistan established the ''National Center in Big Data and Cloud Computing (NCBC),'' which includes 11 universities and 12 laboratories around the nation focusing the agriculture, medical distribution, and energy management [103]. In 2002, Pakistan was the first SAARC country to implement e-government and it is also one of the countries with a large IT contribution. Pakistan is particularly in the early stages of BDA implementation in the healthcare sector [14].

Several advantages of BDA are mentioned including Data gathering from different origins such as social media, databases, e-commerce websites, external third-party sources, and so on. Facilitating services and product delivery to satisfy or beyond client expectations. Responding in

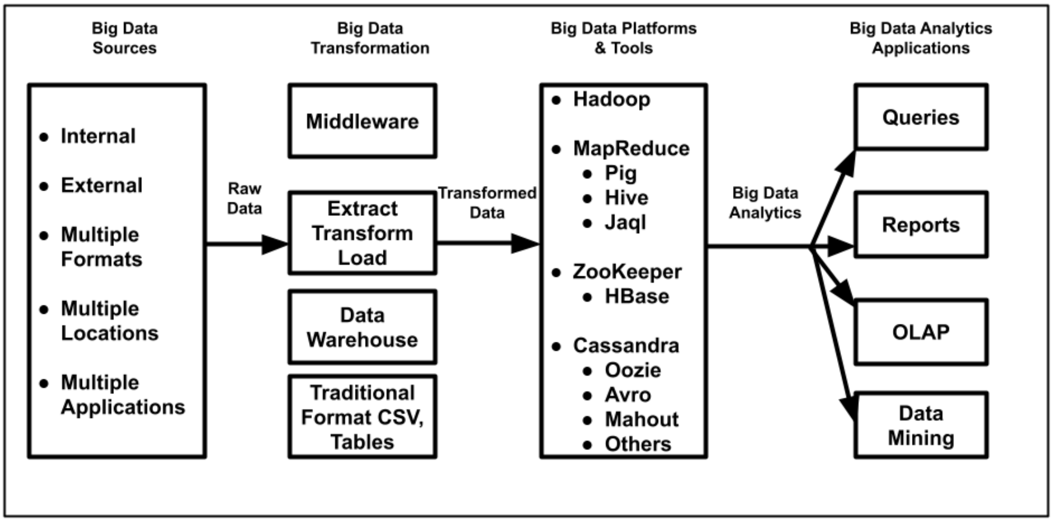

FIGURE 6. Big data analytics: A conceptual architecture [96], [97].

real-time to client requests, concerns, and queries. The sleep monitoring system is designed to monitor the patient's sleeping conditions on a regular basis. The data collected from signals of ECG from ECG sensors are transferred to the cloud for signal processing [104], [105].

Following are additional advantages:

Optimize Cost - Hadoop and Spark both are freely available sources used by corporations to analyze, store, and evaluate huge volumes of data. For example, the transport industry also used data to demonstrate the advantages of cost-cutting. In the transport industry, the cost of the returned product is 1.5 times higher than real transportation expenses. Transport companies can reduce customer return costs by removing the ability to return products in BDA. By using analytics techniques which product return is high to avoid the purchase.

Improve Effectiveness- Big Data can substantially increase operational efficiency. Big Data solutions may collect a great amount of valuable consumer data by engaging them and soliciting their feedback. In order to identify pertinent patterns, such as customer likes and preferences, purchase trends, etc. This data may then be evaluated and analyzed. Consequently, organizations may create customized or modified goods and services.

Novelty - Big Data insight may be leveraged to modify business strategies, generate innovative products and services, increase customer satisfaction, and more.

As recently indicated, corporations have largely benefited from Big Data Analytics, as other industries have also benefited. Particularly, in the healthcare industry, many provinces are already using the potential of Big Data to anticipate and stop epidemics, reduce expenses, cure diseases, etc. This data has also been employed for developing a number of effective therapy models. Big Data created more complete reports, which were subsequently turned into useful essential insights to give better care. The use of Big Data in Education enables teachers to monitor, respond, and measure comprehension of the content in real-time. Professors have developed tailored resources with students to check their knowledge levels in order to grab their interest [13].

Google BigQuery is a cloud computing platform that is cost efficient serverless data warehouse with built-in ML abilities. It's quite versatile and includes a plethora of capabilities for facilitating analytics in various sizes and data types. In order to provide flexible resources, faster innovation, and scale economies, cloud computing is the distribution of computer services, servers, networking, databases, storage, software analytics, and so on across the internet [36]. Cloud computing has changed the way computer infrastructure is utilized and abstracted. The concept of the cloud has been broadened to embrace anything that may be thought of as a service [36], [106].

In healthcare, the cloud computing development and deployment of telemedicine allows healthcare providers to deliver care remotely, using video conferencing and other digital tools. By using cloud-based telemedicine services, without the need for expensive infrastructure. This can improve access to care, particularly in underserved or rural areas, and reduce the need for in-person visits. Cloud computing is also the development and deployment of EHR systems in medical data [107].

EHR systems allow healthcare providers to store and access patient medical records electronically, improving the efficiency and accuracy of care. By using cloud-based EHR systems, healthcare organizations can quickly and easily deploy EHRs, without the need for expensive infrastructure. This can enhance coordination and quality care and decrease the risk of errors and omissions. A third potential use of cloud computing in healthcare is the analysis of large datasets for research [108]. The healthcare sector generates a huge volume of data, including EHR, medical images, and genomics data that can be difficult to analyze and store using traditional data in a centralized location, making it easier to identify trends and patterns that can inform the development of new treatments and therapies [109].

VI. BIG DATA IN HEALTHCARE AND ITS NATURE

Electronic Health Records (EHRs) can enhance advanced analysis and clinical decision-making by storing massive amounts of data. Nevertheless, a significant percentage of all this data has become unstructured. Unstructured data refers to information that does not follow a particular model or organizational system. This selection can easily be due to the fact it is possible to preserve it in a variety of forms.

Another cause for using an unstructured form seems to be that structured data alternatives frequently fail to meet of collecting complicated data. For instance, capture unusual data such as a patient's values and preferences, essential personal characteristics, and some other associated data in a different format than an unstructured manner.

It is essential to combine various disparate, crucial, information sources into a consistent or coherent data format for subsequent analysis utilizing algorithms to better understand and exploit the patient's treatment. However, the healthcare sector must leverage the complete capacity of such rich sources of data to improve patient satisfaction.

In the healthcare industry, clinical data (including personal records, Electronic Medical Records (EMRs), Pharmaceutical records, genetic records, patient personal records, financial records, etc.) are the sources of Big Data [110] and healthcare records (including health management, doctors, and clinical decision support system or patient feedback, medicine, Hospital, disease surveillance) [111], [112]. According to the report, the Compound Annual Growth Rate (CAGR) data for health coverage would exceed 36% by 2025, which is predicted to increase as compared to other sectors [113]. The study examined the long history of healthcare using Big Data. Despite its diversity, the evaluated literature seems to fall specific areas of healthcare into basic categories like Big Data and its significance, analytics and technology development, and research about Big Data. The category of research addresses the complexity, problems, usefulness, and facility of using Big Data to enhance health end results. Now next physical systems development that integrates Big Data Analytics is the next category is data analytics technology framework.

This framework performs heterogeneity from data collection to visualization in the healthcare industry. In the healthcare industry consider lessons learned and general practice suggestions on Big Data Analytics.

A. ELECTRONIC HEALTH RECORDS

In 2017, the National Institutes of Health (NIH) announced a Big Data program to collect patient data including, genetic, imaging, EHR, socio-behavioral, and environmental data. Researchers in the field of health have also advocated for further research in this area with the goal of generating novel significance and value for enhancing healthcare practices.

Electronic Health Records (EHRs) can enhance detailed analysis and clinical decision-making by Storing huge amounts of data. Nevertheless, a significant percentage of all this data has become unstructured. Unstructured information is data that does not correspond to a certain modal or organizing scheme. This selection can easily be due to the fact that it is possible to preserve it in various forms [18], [19].

B. BIOMEDICAL DATA

Research such as life science research firms' data production capabilities, filling gaps in genomes, transcriptomics, proteomics, and metabolomics [114], [115]. Applications about data in agriculture, medicine, and health, when paired with pathological data, it simpler to identify specific persons, compromising patient privacy and causing genetic discrimination. Data generated during medical treatment includes personal information, medical image data, Electric Medical Records (EHRs) data, and drug user data [116].

C. CLINICAL RECORDS

Data generated during medical treatment includes detailed personal information including medical image data, Electronic Medical Record data, and drug user data [117]. The data may be used to produce new worth through analysis but it contains a massive quantity of private data that an unauthorized thirty party could gain, threatening patient privacy [118].

D. PUBLIC RECORD

Having studied disease trends and tracking epidemics to enhance safety for the public, producing more precisely targeted vaccinations, and converting massive volumes of data into usable information to detect needs, avoid crises, and offer service are all examples of what we do [119]. Informatics is an "Applied Information Science" that combines computer science, information technology, management sciences, behavioral science practices, and theories to provide concepts, tools, and techniques for integrating information systems into public health [120]. Healthcare informatics research is a scientific endeavor to improve the operation of the healthcare organization as well as patient care outcomes [121]. Online social media paired with epidemiological data aided monitoring of public health. Forecast infectious illnesses using a social network. Digital media is increasingly utilized to enhance the monitoring and efficacy of Healthcare [122].

VII. BIG DATA CHALLENGES IN HEALTHCARE

In the medical sector, Big Data is generated by clinical data (such as personal records, Electronic Medical Records (EMRs), Pharmaceutical records, genetic records, patient personal records, financial records, etc. [110] and healthcare records (including health management, doctors, and clinical

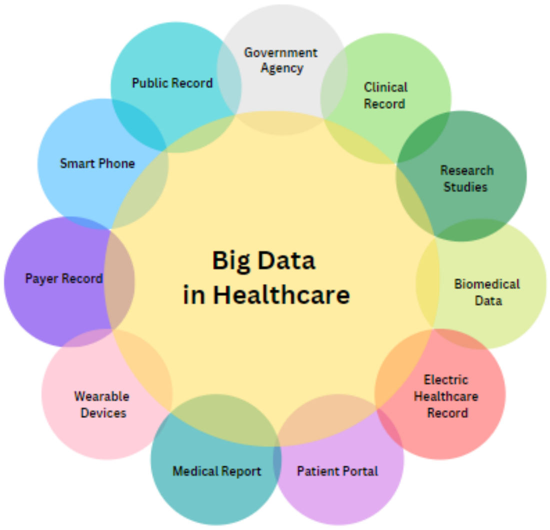

FIGURE 7. Sources of big data in healthcare [123], [124].

decision support system or patient feedback, medicine, hospital, disease surveillance) [111], [112]. According to projections, the Compound Annual Growth Rate (CAGR) data for health coverage would exceed 36% by 2025, which is predicted to increase as compared to other sectors [113]. This study evaluated the background of healthcare and Big Data. Despite its diversity, the evaluated literature seems to fall specific areas of healthcare into basic categories like Big Data and its significance, analytics and technology development, and research about Big Data. The category of research addresses the complexity, problems, usefulness, and facility of using Big Data to enhance health end results.

A. HEALTHCARE BIG DATA CHALLENGES

Currently, the medical sector produces data massively from different sources and different formats, including qualitative sources that are free text and demographics, as well as quantitative sources including gene arrays, laboratory tests, and, sensor data [125]. Medical data is extremely hypersensitive and challenging to approach [126]. Big Data has evolved, introducing with it new challenges and issues brought on by the exponential increase of medical data. The ever-changing structure of data presents several obstacles during the storage, evaluation, and retrieval of massive volumes of data. Due to their large and tremendous volume of data cannot be handled by traditional database systems [127].

Big Data difficulties that commonly occur in healthcare institutions fall into broad categories due to the large quantity of unstructured data, that is traditional language, clinical, and prescriptions which are hand-written data analysis, integration, and storage provide a reasonable level of complexity. The primary goal of a medical decision system is to use social medical data from different websites to predict the onset of illnesses in various geographic locations. To do this, a comprehensive prediction system that takes into account a collection of features connected to a certain epidemic, such as dengue fever and influenza must be created [128], [129]. In healthcare, a significant number of devices generate massive and complicated data from the human body to store for later use and analysis. To resolve these restrictions, various cloud-based technologies for data storage are available, including Nimbit [130], ThingWorx [131], GENI [132], [133], Amazon [134], and, Google Cloud [135].

These platforms improve data storage and management. Practically everywhere and anyplace, data may be obtained and reviewed by healthcare practitioners and researchers for growth and increased knowledge about the healthcare sector. Several hurdles must be addressed before cloud computing can become even more feasible [136]. To begin, cloud computing provides a concise and versatile technique for mining resources. It does, however, increase the potential of privacy exposure. It is a clinically obvious fact in clinical informatics. Subsequently, the importing and exporting of massive amounts (petabytes) of data in medical data to the cloud. Network capacity raises the cost of data while limiting its mobility [137]. Data quality and data integration used for analysis often outcomes from multiple sources and they are not easy to analyze due to data being in different formats. Data cleaning and preprocessing are required to identify and correct errors and inconsistencies and they are not ensured to analyze different types of data. Healthcare data is sensitive and confidential and it is ensured that data is accessed by not every individual besides practitioners and data analysts, this requires robust data governance policies and strict data security controls, such as encryption and secure access controls.

B. OPPORTUNITIES FOR BIG DATA ANALYTICS

By conducting literature work of previous studies medical sector gained potential in Big Data Analytics, some of the following are mentioned below: [138].

1) MEDICAL DIAGNOSIS

A data-driven diagnostics may detect disease at its initial stages and eliminate treatment complications [139].

2) ELECTRONIC HEALTHCARE RECORDS

Big Data Analytics may be utilized to extract useful information from EHRs, which can enhance patient findings and decrease healthcare costs [41], [140].

3) CLINICAL DECISION MAKING

Big Data analytics help healthcare professionals make more informed decisions about patient care. For example, by analyzing data from clinical trials, healthcare practitioners can identify the best treatment option for different patient populations, which can enhance patient findings and decrease healthcare costs [141], [142].

4) POPULATION HEALTH MANAGEMENT

Big Data Analytics identifies trends and patterns in healthcare data that can help to enhance population health [143], [144], [145]. For instance, electronic health records are analyzing and claim data, healthcare providers can identify areas where patients are not receiving appropriate care, and take steps to address those issues.

5) DETECTION OF FRAUD

In the healthcare system, identifying patterns in healthcare data can indicate fraudulent activity. This can help reduce healthcare costs by identifying and preventing fraud before it occurs [96], [97].

6) PERSONALIZED MEDICINE

Big Data Analytics leverages healthcare providers to recognize data that can aid in tailoring treatments to the specific needs of individual patients. This can improve patient outcomes and effective treatments for their specific conditions [146], [147].

7) PREDICTIVE ANALYTICS

Predictive analytics is utilized in the healthcare industry to recognize individuals who are at high risk of developing chronic diseases or readmissions. It might also be used to forecast which patients would have the most complex cases and which are the most at high risk of poor outcomes [148], [149].

8) REMOTE MONITORING

Big Data Analytics are used to monitor patients remotely and identify signs of deterioration in their health. For example, by analyzing data from wearable devices, healthcare practitioners can diagnose those who are at high risk of acquiring chronic diseases and perform preventive or management measures before the conditions become serious. Overall, Big Data Analytics has the possibility to revolutionize healthcare by providing healthcare facilitators with insights that can help them make more informed decisions about patient care, enhance patient findings, and decrease healthcare costs.

9) COMMUNITY HEALTH

Authorities may adopt preventative actions in a community to mitigate the danger of chronic disease [150] and contagious disease epidemic [151].

10) HOSPITAL SURVEILLANCE

Real-time surveillance of hospitals can help government officials preserve the highest level of service quality [152].

11) PATIENT CARE

Big Data Analytics enable personalized patient treatment which can bring immediate relief [153] and minimize hospital readmission rates [154]

C. BIG DATA ANALYTICS IN HEALTHCARE SYSTEMS

The Application of Big Data Analytics in healthcare potentially confronts a variety of problems [155], [156]. In this field, the challenges are usually mentioned.

1) INITIAL INVESTMENT

Implementing the infrastructure required to realize the benefits of Big Data incurs large upfront costs for healthcare sectors [157], [158].

2) DATA QUALITY AND ACCURACY

The scarcity of skilled manpower as well as make a change in reluctance to change in organizational practices [159], [160]. The quality and accuracy of data can be inconsistent, which can make it difficult to draw accurate conclusions from the data.

3) QUALITY OF INSIGHTS

Poorly executed diverse healthcare data may lead to inadequate insight and inaccurate recommendations [161], [162].

4) DATA INTEGRATION

Data is often stored in multiple different systems and formats, which can make it difficult to integrate and analyze.

5) LACK OF STANDARDIZATION

There are many standards to gather, record, and store data which makes it hard to compare data from different sources and make of it.

6) ANALYTICAL COMPLEXITY

Analyzing Big Data advanced skills and sophisticated tools, which can be a barrier for many healthcare organizations.

7) HANDLING OF UNSTRUCTURED DATA

Healthcare data contains many unstructured data such as images and medical notes, which can be difficult to understand and analyze.

8) ACTIONABLE INSIGHTS

Extracting actionable insights from Big Data requires advanced analytical and data mining techniques, which can be time-consuming and resource-intensive.

9) PRIVACY AND SECURITY

Researchers caution against exposing patients to unauthorized data access during inter-system transfers due to their privacy and security concerns [42], [163].

VIII. HEALTHCARE BIG DATA ATTRIBUTES AND SOURCES

In today's digital world, data is important to become a more valuable asset than oil. Data is very valuable and effective as many industries use it for their benefits [3], [164]. Big Data in multiple domains like healthcare, education, agriculture, business, social media, and sports, produced large values of data for testing and analysis purposes. Data generation sources in different and multiple fields, like network sensors, smartphone applications, social media, and financial matters, particularly in healthcare [3]. Each person generated 1.7 MB data in each second [13] on the internet, 1826 PB (petabytes) data are processed in a day reported by the National Security Agency(NSA) [165], [166]. Forecasts by global data experts revealed that by the end of 2022, humans would produce and consume about 94 zettabytes of data. Every two years the data will double according to the International Data Corporation (IDC) [21], [166].

Data production on a daily basis is astonishing. Correspondingly, approaches for analyzing and understanding this massive size of data are necessary, an incredible source of important information. To convert Big Data into useful data information using Big Data Analytics techniques can be performed on critical data from datasets [167]. Researchers describe the concept of "V" using the term Big Data [168], [169].

In 2001 [170], evolving Big Data characteristics were used to identify the Variety, Volume, and Velocity in the three V's, [39], [169], [170], [171], [172], [173]. IDC defines Four more V's (Variety, Velocity, Volume, and Value) using the Big Data's characteristics in 2011 [166].

In 2012, a new characteristics name Veracity was added as another Big Data characteristic [107], [166], [174], [175], [176], 5V as presented as a patient data attribute in another study [177]5V, according to Van and Algar, describe Big Data and motivate its significance to healthcare data [176], 7 Vs [5], [178], [179] added(Variability and Visualization), and 10Vs (Validity, Vulnerability, and Viability) [41], [97], [180]. Few scholars have concentrated on the healthcare sector and they explored the 5V characteristics to show the properties of Big Data [168], [175]. It is presented as a patient data feature in another study. 5V, according to Van and Alagar, described Big Data and drove their applicability to medical data [168], [176]. Although there are other Vs present in literature, in this review only 10 Vs are discussed below:-

A. VOLUME (SIZE)

It describes the massive quantity of data generated every second, reflecting the dataset's volume which is the main challenge in developing general criteria for enormous data size since the timing and kind of data may have an impact on its definition [107], [163] (e.g. definition of "Big dataset"). Exabyte (ED) or Zettabyte (ZB)-sized datasets are currently classified as "Big Data" but issues regarding small size range datasets are still existing [39], [166], [181]. For example, 2.5 Petabyte (PB) is generated from more than a million clients per hour by Walmart [166]. The community discusses leveraging the majority of hospital records, by analyzing datasets it is costly to obtain, and perception about is currently restricted [39], [146], [182].

B. VELOCITY (SPEED)

It is the speed of data processing (defined as a streaming, real-time, emphasizes the data processing in which speed of data is created [39], [166], [183]. IoT devices receive data from sensors regularly. If a delay occurs during data processing using medical monitoring record devices, may cause receiving late results to the physician's harm and death of the patient (e.g. notifying a doctor or institution of emergencies) [184].

C. VARIETY (COMPLEXITY)

It means the different types of datasets, for example, unstructured dataset, semi-structured dataset, and structured dataset [31], [96], [107], [185].

(i) Unstructured datasets (for example, multimedia information, and text) that have not been predefined structured [33], and in the healthcare industry, a different kind of data including computed tomography diagnostic, IoT sensors, laboratory examinations, patients records, and findings. Every day, unstructured healthcare records such as patient information, doctor notes, prescriptions, an image of MRIs, CTs, radio films, clinical or official medical records, and so on are generated. Furthermore, electronic apps, actuarial data, automated database information, electronic billings are accounting, and some clinical and laboratory instrument reading observations are included in the structured and semi-structured variations of EHS and EMS. Data analytics provides several tools for covering unstructured data into structured datasets, particularly NLP in healthcare [186].

(ii) Semi-structured data contains tags to divide data items (such as in NoSQL databases) [107], [187], but it is up to the database user to enforce this structure. Data generated by sensors or other devices to efficiently monitor patient behavior are examples of such data.

(iii) Structured dataset (e.g., a relational database is used for data recording) is very particular and is recorded in the specified format. In the healthcare domain, examples of such data include hierarchical terminology of numerous diseases, their diagnosis information, and symptoms, Results from the laboratory, patient data such as admission records and prescription histories, billing details for therapeutic services provided [14].

D. VERACITY (QUALITY)

The quality of data is represented by its veracity [163]. For example, IBM bears very high expenses for this purpose [188]. It is classified as horrible, outstanding, or ambiguous since data can be noisy, undecided, inconsistent, and imperfect. Data analytics accuracy is getting more challenging to establish as data source diversifies. When evaluating millions of healthcare records from a dataset to discover and predict illness patterns or to minimize an epidemic that may affect several people, any inconsistencies and uncertainties might be delayed and the analytics process' accuracy will be reduced [188]. Veracity in clinical practice typically results have accurate information about their patients, and they are better able to make informed treatment decisions. This can lead to more effective treatments and faster recovery times. In addition, patients who receive accurate information about their health and treatment are more likely to understand their condition and follow their treatment plan, which can also contribute to better outcomes. Ensuring veracity in clinical practice also helps to build trust between healthcare providers and patients, which is essential for a positive patient-provider relationship. Overall, maintaining veracity in clinical practice is crucial for providing high-quality healthcare and achieving optimal patient outcomes. High-quality videos and streaming platforms are some examples of massive data [189].

E. VALUE (EFFECTIVENESS)

The role in data characterization is used for decision making, as opposed to the proceeding Vs, which is primarily concerned with large data challenges [163] Amazon, Facebook, and Google have all utilized analytics to maximize their respective products. Big Data can bring value to healthcare by improving the effectiveness of systems in healthcare. Patient data which is analyzed by healthcare providers can detect obstacles and inadequacies in the system. This can result in improved patient outcomes and reduced costs. To predict the future, data scientists can check trends and make patterns such as the likelihood of a patient developing a certain condition or the likelihood of a patient responding to a particular treatment. This information can be used to intervene with preventive measures, potentially avoiding costly and complex treatments [190].

F. VARIABILITY (CHANGE)

Modeling changeable data sources is common in data science. Models deployed into production may run across, particularly erratic data [3].

G. VISUALIZATION (VISUAL CONTEXT)

Healthcare data must not only be reliable but they must also be accessible clearly and aesthetically to the end users. Complicated hospital reports must be represented in a relevant and less time-consuming manner. Proper visualization aids in the discovery of key ideas by reflecting the information in an expressive and useable manner [3], [191], [192].

H. VALIDITY (ACCURACY AND CORRECTNESS)

In [3], [5], [191], and [192], the validity of data is also determined by its timeliness and if it was obtained using acceptable scientific techniques and methodologies [193].

I. VULNERABILITY (SECURITY CONCERN)

In literatures, [3], [191], and [192] are concerned about data breaches of critical data in healthcare, especially when data is transported from several junctions and electronically stored in the cloud. The patient's sensitive and personal data must be protected in the cloud from unauthorized users. This includes personal identification information, medical records, financial information, and other sensitive data. Data breaches can occur due to a variety of factors such as weak security controls, lack of employee training, and human error. Hackers may also target healthcare organizations specifically, as they often have the health of sensitive data that can be used for financial gain or identity theft.

To prevent data vulnerabilities in healthcare systems, organizations should implement robust cybersecurity measures such as encryption, secure data storage, and regular security audits. They should also provide regular employee training education on data security and have incident response plans in place to quickly detect and respond to breaches.

J. VOLATILITY (CURRENCY AND AVAILABILITY)

References [3], [5], [191], and [192] relates to the duration of data validity and storage. Healthcare data is created and altered at a quick pace. As a consequence, the data life is comparatively short which is be analyzed [193].

IX. DISCUSSION

Big Data has revolutionized the healthcare industry, offering specific approaches and solutions to improve individual health outcomes and enhance the performance of healthcare organizations. In order to continually advance the field's knowledge, sharing additional findings from past research is of utmost importance. The integration of Big Data in healthcare indicates the important progress of utilizing vast datasets within the healthcare sector, made possible by advancements in computer technologies like real-time processing and cloud computing.

In addition, developers, researchers, and industry professionals are increasingly fascinated by the potential value of Big Data in the healthcare industry. Despite various review techniques addressing Big Data and its significance, the findings remain fragmented, lacking comprehensive insights. Expanding the scope of research is essential to identify interconnections, enabling a comprehensive understanding of Big Data research and addressing the research questions previously posed. Consequently, future studies should strive to bridge the gap among diverse Big Data domains in healthcare. This review, to the best of our knowledge, is the most comprehensive and extends the number of relevant publications to be examined.

In this comprehensive review study, we discussed the distinctive features of Big Data in the healthcare sector, highlighting the specific challenges and opportunities that arise from its unique characteristics. We then provide a brief overview of the challenges encountered when integrating Big Data Analytics into healthcare, shedding light on the intricacies of deploying advanced data techniques within this domain. Additionally, we explore the potential for improved decision-making through real-world applications of Big Data Analytics, offering insights into how these methodologies can empower informed and efficient decision-making processes. Furthermore, we elucidate the various tools and techniques commonly utilized for the implementation of Big Data Analytics in healthcare, offering valuable insights into the technological landscape of the field and facilitating a deeper understanding of its evolving advancements.

X. CONCLUSION

The primary aim of this review article has been to explore the significant implications of Big Data techniques and technology on the performance and outcomes of healthcare systems. It introduced the innovative concept of Big Data, its evolutionary trajectory, and its essential characteristics such as Volume, Velocity, Variety, Veracity, Value, Variability, Visualization, Validity, Vulnerability, and Volatility. This comprehensive review serves as a foundational step toward conducting Big Data Analytics research work and expands the understanding of its properties, attributes, and challenges. Big Data Analytics has the potential to elevate the strategic capabilities of healthcare organizations and improve the quality of treatments. By implementing Big Data Analytics in healthcare, organizations can enhance productivity and reduce healthcare costs that may arise from not leveraging the benefits of Big Data Analytics.

REFERENCES

[1] P. Galetsi and K. Katsaliaki, "A review of the literature on big data analytics in healthcare," *J. Oper. Res. Soc.*, vol. 71, no. 10, pp. 1511–1529, 2020.

[2] G. Wang, A. Gunasekaran, E. W. T. Ngai, and T. Papadopoulos, "Big data analytics in logistics and supply chain management: Certain investigations for research and applications," *Int. J. Prod. Econ.*, vol. 176, pp. 98–110, Jun. 2016.

[3] M. Bansal, I. Chana, and S. Clarke, "A survey on IoT big data: Current status, 13V's challenges, and future directions," *ACM Comput. Surv.*, vol. 53, no. 6, pp. 1–59, Dec. 2020, doi: 10.1145/3419634.

[4] L. A. Tawalbeh, R. Mehmood, E. Benkhlifa, and H. Song, "Mobile cloud computing model and big data analysis for healthcare applications," *IEEE Access*, vol. 4, pp. 6171–6180, 2016.

[5] S. J. Miah, E. Camilleri, and H. Q. Vu, "Big data in healthcare research: A survey study," *J. Comput. Inf. Syst.*, vol. 62, no. 3, pp. 480–492, May 2022.

[6] W.-K. Liu and C.-C. Yen, "Optimizing bus passenger complaint service through big data analysis: Systematized analysis for improved public sector management," *Sustainability*, vol. 8, no. 12, p. 1319, Dec. 2016.

[7] H. Daki, A. El Hannani, A. Aqqal, A. Haidine, and A. Dahbi, "Big data management in smart grid: Concepts, requirements and implementation," *J. Big Data*, vol. 4, no. 1, pp. 1–19, Dec. 2017.

[8] R. J. Watson and J. L. Christensen, "Big data and student engagement among vulnerable youth: A review," *Current Opinion Behav. Sci.*, vol. 18, pp. 23–27, Dec. 2017.

[9] M. Nauman, N. Akhtar, A. Alhudhaif, and A. Alothaim, "Guaranteeing correctness of machine learning based decision making at higher educational institutions," *IEEE Access*, vol. 9, pp. 92864–92880, 2021.

[10] E. Kasturi, S. P. Devi, S. V. Kiran, and S. Manivannan, "Airline route profitability analysis and optimization using BIG DATA analyticson aviation data sets under heuristic techniques," *Proc. Comput. Sci.*, vol. 87, pp. 86–92, Jan. 2016.

[11] M. Cottle, W. Hoover, S. Kanwal, M. Kohn, T. Strome, and N. Treister. (2013). Transforming health care through big data strategies for leveraging big data in the health care industry. Institute for Health Technology Transformation. [Online]. Available: http://ihealthtran.com/big-data-in-healthcare

[12] A. A. Phatak, F.-G. Wieland, K. Vempala, F. Volkmar, and D. Memmert, "Artificial intelligence based body sensor network framework—Narrative review: Proposing an end-to-end framework using wearable sensors, real-time location systems and artificial Intelligence/Machine learning algorithms for data collection, data mining and knowledge discovery in sports and healthcare," *Sports Med., Open*, vol. 7, no. 1, pp. 1–15, Dec. 2021.

[13] B. Berisha, E. Mëziu, and I. Shabani, "Big data analytics in cloud computing: An overview," *J. Cloud Comput.*, vol. 11, no. 1, pp. 1–10, Aug. 2022.

[14] M. Shahbaz, C. Gao, L. Zhai, F. Shahzad, and Y. Hu, "Investigating the adoption of big data analytics in healthcare: The moderating role of resistance to change," *J. Big Data*, vol. 6, no. 1, pp. 1–20, Dec. 2019.

[15] S. Connolly, S. Wooledge, and T. Aster, "Harnessing the value of big data analytics," Teradata, Hortonworks, Santa Clara, CA, USA, 2013.

[16] K. Bakshi, "Considerations for big data: Architecture and approach," in *Proc. IEEE Aerosp. Conf.*, Mar. 2012, pp. 1–7.

[17] M. Adibuzzaman, P. DeLaurentis, J. Hill, and B. D. Benneyworth, "Big data in healthcare—The promises, challenges and opportunities from a research perspective: A case study with a model database," in *Proc. AMIA Annu. Symp.* Bethesda, MD, USA: American Medical Informatics Association, 2017, p. 384.

[18] D. PeterAugustine, "Leveraging big data analytics and Hadoop in developing India's healthcare services," *Int. J. Comput. Appl.*, vol. 89, no. 16, pp. 44–50, Mar. 2014.

[19] C. S. Kruse, R. Goswamy, Y. Raval, and S. Marawi, "Challenges and opportunities of big data in health care: A systematic review," *JMIR Med. Informat.*, vol. 4, no. 4, p. e38, Nov. 2016.

[20] R. Pastorino, C. De Vito, G. Migliara, K. Glocker, I. Binenbaum, W. Ricciardi, and S. Boccia, "Benefits and challenges of big data in healthcare: An overview of the European initiatives," *Eur. J. Public Health*, vol. 29, pp. 23–27, Oct. 2019.

[21] S. Bahri, N. Zoghlami, M. Abed, and J. M. R. S. Tavares, "BIG DATA for healthcare: A survey," *IEEE Access*, vol. 7, pp. 7397–7408, 2019.

[22] D. Rajeshwari, "State of the art of big data analytics: A survey," *Int. J. Comput. Appl.*, vol. 120, no. 22, pp. 39–46, Jun. 2015.

[23] P. Galetsi, K. Katsaliaki, and S. Kumar, "Big data analytics in health sector: Theoretical framework, techniques and prospects," *Int. J. Inf. Manage.*, vol. 50, pp. 206–216, Feb. 2020.

[24] P. de Camargo Fiorini, B. M. R. P. Seles, C. J. C. Jabbour, E. B. Mariano, and A. B. L. de Sousa Jabbour, "Management theory and big data literature: From a review to a research agenda," *Int. J. Inf. Manage.*, vol. 43, pp. 112–129, Dec. 2018.

[25] H. Chen, R. H. Chiang, and V. C. Storey, "Business intelligence and analytics: From big data to big impact," *MIS Quart.*, vol. 36, no. 4, pp. 1165–1188, 2012.

[26] S. F. Wamba, A. Gunasekaran, S. Akter, S. J.-F. Ren, R. Dubey, and S. J. Childe, "Big data analytics and firm performance: Effects of dynamic capabilities," *J. Bus. Res.*, vol. 70, pp. 356–365, Jan. 2017.

[27] R. Srinivasan and M. Swink, "An investigation of visibility and flexibility as complements to supply chain analytics: An organizational information processing theory perspective," *Prod. Oper. Manage.*, vol. 27, no. 10, pp. 1849–1867, Oct. 2018.

[28] S. Shafqat, S. Kishwer, R. U. Rasool, J. Qadir, T. Amjad, and H. F. Ahmad, "Big data analytics enhanced healthcare systems: A review," *J. Supercomput.*, vol. 76, no. 3, pp. 1754–1799, 2020.

[29] A. Kankanhalli, J. Hahn, S. Tan, and G. Gao, "Big data and analytics in healthcare: Introduction to the special section," *Inf. Syst. Frontiers*, vol. 18, no. 2, pp. 233–235, Apr. 2016.

[30] V. Rajaraman, "Big data analytics," *Resonance*, vol. 21, pp. 695–716, Sep. 2016.

[31] V. Palanisamy and R. Thirunavukarasu, "Implications of big data analytics in developing healthcare frameworks—A review," *J. King Saud Univ., Comput. Inf. Sci.*, vol. 31, no. 4, pp. 415–425, Oct. 2019.

[32] M. J. Ward, K. A. Marsolo, and C. M. Froehle, "Applications of business analytics in healthcare," *Bus. Horizons*, vol. 57, no. 5, pp. 571–582, Sep. 2014.

[33] W. Raghupathi and V. Raghupathi, "An overview of health analytics," *J. Health Med. Informat.*, vol. 4, no. 132, p. 2, 2013.

[34] R. B. D'Agostino, S. Grundy, L. M. Sullivan, and P. Wilson, "Validation of the Framingham coronary heart disease prediction scores: Results of a multiple ethnic groups investigation," *J. Amer. Med. Assoc.*, vol. 286, no. 2, pp. 180–187, 2001.

[35] M. Waqialla and M. I. Razzak, "An ontology-based framework aiming to support cardiac rehabilitation program," *Proc. Comput. Sci.*, vol. 96, pp. 23–32, Jan. 2016.

[36] S. L. Yadav and A. Sohal, "Review paper on big data analytics in cloud computing," *Int. J. Comput. Trends Technol.*, vol. 49, no. 3, pp. 156–160, 2017.

[37] P. Mell and T. Grance, "The NIST definition of cloud computing," Nat. Inst. Standards Technol., Special Publication (NIST SP), Gaithersburg, MD, USA, Sep. 2011, doi: 10.6028/NIST.SP.800-145.

[38] E. A. Bayrak and P. Kirci, "A brief survey on big data in healthcare," *Int. J. Big Data Anal. Healthcare*, vol. 5, no. 1, pp. 1–18, Jan. 2020.

[39] M. Chen et al., *Big Data: Related Technologies, Challenges and Future Prospects*, vol. 100. Springer, 2014.

[40] A. K. Bhadani and D. Jothimani, "Big data: Challenges, opportunities, and realities," in *Effective Big Data Management and Opportunities for Implementation*. Hershey, PA, USA: IGI Global, 2016, pp. 1–24.

[41] G. Manogaran, D. Lopez, C. Thota, K. M. Abbas, S. Pyne, and R. Sundarasekar, "Big data analytics in healthcare Internet of Things," in *Innovative Healthcare Systems for the 21st Century*, H. Qudrat-Ullah and P. Tsasis, Eds. Cham, Switzerland: Springer, 2017, pp. 263–284, doi: 10.1007/978-3-319-55774-8_10.

[42] E. A. Mohammed, B. H. Far, and C. Naugler, "Applications of the MapReduce programming framework to clinical big data analysis: Current landscape and future trends," *BioData Mining*, vol. 7, no. 1, pp. 1–23, Dec. 2014.

[43] P. Zikopoulos, D. Deroos, K. Parasuraman, T. Deutsch, J. Giles, and D. Corrigan, *Harness the Power of Big Data The IBM Big Data Platform*. New York, NY, USA: McGraw-Hill, 2012.

[44] P. Zikopoulos and C. Eaton, *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*. New York, NY, USA: McGraw-Hill, 2011.

[45] X. Zhang and X. Wang, "Intelligent prediction and optimization algorithm for chronic disease rehabilitation in sports using big data," *J. Healthcare Eng.*, vol. 2021, pp. 1–7, Apr. 2021.

[46] A. Oussous, F. Z. Benjelloun, A. A. Lahcen, and S. Belfkih, "Big data technologies: A survey," *J. King Saud Univ., Comput. Inf. Sci.*, vol. 30, no. 4, pp. 431–448, Oct. 2018.

[47] D. Usha and A. Jenil, "A survey of big data processing in perspective of Hadoop and MapReduce," *Int. J. Current Eng. Technol.*, vol. 4, no. 2, pp. 602–606, 2014.

[48] M. Ghasemaghaei, K. Hassanein, and O. Turel, "Increasing firm agility through the use of data analytics: The role of fit," *Decis. Support Syst.*, vol. 101, pp. 95–105, Sep. 2017.

[49] D. P. Acharjya and A. P. Kauser, "A survey on big data analytics: Challenges, open research issues and tools," *Int. J. Adv. Comput. Sci. Appl.*, vol. 7, no. 2, pp. 511–518, 2016.

[50] H. Torkey, M. Atlam, N. El-Fishawy, and H. Salem, "Machine learning model for cancer diagnosis based on RNAseq microarray," *Menoufia J. Electron. Eng. Res.*, vol. 30, no. 1, pp. 65–75, Jan. 2021.

[51] C. A. Alexander and L. Wang, "Big data analytics in heart attack prediction," *J. Nursing Care*, vol. 6, no. 2, pp. 1–9, 2017.

[52] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, "The Hadoop distributed file system," in *Proc. IEEE 26th Symp. Mass Storage Syst. Technol. (MSST)*, May 2010, pp. 1–10.

[53] L. Rajabion, A. A. Shaltooki, M. Taghikhah, A. Ghasemi, and A. Badfar, "Healthcare big data processing mechanisms: The role of cloud computing," *Int. J. Inf. Manage.*, vol. 49, pp. 271–289, Dec. 2019.

[54] T. White, *Hadoop: The Definitive Guide*. Sebastopol, CA, USA: O'Reilly Media, 2012.

[55] N. N. Mall and S. Rana, "Overview of big data and Hadoop," *Imperial J. Interdiscipl. Res.*, vol. 2, no. 5, pp. 1399–1406, 2016.

[56] J. H. Holmes, J. Sun, and N. Peek, "Technical challenges for big data in biomedicine and health: Data sources, infrastructure, and analytics," *Yearbook Med. Informat.*, vol. 23, no. 1, pp. 42–47, 2014.

[57] B. R. Prasad and S. Agarwal, "Comparative study of big data computing and storage tools: A review," *Int. J. Database Theory Appl.*, vol. 9, no. 1, pp. 45–66, Jan. 2016.

[58] N. Khan, I. Yaqoob, I. A. T. Hashem, Z. Inayat, W. K. M. Ali, M. Alam, M. Shiraz, and A. Gani, "Big data: Survey, technologies, opportunities, and challenges," *Sci. World J.*, vol. 2014, Jul. 2014, Art. no. 712826.

[59] C. Coronel and S. Morris, *Database Systems: Design, Implementation, & Management*. Boston, MA, USA: Cengage Learning, 2016.

[60] J.-P. Dijcks, "Oracle: Big data for the enterprise," Oracle, Austin, TX, USA, White Paper 16, 2012.

[61] N. Maheswari and M. Sivagami, "Large-scale data analytics tools: Apache Hive, Pig, and HBase," in *Data Science and Big Data Computing: Frameworks and Methodologies*, Z. Mahmood, Ed. Cham, Switzerland: Springer, 2016, pp. 191–220, doi: 10.1007/978-3-319-31861-5_9.

[62] J. Dean and S. Ghemawat, "MapReduce: Simplified data processing on large clusters," *Commun. ACM*, vol. 51, no. 1, pp. 107–113, Jan. 2008.

[63] E. L. Lydia and M. B. Swarup, "Big data analysis using Hadoop components like flume, MapReduce, pig and hive," *Int. J. Sci., Eng. Comput. Technol.*, vol. 5, no. 11, p. 390, 2015.

[64] H. Dhayne, R. Haque, R. Kilany, and Y. Taher, "In search of big medical data integration solutions—A comprehensive survey," *IEEE Access*, vol. 7, pp. 91265–91290, 2019.

[65] J. Ni, Y. Chen, J. Sha, and M. Zhang, "Hadoop-based distributed computing algorithms for healthcare and clinic data processing," in *Proc. 8th Int. Conf. Internet Comput. Sci. Eng. (ICICSE)*, Nov. 2015, pp. 188–193.

[66] S. V. Poucke, Z. Zhang, M. Schmitz, M. Vukicevic, M. V. Laenen, L. A. Celi, and C. D. Deyne, "Scalable predictive analysis in critically ill patients using a visual open data analysis platform," *PLoS ONE*, vol. 11, no. 1, Jan. 2016, Art. no. e0145791.

[67] W. McClay, N. Yadav, Y. Ozbek, A. Haas, H. Attias, and S. Nagarajan, "A real-time magnetoencephalography brain–computer interface using interactive 3D visualization and the Hadoop ecosystem," *Brain Sci.*, vol. 5, no. 4, pp. 419–440, Sep. 2015.

[68] P. Nathan, *Enterprise Data Workflows with Cascading: Streamlined Enterprise Data Management and Analysis*. Sebastopol, CA, USA: O'Reilly Media, 2013.

[69] S. Mazumder, "Big data tools and platforms," in *Big Data Concepts, Theories, and Applications*, S. Yu and S. Guo, Eds. Cham, Switzerland: Springer, 2016, pp. 29–128, doi: 10.1007/978-3-319-27763-9_2.

[70] K. Krishnan, *Data Warehousing in the Age of Big Data*. Boston, MA, USA: Newnes, 2013.

[71] K. S. Beyer, V. Ercegovac, R. Gemulla, A. Balmin, M. Eltabakh, C.-C. Kanne, F. Ozcan, and E. J. Shekita, "Jaql: A scripting language for large scale semistructured data analysis," *Proc. VLDB Endowment*, vol. 4, no. 12, pp. 1272–1283, Aug. 2011.

[72] H. Karau, *Fast Data Processing With Spark*. Packt Publishing, 2013.

[73] S. S. Sadhana and S. Shetty, "Analysis of diabetic data set using hive and R," *Int. J. Emerg. Technol. Adv. Eng.*, vol. 4, no. 7, pp. 626–629, 2014.

[74] T. Daghistani, R. Shammari, and M. Razzak, "Discovering diabetes complications: An ontology based model," *Acta Inf. Medica*, vol. 23, no. 6, p. 385, 2015.

[75] S. Shaw, A. F. Vermeulen, A. Gupta, and D. Kjerrumgaard, "Hive architecture," in *Practical Hive: A Guide to Hadoop's Data Warehouse System*. Berkeley, CA, USA: Apress, 2016, pp. 37–48, doi: 10.1007/978-1-4842-0271-5_3.

[76] S. Sakr, *Big Data 2.0 Processing Systems: A Survey*, vol. 2142. Springer, 2016.

[77] E. Capriolo, D. Wampler, and J. Rutherglen, *Programming Hive: Data Warehouse and Query Language For Hadoop*. Sebastopol, CA, USA: O'Reilly Media, 2012.

[78] S. Hoffman, *Apache Flume: Distributed Log Collection for Hadoop*. Birmingham, U.K.: Packt Publishing, 2013.

[79] S. Hoffman, *Apache Flume: Distributed Log Collection for Hadoop*. Birmingham, U.K.: Packt Publishing, 2015.

[80] D. Vohra, "Using Apache Sqoop," in *Pro Docker*. Berkeley, CA, USA: Apress, 2016, pp. 151–183, doi: 10.1007/978-1-4842-1830-3_11.

[81] A. Jain, *Instant Apache Sqoop*. Birmingham, U.K.: Packt Publishing, 2013.

[82] R. Shireesha and S. Bhutada, "A study of tools, techniques, and trends for big data analytics," *Int. J. Adv. Comput. Techn. Appl.*, vol. 4, no. 1, pp. 152–158, 2016.

[83] O. M. De Carvalho, E. Roloff, and O. Navaux, "A survey of the state-of-the-art in event processing," in *Proc. 11th Workshop Parallel Distrib. Process. (WSPPD)*, 2013, p. 16.

[84] A. Rabkin and R. Katz, "Chukwa: A system for reliable large-scale log collection," in *Proc. 24th Large Installation Syst. Admin. Conf. (LISA)*, 2010, pp. 1–15.

[85] S. Wadkar and M. Siddalingaiah, "HCatalog and Hadoop in the enterprise," in *Pro Apache Hadoop*. Berkeley, CA, USA: Apress, 2014, pp. 271–282, doi: 10.1007/978-1-4302-4864-4_12.

[86] *Confluent Documentation*. Accessed: Nov. 24, 2022. [Online]. Available: https://docs.confluent.io/home/overview.html

[87] T. W. Dinsmore, "Streaming analytics," in *Disruptive Analytics: Charting Your Strategy for Next-Generation Business Analytics*. Berkeley, CA, USA: Apress, 2016, pp. 117–144, doi: 10.1007/978-1-4842-1311-7_6.

[88] M. Grover, T. Malaska, J. Seidman, and G. Shapira, *Hadoop Application Architectures: Designing Real-World Big Data Applications*. Sebastopol, CA, USA: O'Reilly Media, 2015.

[89] K. Maeda, "Comparative survey of object serialization techniques and the programming supports," *J. Commun. Comput.*, vol. 9, no. 8, pp. 920–928, 2012.

[90] M. K. Islam and A. Srinivasan, *Oozie: The Workflow Scheduler for Hadoop*. Sebastopol, CA, USA: O'Reilly Media, 2015.

[91] B. Lublinsky, K. T. Smith, and A. Yakubovich, *Professional Hadoop Solutions*. Hoboken, NJ, USA: Wiley, 2013.

[92] C. P. Chullipparambil, "Big data analytics using Hadoop tools," M.S. thesis, Dept. Comput. Sci., San Diego State Univ., 2016.

[93] S. Landset, T. M. Khoshgoftaar, A. N. Richter, and T. Hasanin, "A survey of open source tools for machine learning with big data in the Hadoop ecosystem," *J. Big Data*, vol. 2, no. 1, pp. 1–36, Dec. 2015.

[94] S. Lovalekar, "Big data: An emerging trend in future," *Int. J. Comput. Sci. Inf. Technol.*, vol. 5, no. 1, pp. 538–541, 2014.

[95] J. Sangeetha and V. S. J. Prakash, "A survey on big data mining techniques," *Int. J. Comput. Sci. Inf. Secur.*, vol. 15, no. 1, p. 482, 2017.

[96] W. Raghupathi and V. Raghupathi, "Big data analytics in healthcare: Promise and potential," *Health Inf. Sci. Syst.*, vol. 2, no. 1, pp. 1–10, Dec. 2014.

[97] R. Raja, I. Mukherjee, and B. K. Sarkar, "A systematic review of healthcare big data," *Sci. Program.*, vol. 2020, pp. 1–15, Jul. 2020.

[98] A. M.-H. Kuo, "Opportunities and challenges of cloud computing to improve health care services," *J. Med. Internet Res.*, vol. 13, no. 3, p. e1867, Sep. 2011.

[99] S. P. Ahuja, S. Mani, and J. Zambrano, "A survey of the state of cloud computing in healthcare," *Netw. Commun. Technol.*, vol. 1, no. 2, p. 12, Sep. 2012.

[100] G. Aceto, V. Persico, and A. Pescapé, "The role of information and communication technologies in healthcare: Taxonomies, perspectives, and challenges," *J. Netw. Comput. Appl.*, vol. 107, pp. 125–154, Apr. 2018.

[101] D. Uppal, R. Sinha, V. Mehra, and V. Jain, "Malware detection and classification based on extraction of API sequences," in *Proc. Int. Conf. Adv. Comput., Commun. Informat. (ICACCI)*, Sep. 2014, pp. 2337–2342.

[102] P. D. Kaur and I. Chana, "Cloud based intelligent system for delivering health care as a service," *Comput. Methods Programs Biomed.*, vol. 113, no. 1, pp. 346–359, Jan. 2014.

[103] V. Assadi and K. Hassanein, "Consumer adoption of personal health record systems: A self-determination theory perspective," *J. Med. Internet Res.*, vol. 19, no. 7, p. e270, Jul. 2017.

[104] N. Surantha, T. F. Lesmana, and S. M. Isa, "Sleep stage classification using extreme learning machine and particle swarm optimization for healthcare big data," *J. Big Data*, vol. 8, no. 1, pp. 1–17, Dec. 2021.

[105] N. Surantha, G. P. Kusuma, and S. M. Isa, "Internet of Things for sleep quality monitoring system: A survey," in *Proc. 11th Int. Conf. Knowl., Inf. Creativity Support Syst. (KICSS)*, Nov. 2016, pp. 1–6.

[106] L. Qian, Z. Luo, Y. Du, and L. Guo, "Cloud computing: An overview," in *Cloud Computing*, M. G. Jaatun, G. Zhao, and C. Rong, Eds. Berlin, Germany: Springer, Dec. 2009, pp. 626–631.

[107] A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," *Int. J. Inf. Manage.*, vol. 35, no. 2, pp. 137–144, Apr. 2015.

[108] K. M. Hassan, A. Abdo, and A. Yakoub, "Enhancement of health care services based on cloud computing in IoT environment using hybrid swarm intelligence," *IEEE Access*, vol. 10, pp. 105877–105886, 2022.

[109] G.-H. Kim, S. Trimi, and J.-H. Chung, "Big-data applications in the government sector," *Commun. ACM*, vol. 57, no. 3, pp. 78–85, Mar. 2014.

[110] E. Vayena, M. Salathé, L. C. Madoff, and J. S. Brownstein, "Ethical challenges of big data in public health," *PLOS Comput. Biol.*, vol. 11, no. 2, Feb. 2015, Art. no. e1003904.

[111] C. Burghard, "Big data and analytics key to accountable care success," *IDC Health Insights*, vol. 1, pp. 1–9, Oct. 2012.

[112] L. M. Fernandes, M. O'Connor, and V. Weaver, "Big data, bigger outcomes," *J. AHIMA*, vol. 83, no. 10, pp. 38–43, 2012.

[113] B. N. Subudhi, D. K. Rout, and A. Ghosh, "Big data analytics for video surveillance," *Multimedia Tools Appl.*, vol. 78, no. 18, pp. 26129–26162, Sep. 2019.

[114] A. Juneja, S. Juneja, V. Bali, and S. Mahajan, "Multi-criterion decision making for wireless communication technologies adoption in IoT," *Int. J. Syst. Dyn. Appl.*, vol. 10, no. 1, pp. 1–15, Jan. 2021.

[115] M. Uppal, D. Gupta, S. Juneja, G. Dhiman, and S. Kautish, "Cloud-based fault prediction using IoT in office automation for improvisation of health of employees," *J. Healthcare Eng.*, vol. 2021, pp. 1–13, Nov. 2021.

[116] L. Kansal, G. S. Gaba, A. Sharma, G. Dhiman, M. Baz, and M. Masud, "Performance analysis of WOFDM-WiMAX integrating diverse wavelets for 5G applications," *Wireless Commun. Mobile Comput.*, vol. 2021, pp. 1–14, Nov. 2021.

[117] A. Juneja, S. Juneja, S. Kaur, and V. Kumar, "Predicting diabetes mellitus with machine learning techniques using multi-criteria decision making," *Int. J. Inf. Retr. Res.*, vol. 11, no. 2, pp. 38–52, Apr. 2021.

[118] S. Juneja, A. Juneja, G. Dhiman, S. Jain, A. Dhankhar, and S. Kautish, "Computer vision-enabled character recognition of hand gestures for patients with hearing and speaking disability," *Mobile Inf. Syst.*, vol. 2021, pp. 1–10, Dec. 2021.

[119] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers, *Big Data: The Next Frontier for Innovation, Competition, and Productivity*. Washington, DC, USA: McKinsey Global Institute, 2011.

[120] B. F. Hankey, L. A. Ries, and B. K. Edwards, "The surveillance, epidemiology, and end results program: A national resource," *Cancer Epidemiol. Biomarkers Prevention*, vol. 8, no. 12, pp. 1117–1121, 1999.

[121] S. Tsugawa, Y. Mogi, Y. Kikuchi, F. Kishino, K. Fujita, Y. Itoh, and H. Ohsaki, "On estimating depressive tendencies of Twitter users utilizing their tweet data," in *Proc. IEEE Virtual Reality (VR)*, Mar. 2013, pp. 1–4.

[122] A. Sadilek, H. Kautz, and V. Silenzio, "Modeling spread of disease from social interactions," in *Proc. Int. AAAI Conf. Web Social Media*, 2012, vol. 6, no. 1, pp. 322–329.

[123] G. Dhiman, S. Juneja, H. Mohafez, I. El-Bayoumy, L. K. Sharma, M. Hadizadeh, M. A. Islam, W. Viriyasitavat, and M. U. Khandaker, "Federated learning approach to protect healthcare data over big data scenario," *Sustainability*, vol. 14, no. 5, p. 2500, Feb. 2022.

[124] M. A. Al-Khasawneh, A. Bukhari, and A. M. Khasawneh, "Effective of smart mathematical model by machine learning classifier on big data in healthcare fast response," *Comput. Math. Methods Med.*, vol. 2022, pp. 1–9, Feb. 2022.

[125] J. Andreu-Perez, C. C. Y. Poon, R. D. Merrifield, S. T. C. Wong, and G.-Z. Yang, "Big data for health," *IEEE J. Biomed. Health Informat.*, vol. 19, no. 4, pp. 1193–1208, Jul. 2015.

[126] C. H. Lee and H.-J. Yoon, "Medical big data: Promise and challenges," *Kidney Res. Clin. Pract.*, vol. 36, no. 1, pp. 3–11, Mar. 2017.

[127] A. R. Reddy and P. S. Kumar, "Predictive big data analytics in healthcare," in *Proc. 2nd Int. Conf. Comput. Intell. Commun. Technol. (CICT)*, Feb. 2016, pp. 623–626.

[128] D. V. Dimitrov, "Medical Internet of Things and big data in healthcare," *Healthcare Inform. Res.*, vol. 22, no. 3, pp. 156–163, 2016.

[129] B. K. Sarkar, "Big data for secure healthcare system: A conceptual design," *Complex Intell. Syst.*, vol. 3, no. 2, pp. 133–151, Jun. 2017.

[130] P. P. Ray, "A survey of IoT cloud platforms," *Future Comput. Informat. J.*, vol. 1, nos. 1–2, pp. 35–46, Dec. 2016.

[131] J. L. Shah and H. F. Bhat, "CloudIoT for smart healthcare: Architecture, issues, and challenges," in *Internet of Things Use Cases for the Healthcare Industry*. Cham, Switzerland: Springer, 2020, pp. 87–126, doi: 10.1007/978-3-030-37526-3_5.

[132] F. Khan, A. U. Rehman, A. Yahya, M. A. Jan, J. Chuma, Z. Tan, and K. Hussain, "A quality of service-aware secured communication scheme for Internet of Things-based networks," *Sensors*, vol. 19, no. 19, p. 4321, Oct. 2019.

[133] M. Bowya and V. Karthikeyan, "A novel secure IoT based optimizing sensor network for automatic medicine composition prescribe system," in *Inventive Communication and Computational Technologies*, G. Ranganathan, J. Chen, and Á. Rocha, Eds. Singapore: Springer, 2020, pp. 1109–1118.

[134] T. Pflanzner and A. Kertész, "A survey of IoT cloud providers," in *Proc. 39th Int. Conv. Inf. Commun. Technol., Electron. Microelectron. (MIPRO)*, May 2016, pp. 730–735.

[135] P. P. Jayaraman, C. Perera, D. Georgakopoulos, S. Dustdar, D. Thakker, and R. Ranjan, "Analytics-as-a-service in a multi-cloud environment through semantically-enabled hierarchical data processing," *Softw., Pract. Exp.*, vol. 47, no. 8, pp. 1139–1156, Aug. 2017.

[136] J. Chen, F. Qian, W. Yan, and B. Shen, ''Translational biomedical informatics in the cloud: Present and future,'' *BioMed Res. Int.*, vol. 2013, pp. 1–8, Jan. 2013.

[137] L. Hong, M. Luo, R. Wang, P. Lu, W. Lu, and L. Lu, ''Big data in health care: Applications and challenges,'' *Data Inf. Manage.*, vol. 2, no. 3, pp. 175–197, Dec. 2018.

[138] P. Kaur, M. Sharma, and M. Mittal, ''Big data and machine learning based secure healthcare framework,'' *Proc. Comput. Sci.*, vol. 132, pp. 1049–1059, Jan. 2018.

[139] A. Gunasekaran, T. Papadopoulos, R. Dubey, S. F. Wamba, S. J. Childe, B. Hazen, and S. Akter, ''Big data and predictive analytics for supply chain and organizational performance,'' *J. Bus. Res.*, vol. 70, pp. 308–317, Jan. 2017.

[140] J. N. Undavia and A. M. Patel, ''Big data analytics in healthcare: Applications and challenges,'' *Int. J. Big Data Anal. Healthcare*, vol. 5, no. 1, pp. 19–27, 2020.

[141] P. E. Beeler, D. W. Bates, and B. L. Hug, ''Clinical decision support systems,'' *Swiss Med. Weekly*, vol. 144, no. 5152, 2014, Art. no. w14073.

[142] C. S. Mayo et al., ''American association of physicists in medicine task group 263: Standardizing nomenclatures in radiation oncology,'' *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 100, no. 4, pp. 1057–1066, Mar. 2018.

[143] J. Luo, M. Wu, D. Gopukumar, and Y. Zhao, ''Big data application in biomedical research and health care: A literature review,'' *Biomed. Informat. Insights*, vol. 8, Jan. 2016, Art. no. BII.S31559.

[144] P. Galetsi, K. Katsaliaki, and S. Kumar, ''Values, challenges and future directions of big data analytics in healthcare: A systematic review,'' *Social Sci. Med.*, vol. 241, Nov. 2019, Art. no. 112533.

[145] T. Heart, O. Ben-Assuli, and I. Shabtai, ''A review of PHR, EMR and EHR integration: A more personalized healthcare and public health policy,'' *Health Policy Technol.*, vol. 6, no. 1, pp. 20–25, Mar. 2017.

[146] M. Viceconti, P. Hunter, and R. Hose, ''Big data, big knowledge: Big data for personalized healthcare,'' *IEEE J. Biomed. Health Informat.*, vol. 19, no. 4, pp. 1209–1215, Jul. 2015.

[147] Y. Zhang, L. Zhang, E. Oki, N. V. Chawla, and A. Kos, ''IEEE access special section editorial: Big data analytics for smart and connected health,'' *IEEE Access*, vol. 4, pp. 9906–9909, 2016.

[148] G. Phillips-Wren, L. S. Iyer, U. Kulkarni, and T. Ariyachandra, ''Business analytics in the context of big data: A roadmap for research,'' *Commun. Assoc. Inf. Syst.*, vol. 37, p. 23, 2015.

[149] H. J. Watson, ''Tutorial: Big data analytics: Concepts, technologies, and applications,'' *Commun. Assoc. for Inf. Syst.*, vol. 34, no. 1, p. 65, 2014.

[150] Y.-K. Lin, H. Chen, R. A. Brown, S.-H. Li, and H.-J. Yang, ''Healthcare predictive analytics for risk profiling in chronic care: A Bayesian multitask learning approach,'' *MIS Quart.*, vol. 41, no. 2, pp. 473–495, Feb. 2017.

[151] N. Antoine-Moussiaux, O. Vandenberg, Z. Kozlakidis, C. Aenishaenslin, M. Peyre, M. Roche, P. Bonnet, and A. Ravel, ''Valuing health surveillance as an information system: Interdisciplinary insights,'' *Frontiers Public Health*, vol. 7, p. 138, Jun. 2019.

[152] J. Archenaa and E. A. M. Anita, ''A survey of big data analytics in healthcare and government,'' *Proc. Comput. Sci.*, vol. 50, pp. 408–413, Jan. 2015.

[153] M. Salomi and S. A. A. Balamurugan, ''Need, application and characteristics of big data analytics in healthcare—A survey,'' *Indian J. Sci. Technol.*, vol. 9, no. 16, pp. 1–5, May 2016.

[154] M. Gowsalya, K. Krushitha, and C. Valliyammai, ''Predicting the risk of readmission of diabetic patients using MapReduce,'' in *Proc. 6th Int. Conf. Adv. Comput. (ICoAC)*, Dec. 2014, pp. 297–301.

[155] M. Aiello, C. Cavaliere, A. D'Albore, and M. Salvatore, ''The challenges of diagnostic imaging in the era of big data,'' *J. Clin. Med.*, vol. 8, no. 3, p. 316, Mar. 2019.

[156] F. Amalina, I. A. Targio Hashem, Z. H. Azizul, A. T. Fong, A. Firdaus, M. Imran, and N. B. Anuar, ''Blending big data analytics: Review on challenges and a recent study,'' *IEEE Access*, vol. 8, pp. 3629–3645, 2020.

[157] N. Szlezák, M. Evers, J. Wang, and L. Pérez, ''The role of big data and advanced analytics in drug discovery, development, and commercialization,'' *Clin. Pharmacol. Therapeutics*, vol. 95, no. 5, pp. 492–495, May 2014.

[158] J. Wu, H. Li, S. Cheng, and Z. Lin, ''The promising future of healthcare services: When big data analytics meets wearable technology,'' *Inf. Manage.*, vol. 53, no. 8, pp. 1020–1033, Dec. 2016.

[159] Y. Wang, L. Kung, W. Y. C. Wang, and C. G. Cegielski, ''An integrated big data analytics-enabled transformation model: Application to health care,'' *Inf. Manage.*, vol. 55, no. 1, pp. 64–79, Jan. 2018.

[160] F. Zhang, J. Cao, S. U. Khan, K. Li, and K. Hwang, ''A task-level adaptive MapReduce framework for real-time streaming data in healthcare applications,'' *Future Gener. Comput. Syst.*, vols. 43–44, pp. 149–160, Feb. 2015.

[161] T. R. McNutt, K. L. Moore, and H. Quon, ''Needs and challenges for big data in radiation oncology,'' *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 95, no. 3, pp. 909–915, Jul. 2016.

[162] C. Sáez and J. M. García-Gómez, ''Kinematics of big biomedical data to characterize temporal variability and seasonality of data repositories: Functional data analysis of data temporal evolution over non-parametric statistical manifolds,'' *Int. J. Med. Informat.*, vol. 119, pp. 109–124, Nov. 2018.

[163] S. Khanra, A. Dhir, A. K. M. N. Islam, and M. Mäntymäki, ''Big data analytics in healthcare: A systematic literature review,'' *Enterprise Inf. Syst.*, vol. 14, no. 7, pp. 878–912, Aug. 2020.

[164] H. Liu and H. Motoda, *Computational Methods of Feature Selection*. Boca Raton, FL, USA: CRC Press, 2007.

[165] K. U. Jaseena and J. M. David, ''Issues, challenges, and solutions: Big data mining,'' *Comput. Sci. Inf. Technol.*, vol. 4, no. 13, pp. 131–140, Dec. 2014.

[166] R. H. Hariri, E. M. Fredericks, and K. M. Bowers, ''Uncertainty in big data analytics: Survey, opportunities, and challenges,'' *J. Big Data*, vol. 6, no. 1, pp. 1–16, Dec. 2019.

[167] A. Lenk, L. Bonorden, A. Hellmanns, N. Roedder, and S. Jaehnichen, ''Towards a taxonomy of standards in smart data,'' in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Oct. 2015, pp. 1749–1754.

[168] F. Soleimani-Roozbahani, A. R. Ghatari, and R. Radfar, ''Knowledge discovery from a more than a decade studies on healthcare big data systems: A scientometrics study,'' *J. Big Data*, vol. 6, no. 1, pp. 1–15, Dec. 2019.

[169] P. Russom, ''Big data analytics,'' *TDWI Best Practices Rep.*, vol. 19, no. 4, pp. 1–34, 2011.

[170] D. Laney, ''3D data management: Controlling data volume, velocity and variety,'' *META Group Res. Note*, vol. 6, no. 70, p. 1, 2001.

[171] A. Sathi, ''Big data analytics: Disruptive technologies for changing the game,'' 2012. [Online]. Available: https://api.semanticscholar.org/CorpusID:86848357

[172] C.-W. Tsai, C.-F. Lai, H.-C. Chao, and A. V. Vasilakos, ''Big data analytics: A survey,'' *J. Big Data*, vol. 2, no. 1, pp. 1–32, Dec. 2015.

[173] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, ''Internet of Things (IoT): A vision, architectural elements, and future directions,'' *Future Gener. Comput. Syst.*, vol. 29, no. 7, pp. 1645–1660, Sep. 2013.

[174] M. White, ''Digital workplaces: Vision and reality,'' *Bus. Inf. Rev.*, vol. 29, no. 4, pp. 205–214, Dec. 2012.

[175] W. Jatmiko, D. M. S. Arsa, H. Wisesa, G. Jati, and M. A. Ma'sum, ''A review of big data analytics in the biomedical field,'' in *Proc. Int. Workshop Big Data Inf. Secur. (IWBIS)*, Oct. 2016, pp. 31–41.

[176] K. Wan and V. Alagar, ''Characteristics and classification of big data in health care sector,'' in *Proc. 12th Int. Conf. Natural Comput., Fuzzy Syst. Knowl. Discovery (ICNC-FSKD)*, Aug. 2016, pp. 1439–1446.

[177] P. K. Sahoo, S. K. Mohapatra, and S.-L. Wu, ''Analyzing healthcare big data with prediction for future health condition,'' *IEEE Access*, vol. 4, pp. 9786–9799, 2016.

[178] U. Sivarajah, M. M. Kamal, Z. Irani, and V. Weerakkody, ''Critical analysis of big data challenges and analytical methods,'' *J. Bus. Res.*, vol. 70, pp. 263–286, Jan. 2017.

[179] P. Pant and R. Tanwar, ''An overview of big data opportunity and challenges,'' in *Smart Trends in Information Technology and Computer Communications*, A. Unal, M. Nayak, D. K. Mishra, D. Singh, and A. Joshi, Eds. Singapore: Springer, 2016, pp. 691–697.

[180] K. F. Tiampo, S. McGinnis, Y. Kropivnitskaya, J. Qin, and M. A. Bauer, ''Big data challenges and hazards modeling,'' in *Risk Modeling for Hazards and Disasters*. Amsterdam, The Netherlands: Elsevier, 2018, pp. 193–210.

[181] N. R. Vajjhala, K. D. Strang, and Z. Sun, ''Statistical modeling and visualizing open big data using a terrorism case study,'' in *Proc. 3rd Int. Conf. Future Internet Things Cloud*, Aug. 2015, pp. 489–496.

[182] D. Saidulu and R. Sasikala, ''Machine learning and statistical approaches for big data: Issues, challenges and research directions,'' *Int. J. Appl. Eng. Res.*, vol. 12, no. 21, pp. 11691–11699, 2017.

[183] B. Cyganek, M. Graña, B. Krawczyk, A. Kasprzak, P. Porwik, K. Walkowiak, and M. Woźniak, ''A survey of big data issues in electronic health record analysis,'' *Appl. Artif. Intell.*, vol. 30, no. 6, pp. 497–520, Jul. 2016.

[184] A. Jain. (2017). *The 5 V's of Big Data—Watson Health Perspectives*. Accessed: Oct. 5, 2023. [Online]. Available: https://deepstash.com/article/113400/the-5-vs-of-big-data-watson-health-perspectives

[185] K. Verspoor and F. Martin-Sanchez, "Big data in medicine is driving big changes," *Yearbook Med. Informat.*, vol. 23, no. 1, pp. 14–20, 2014.

[186] A. Rehman, S. Naz, and I. Razzak, "Leveraging big data analytics in healthcare enhancement: Trends, challenges and opportunities," *Multimedia Syst.*, vol. 28, pp. 1339–1371, Aug. 2022.

[187] J. Pokornỳ, P. Škoda, I. Zelinka, D. Bednárek, F. Zavoral, M. Kruliš, and P. Šaloun, "Big data movement: A challenge in data processing," in *Big Data in Complex Systems: Challenges and Opportunities*, A. E. Hassanien, A. T. Azar, V. Snasael, J. Kacprzyk, and J. H. Abawajy, Eds. Cham, Switzerland: Springer, 2015, pp. 29–69, doi: 10.1007/978-3-319-11056-1_2.

[188] IBMB Data, "Extracting business value from the 4 V's of big data," *Retrieved July*, vol. 19, p. 2017, Jun. 2016.

[189] N. Deepa, Q.-V. Pham, D. C. Nguyen, S. Bhattacharya, B. Prabadevi, T. R. Gadekallu, P. K. R. Maddikunta, F. Fang, and P. N. Pathirana, "A survey on blockchain for big data: Approaches, opportunities, and future directions," *Future Gener. Comput. Syst.*, vol. 131, pp. 209–226, Jun. 2022.

[190] D. Court, "Getting big impact from big data," *McKinsey Quart.*, vol. 1, no. 1, pp. 52–60, 2015.

[191] K. Yadav and Y. Hasija, "IoT and big data inter-relation: A boom in biomedical healthcare," in *Proc. IEEE Delhi Sect. Conf. (DELCON)*, Feb. 2022, pp. 1–6.

[192] A. Nargundkar and A. J. Kulkarni, "Big data in supply chain management and medicinal domain," in *Big Data Analytics in Healthcare*, A. J. Kulkarni, P. Siarry, P. K. Singh, A. Abraham, M. Zhang, A. Zomaya, and F. Baki, Eds. Cham, Switzerland: Springer, 2020, pp. 45–54, doi: 10.1007/978-3-030-31672-3_3.

[193] J. Bresnick, "Understanding the many V's of healthcare big data analytics," *Retrieved July*, vol. 20, p. 2018, Jun. 2017.

FATIMA HUSSAIN is currently pursuing the Ph.D. degree in computer science with The Islamia University of Bahawalpur (IUB), Pakistan. Her current research interests include big data analytics, medical diagnostics, clinical decision-making, machine learning, artificial neural networks, explainable artificial intelligence, data mining, and decision trees.

MUHAMMAD NAUMAN received the Ph.D. degree in computer science from The Islamia University of Bahawalpur (IUB), Pakistan. He is currently a Lecturer with IUB. His current research interests include formal approaches, formal methods, big data, big data analytics, machine learning, data mining, explainable artificial intelligence, artificial neural networks, decision trees, and predictive models.

ABDULLAH ALGHURIED received the Ph.D. degree in industrial engineering from the University of Miami, in 2020. He is currently an Assistant Professor of industrial engineering with the University of Tabuk. His current research interests include stochastic modeling and optimization, data mining, big data, data analytics, decision-making under uncertainty, lean six sigma, and sustainability.

ADI ALHUDHAIF received the bachelor's degree in computer science from King Saud University, the master's degree in computer science (information security and big data) from George Washington University, Washington D.C., the master's degree in law LLM (internet law) from The University of Strathclyde, Glasgow, U.K., and the Ph.D. degree in computer science (information security and big data) from George Washington University. He received several IT professional certificates in IT governance, risk management, and project management.

NADEEM AKHTAR received the master's degree in information systems from Archite Institut Universitaire Professionnalis Vannes, France, and the Ph.D. degree (Hons.) from the Laboratory VALORIA of Computer Science, University of South Brittany (UBS), France. He was a recipient of several awards, scholarships, and research grants, such as the 2004 French Embassy Scholarship for the master's studies in France; the HEC, France, the Teaching Assistant for ENSIBS=million; the 2014 Research Award from the Directorate of Research and Development, The Islamia University of Bahawalpur.