IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 23 September 2023, accepted 7 October 2023, date of publication 10 October 2023, date of current version 18 October 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3323574

SURVEY

Harnessing Big Data Analytics for Healthcare: A Comprehensive Review of Frameworks, Implications, Applications, and Impacts

AWAIS AHMED[1], (Member, IEEE), RUI XI[1], (Member, IEEE), MENGSHU HOU[1,2], (Member, IEEE), SYED ATTIQUE SHAH[3], (Senior Member, IEEE), AND SUFIAN HAMEED[4], (Member, IEEE)

[1]School of Computer Science and Engineering, University of Electronic Science and Technology of China (UESTC), Chengdu, Sichuan 611731, China
[2]School of Big Data and Artificial Intelligence, Chengdu Technological University, Chengdu, Sichuan 611730, China
[3]School of Computing and Digital Technology, Birmingham City University, B5 5JU Birmingham, U.K.
[4]Department of Computer Science, National University of Computer and Emerging Sciences, Karachi 75160, Pakistan

Corresponding authors: Rui Xi (ruix.ryan@gmail.com) and Mengshu Hou (meshou@uestc.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 62072075, and in part by the Sichuan Province Key Research and Development Project under Grant 2023YFS0420.

ABSTRACT Big Data Analytics (BDA) has garnered significant attention in both academia and industries, particularly in sectors such as healthcare, owing to the exponential growth of data and advancements in technology. The integration of data from diverse sources and the utilization of advanced analytical techniques has the potential to revolutionize healthcare by improving diagnostic accuracy, enabling personalized medicine, and enhancing patient outcomes. In this paper, we aim to provide a comprehensive literature review on the application of big data analytics in healthcare, focusing on its ecosystem, applications, and data sources. To achieve this, an extensive analysis of scientific studies published between 2013 and 2023 was conducted and overall 180 scientific studies were thoroughly evaluated, establishing a strong foundation for future research and identifying collaboration opportunities in the healthcare domain. The study delves into various application areas of BDA in healthcare, highlights successful implementations, and explores their potential to enhance healthcare outcomes while reducing costs. Additionally, it outlines the challenges and limitations associated with BDA in healthcare, discusses modelling tools and techniques, showcases deployed solutions, and presents the advantages of BDA through various real-world use cases. Furthermore, this study identifies and discusses key open research challenges in the field of big data analytics in healthcare, aiming to push the boundaries and contribute to enhanced healthcare outcomes and decision-making processes.

INDEX TERMS Big data analytics, healthcare information systems, systematic literature review, multimodal big data, data-driven decisions, natural language processing, block-chain, electronic health records.

I. INTRODUCTION

In healthcare, the generation of vast amounts of patient data is aimed at improving care quality and cost reduction. However, effectively analyzing this data presents a significant challenge in identifying trends and patterns for problem-solving [1]. The advancement of information and communication technologies has facilitated the sharing of health information, enabling more sophisticated analysis of big data [2]. Big data analytics is recognized as a valuable tool for knowledge discovery from centralized and distributed databases. Leveraging advanced statistical models and machine learning algorithms, healthcare organizations can develop accurate and personalized treatments while identifying cost-saving opportunities. Additionally, big data analytics can optimize operational efficiency, reducing patient wait times and

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

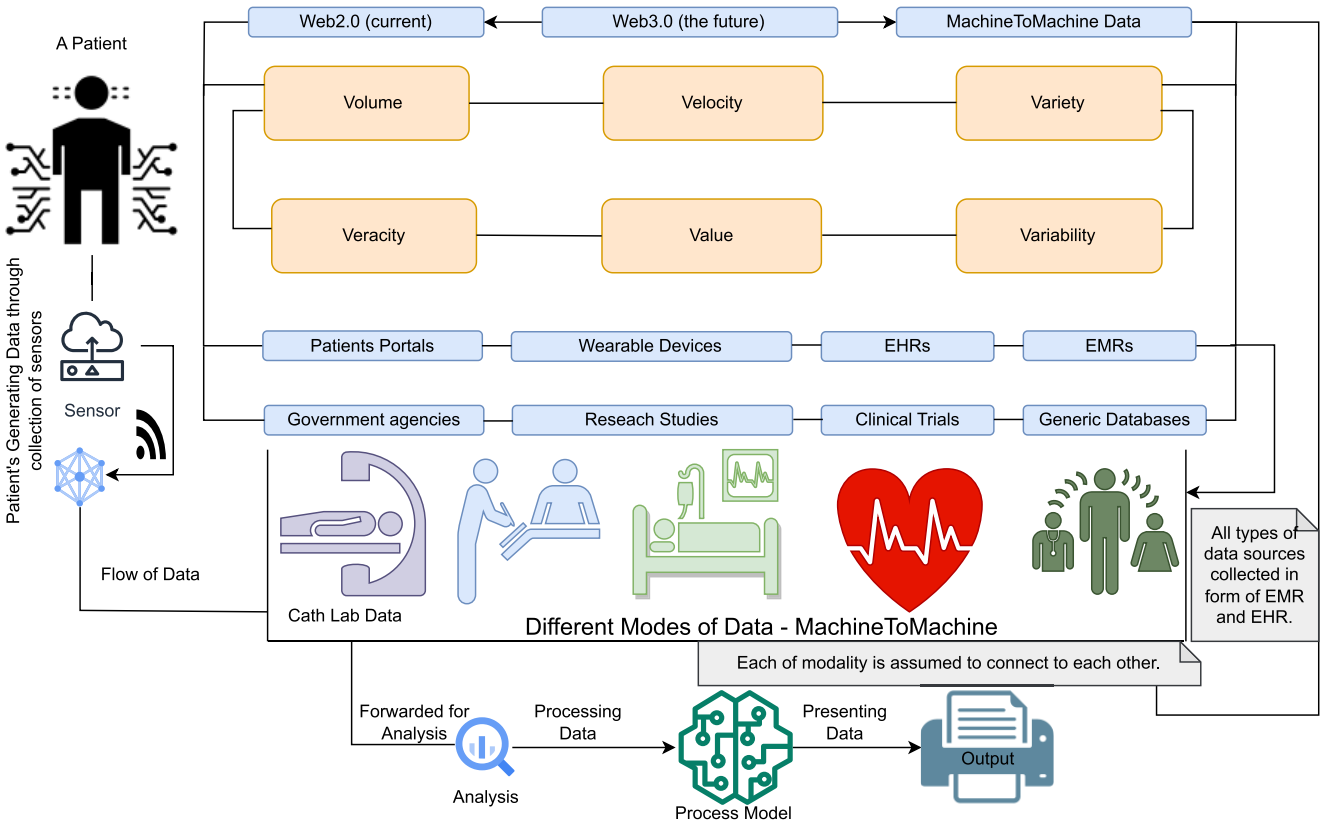

FIGURE 1. The general characteristics of healthcare big data analytics with respect to the sources of data.

enhancing the effectiveness of medical staff. Ensuring data privacy and security is paramount as the adoption of big data analytics expands. Healthcare organizations can mitigate risks associated with data breaches and cyberattacks by implementing robust data governance frameworks and adhering to best practices for data handling [3].

The availability of vast amounts of medical big data has brought about a revolution in the healthcare industry. Leveraging big data analytics to share, analyze, and process this data holds immense potential for identifying treatment patterns and reducing healthcare costs. However, it is of utmost importance to prioritize the privacy and security of medical data, implementing robust measures and regulatory frameworks to safeguard patients' information [4]. Any delay in treatment resulting from concerns about privacy or security can have life-threatening consequences, underscoring the need for healthcare providers to establish reliable infrastructure for systematic big data analytics. The integration of big data analytics in healthcare represents a significant advancement in improving public health, empowering healthcare providers to devise more effective treatment plans and enhance patient outcomes.

The healthcare industry is constantly confronted with vast amounts of data originating from diverse sources like smart devices, electronic health records, medical imaging, and genomics data [5]. Managing and analyzing this data is a challenge due to its unstructured and complex nature. As shown in Figure 1, the diagram illustrates the general flow of data within the big data analytics ecosystem in healthcare. Technological advancements, including cloud computing, machine learning, and natural language processing, have significantly improved the management of healthcare big data by enabling efficient processing and analysis. The COVID-19 pandemic has further highlighted the importance of digitizing medical records and employing telemedicine, resulting in a substantial increase in data volume and placing additional pressure on the healthcare industry to effectively manage and secure this data. Consequently, efforts to develop robust security measures and privacy protocols have intensified to safeguard sensitive patient information. Block-chain technology presents a potential solution by offering a decentralized and secure approach to data storage and sharing. Furthermore, techniques like differential privacy can be employed to anonymous data, preserving patient privacy while allowing meaningful analysis of big data [6]. As the healthcare sector continues to embrace big data technologies, ethical considerations such as data ownership, informed consent, and transparency must also be addressed. Establishing clear guidelines and regulations for big data usage in healthcare ensures its ethical and responsible application, benefiting patients and the healthcare industry as a whole.

A. MOTIVATION AND SCOPE

The healthcare sector is grappling with various challenges, encompassing rising expenses, growing demand for quality healthcare solutions, increasing patient expectations, and a shortfall of professionals, etc [7]. The complexity of healthcare data arises from the fact that it is generated by diverse sources and systems such as electronic health records, medical imaging devices, and wearables, each with its own data structure and format. In addition, the data must comply with privacy regulations such as HIPAA and GDPR, which further complicates the process of data integration and analysis [8]. The increasing availability of healthcare data and advancements in technology have paved the way for harnessing big data analytics in healthcare. However, there is a pressing need for a comprehensive review that synthesizes the existing frameworks, explores the implications, examines the diverse applications, and assesses the overall impacts of big data analytics in healthcare.

The research is motivated by the need to establish a comprehensive view of utilized frameworks. This empowers researchers and practitioners to make informed decisions and adopt appropriate approaches. Additionally, exploring the implications can guide stakeholders in understanding the potential benefits and challenges associated with implementing big data analytics in healthcare. By conducting a Systematic Literature Review (SLR) on this topic, we want to provide valuable insights and contribute to the understanding and advancement of this field and contribute towards identifying successful applications and assessing their impact, researchers can drive further development and exploration. The implementation of BDA in healthcare requires a multidisciplinary approach, involving experts from various fields such as computer science, statistics, and healthcare professionals. The use of specialized technologies such as BDA, machine learning algorithms, and natural language processing can aid in the processing and analysis of healthcare data, enabling healthcare providers to gain insights into patient care and outcomes, disease patterns, and healthcare utilization. Despite the challenges, the potential benefits of BDA in healthcare are enormous, including improved patient outcomes, reduced healthcare costs, and more efficient healthcare delivery.

In this SLR, our primary goal is to provide a comprehensive exploration of the essential elements for harnessing the power of BDA within the healthcare domain. To achieve this, we strategically structured our paper to ensure a clear progression of topics that collectively paint a holistic picture of the intersection between advanced analytics and healthcare. We examined healthcare applications of big data, including how data science, machine learning, natural language processing, and deep learning can be harnessed to address real-world healthcare challenges. It was our intention to provide concrete examples and insights into how these tools can be leveraged to enhance vital signs monitoring, predict diseases, optimize patient management, and improve hospital operations.

The main contributions of this paper are mentioned below:

1) An in-depth analysis of the different components of BDA and how they interact with each other. This helps readers understand the complexities of BDA in healthcare and how it can be utilized effectively.
2) A detailed analysis of the promising application areas of BDA in healthcare. The paper discusses successful implementations of BDA in various healthcare areas and how they have improved healthcare outcomes and reduced costs. This information is valuable for healthcare practitioners and researchers who are interested in implementing BDA in their organizations.
3) A presentation of the challenges and limitations of BDA in healthcare. By highlighting these challenges, the paper helps readers understand the potential barriers to implementing BDA in healthcare and how they can be overcome.
4) A list of reliable and authentic sources of healthcare analytics that researchers and practitioners can use.
5) The paper also provides an inventory of modelling tools, techniques, and deployed solutions.
6) Finally, the paper highlights the advantages of using BDA in healthcare through various use cases. By presenting these use cases, the paper demonstrates the potential impact of BDA on healthcare outcomes and costs.

To the best of our knowledge, this review is among very few comprehensive studies that shed light on above mentioned contributions. This review is a valuable resource for anyone interested in the potential benefits and challenges of BDA in healthcare.

B. ORGANIZATION OF THE REVIEW

The rest of the paper is organized as; Section II presents the published surveys focusing on the same area and their limitations. Section III discusses the overall SLR method used for this paper. Section IV describes an ecosystem for big data employed in healthcare and it answers the RQ-1. Section V discusses the big data applications in Healthcare by answering the RQ-2. Section VI presents a detailed answer to RQ-3 in its subsections. Section VII presents open research challenges that were identified during the course of this research and it also answers RQ4. Section VIII provides a discussion and implications related to the research questions. Finally, Section IX concludes the study. Figure 2 depicts the section layout of the review paper. Table 1 provides a list of acronyms used in this study.

II. EXISTING SURVEYS

While preparing our Systematic Literature Survey we noticed several surveys have been undertaken in the extant literature to investigate the prospects and challenges associated with big data analytics and the healthcare domain. We also observed that current surveys remain focused on foundational basics and challenges in big data healthcare. Such as the

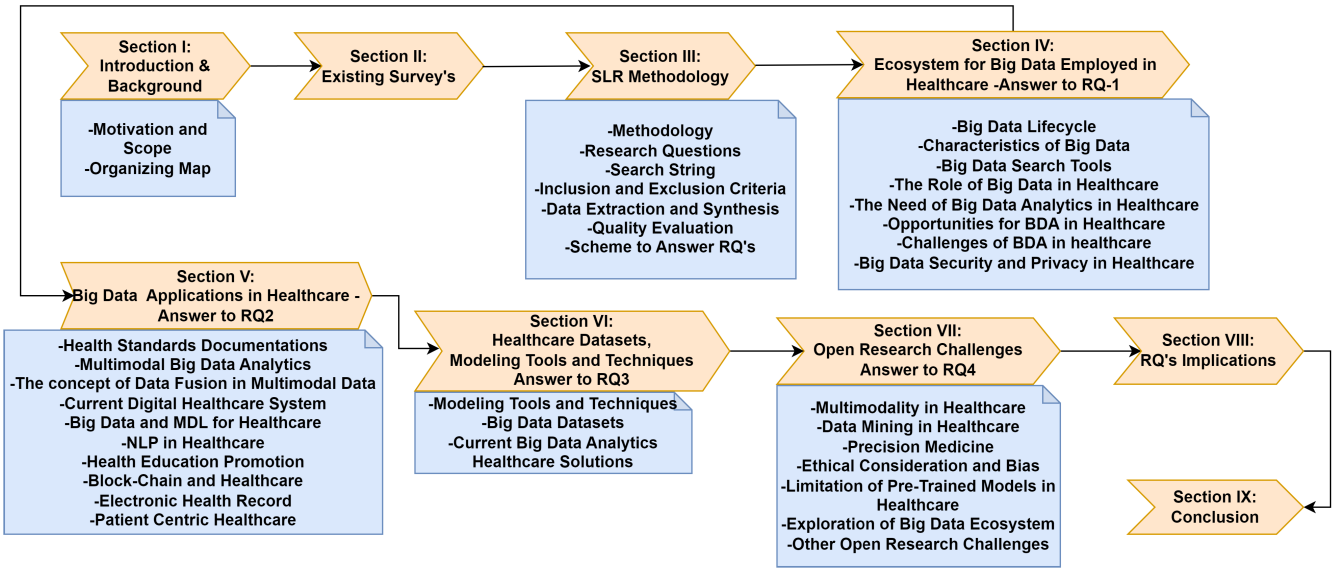

FIGURE 2. Road map with section layout of the review paper.

study conducted by Nambiar et al. [9] primarily centered on the examination and analysis of the challenges and prospects associated with the utilization of big data analytics within the healthcare sector. Further authors discussed big data growth expectations for the year 2015, then statistics shown for spending by geography. Lastly, they enlighten healthcare infrastructure. Following a literature study presented by Raghupathi et al. [10], they presented a comprehensive examination of the key attributes of big data, explored an architectural framework, and elucidated several application possibilities within the healthcare domain. Andreu-Perez et al. [11] performed a systematic literature review (SLR) spanning the years 2008 to 2015. The objective of their study was to offer a thorough examination of advancements in the field of biomedical and health informatics within big data. Luo et al. [12] conducted a comprehensive examination of the recent progress made in the utilization of big data in several healthcare domains. The authors emphasized the substantial expansion observed within the last five years.

Islam et al. [13] had a systematic literature review (SLR) spanning the years 2005 to 2016, with a particular focus on the potential of healthcare analytics through the utilization of data mining and big data. In their study, Bahri et al. [14] directed their attention to examining the many difficulties and possibilities associated with the utilization of big data analytics within the healthcare sector. In their study, Galetsi et al. [15] undertook a systematic literature review (SLR) with the aim of investigating the potential of data-driven methods in enhancing the effectiveness of public health and healthcare organizations. Tandon et al. [16] did a systematic literature review (SLR) examining the utilization of blockchain technology in the healthcare sector. In a similar vein, Imran et al. [17] published a thorough study spanning two decades, offering valuable insights into the application of big data analytics in the field of healthcare. Their work serves as a roadmap for future research and development in this domain.

In their study, Khanra et al. [18] did a systematic literature review (SLR) spanning the years 2013 to 2019. The authors identified and analyzed five distinct viewpoints pertaining to the application of big data analytics in the healthcare domain. Ikegwu et al. [19] conducted a systematic literature review (SLR) on the topic of big data analytics in data-driven industries. Their study aimed to explore the current state of knowledge in this area. In a similar vein, Zhang et al. [20] investigated the primary technologies employed in the rapidly expanding virtual world sector, commonly referred to as the Metaverse. Additionally, they examined the application of big data technology in crucial domains including e-health, transportation, commerce, and finance. After reviewing extensive literature for Big Data Healthcare Analytics for the target period of 2013 - 2023, we came to the conclusion that the current review paper lacks an exact focus on a holistic view of both healthcare and big data. Existing studies either present thoughts on big data or healthcare solely rather than discussing them together. In this review paper, we succinctly summarize the main issue and set the stage for further research Our study offers a more extensive analysis of the current research deficiencies in the domain of big data in healthcare, as compared to the existing surveys. This paper examines the ecosystem of big data in the healthcare sector, focusing on the issues that arise within this context. Additionally, a comprehensive analysis of various uses of big data in healthcare is conducted. Furthermore, a compilation of reliable resources for dataset collecting is presented. Furthermore, we analyze the benefits and applications of big data in the healthcare industry, therefore

TABLE 1. List of acronyms.

Acronyms	Description
ADNI	Alzheimer's Disease Neuroimaging Initiative
AI	Artificial Intelligence
ANSI	American National Standards Institute
BDA	Big Data Analytics
BDCaM	Big Data for Context-Aware Monitoring
BSON	Binary Javascript Object Notation
CCR	Continuity of Care Document
CDA	Clinical Documentation Architecture
CMS	Centers for Medicare Medicaid Services
CQL	Cassandra Structure Language
CRM	Customer Relationship Management
DL	Deep Learning
EMR	Electronic Medical Records
EUS	External Urethral Sphincter
FH	Fathom Health
FHIR	Fast Health Interoperability Resources
GDPR	General Data Protection Regulation
GEO	Gene Expression Omnibus
HCI&A	Healthcare Informatics and Analytics
HCUP	Healthcare Cost and Utilization Project
HDFS	Hadoop Distributed File System
HIPAA	Health Insurance Portability and Accountability Act
ICU	Intensive Care Unit
IoT	Internet of Things
LASA	Life Science Database Archive
LR	Literature Review
LSTM	Long short-term Memory
ML	Machine Learning
MDL	Machine or Deep Learning
MTL	Multitask Learning
NER	Named Entity Recognition
NHS	National Health Service of England
NIST	National Institute of Standards and Technology
PCA	Patient Centric Authorization
PCP	Primary Care Provider
PHI	Personal Health Information
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RIM	Reference Information Model
RNN	Recurrent Neural Network
RQ	Research Question
SLR	Systematic Literature Review
TCGA	The Cancer Genome Atlas
QEQ	Quality Evaluation Question
YOP	Year of Publication

enhancing our comprehensive comprehension of the subject matter.

III. SYSTEMATIC LITERATURE REVIEW - MATERIALS AND METHODS

This section delves into the methodology employed to meticulously craft the overarching framework of our Systematic Literature Review (SLR). The purpose of this section is to present a concise and transparent overview of the methodology and search strategy that was utilized in this review to identify and assess relevant research. The primary aim was to undertake an extensive comparative analysis of previous and ongoing initiatives, critically evaluate the research findings, and uncover any deficiencies or restrictions in current knowledge with regard to addressing the research questions.

A. METHODOLOGY

In line with established best practices, this SLR rigorously follows the methodology outlined by the widely recognized PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [21], [22]. The thoroughly designed methodology of this study unfolds in a systematic manner across three distinct stages: *Input*, *Process*, and *Output*, each intricately interlinked to ensure a comprehensive and transparent approach to our review, as illustrated in Figure 3. For clarity, the sub-steps within each of these stages are further illuminated in this Figure, which follows a three-stage review process inspired by Kitchenham et al. [23]. This structured methodology not only guides our systematic examination of the literature but also enhances the reproducibility and reliability of our findings.

Firstly, we developed our research questions, conducted an initial scoping review of the literature, and consulted with experts in the field. We identified gaps in the literature and areas where further research was needed, which led to the development of our research questions. The research questions were designed to explore the big data ecosystem in healthcare, review the top applications, and provide a comprehensive list of authentic data sources and modelling tools. Furthermore, the motivation behind each individual research question is mentioned in Table 3.

To achieve these objectives, a systematic and rigorous search of prominent academic databases, including Google Scholar, PubMed, Scopus, and Web of Science, was conducted. The search was confined to scholarly articles published within the last ten years, specifically from 2013 to the current year. We searched the search engines with specific keywords related to our research topic to ensure a comprehensive search. The keywords used in the search strategy were carefully selected based on their relevance to the research questions and big data analytics in healthcare. We cross-checked all scientific studies from the original source of publication to ensure accuracy and completeness.

After collecting the initial set of papers, we performed a title and abstract screening to remove irrelevant studies and then conducted a full-text screening to select papers that met our inclusion criteria. We recorded and reported the number of papers screened, assessed for eligibility, and included in the final review. To ensure data accuracy and reliability, we performed a data extraction process using a standardized form that captured essential information from each included paper, such as study design, sample size, data source, and key findings. Lastly, we conducted a quality assessment of the included studies to evaluate the risk of bias and the overall methodological quality of each study. Two reviewers performed the quality assessment independently, and any discrepancies were resolved through discussion and consensus.

B. RESEARCH QUESTIONS

In this subsection, we introduce our research questions (RQs), which seek to investigate a particular facet within the wider context of our study. By constructing a well-defined research inquiry, it is possible to conduct a more comprehensive investigation into the topic at hand, therefore yielding significant

TABLE 2. A comprehensive LR table that reflects work conducted in existing surveys; C is used Criterion followed by numeric numbers as C1, C2, etc. If the particular criterion is present we write Yes else No. All results recorded with respect to big data in healthcare - [Listed papers are sorted to Year of Publication] C1:-Year of Papers Inclusion, C2:-Approach, C3:- Eco System Discussed, C4:-Challenges Privacy and Security, C5:-Applications Discussed, C6:-Data Sources Discussed.

Existing Survey	C1	C2	C3	C4	C5	C6	Main theme of study
[9]	N/A	LR	Yes	No	Yes	No	In this review paper, authors focus challenges and opportunities of Big Data Analytics in Healthcare.
[10]	N/A	LR	Yes	No	Yes	Yes	This study provides an overview of characteristics of big data, then authors discuss architectural framework lastly they describe application perspectives of big data healthcare.
[11]	2008-2015	SLR	Yes	Yes	Yes	No	The purpose of this article is to provide a complete review of current developments (2008-2015) in the field of biomedical and health informatics pertaining to big data.
[12]	2010-2015	SLR	No	Yes	Yes	No	In the past five years, big data applications in healthcare have grown fast, with several discoveries and methods. This paper also reviews recent advances in big data applications in certain healthcare fields.
[13]	2005-2016	SLR	Yes	No	Yes	No	This SLR emphasizes the potential of healthcare analytics using data mining and big data. The study explores the application and theoretical perspectives of healthcare analytics.
[14]	N/A	LR	N	Y	Y	N	In this review paper, authors focus challenges and opportunities of Big Data Analytics in Healthcare.
[15]	N/A	SLR	Yes	No	Yes	No	This SLR study aims to help governments and health policymakers understand how data-driven strategies can improve public health and healthcare organization functioning.
[16]	2015-2018	SLR	No	Yes	Yes	No	This study presents SLR on blockchain applications in the healthcare domain.
[17]	1995-2020	SLR	Yes	Yes	No	No	In this study, authors presented a comprehensive review of two decades. This research presents a road map for BDA insights in healthcare (patient care).
[16]	2015-2018	SLR	Yes	No	Yes	No	In this study, the authors explore various analytical avenues that exist in the context of a patient-centric healthcare system.
[18]	2013-2019	SLR	Yes	No	Yes	No	This study found that BDA can be applied to healthcare from five perspectives: public health awareness, stakeholder interactions, hospital management practices, treatment of specific medical conditions, and technology in healthcare service delivery.
[19]	2013-2021	SLR	No	Yes	No	Yes	This study presents SLR on Big Data Analytics for Data-driven Industry.
[20]	N/A	LR	No	No	Yes	No	This study examined the main technologies utilized in the fast-growing virtual world sector (the Metaverse) and big data technology in critical fields such as e-health, transportation, commerce, and finance.
Our work	2013-2023	SLR	Yes	Yes	Yes	Yes	In comparison to the existing surveys we discuss in detail an ecosystem of big data in terms of healthcare, and we identify the existing research gaps and challenges. We cover scientific studies between 2013 to 2023. We thoroughly review the big data applications in healthcare, and we list the authentic resources for dataset collection. We also discuss the advantages and use cases in our study. Lastly, we uncover a list of potential open research challenges in a detailed manner.

and informative findings. Through the investigation of these research questions, our objective is to make novel contributions as listed in the contribution subsection. These RQs provide a comprehensive examination of the research topic, encompassing the underlying rationales, the methods utilized to address the issue, and the potential ramifications of the results. Our objective is to enhance the body of knowledge and offer significant perspectives to stakeholders in the healthcare industry through a comprehensive response to each research question.

1) RQ-1: What are the components of the big data ecosystem in healthcare, and how do they interact with each other? What are the main challenges and limitations of this ecosystem?
2) RQ-2: What are the most promising application areas of big data analytics in healthcare, and what are some examples of successful implementations in each area? How can these applications improve healthcare outcomes and reduce costs?
3) RQ-3: What are the most reliable and authentic sources of healthcare data, and what are some commonly used modelling tools, techniques, and commercial solutions (presenting use cases) in big data analytics for healthcare?
4) RQ4:- What are the open research challenges reported in the literature in the last three years? Focusing on the solutions for the advancement of Healthcare.

C. SEARCH STRINGS

Our investigation starts by searching the keywords of "big data," "big data analytics," "healthcare," "clinical applications," and "healthcare data." Additionally, we augmented our set of keywords by including the terms "survey," "review," and "literature." In addition, our search query encompassed the terms "multimodal big data," "natural language processing (NLP)," "blockchain," "security," "privacy," and "electronic health records (EHR)." The utilization of logical operators, such as "AND" or "OR", was employed in conjunction with search strings as necessary. We searched on Scopus as it is considered as a standard for retrieving searches from credible databases.

The research question for RQ-1 seeks to investigate the big data ecosystem within the healthcare domain and its interrelationships, including the challenges and constraints inherent in this ecosystem. The search query encompasses

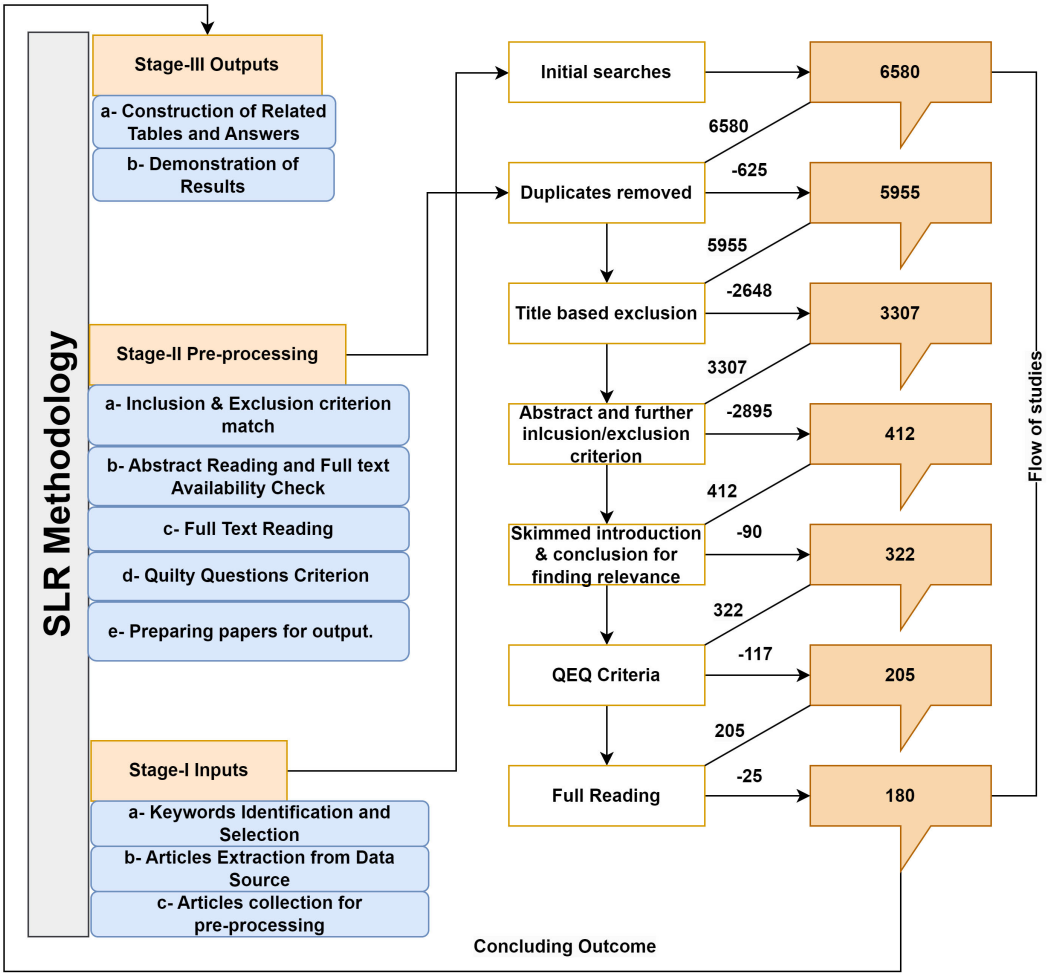

FIGURE 3. SLR Methodology and inclusion/exclusion process for selecting relevant paper in the study.

terms such as "big data ecosystem" and "healthcare ecosystem," in conjunction with keywords such as "components," "challenges," and "limitations." Furthermore, it encompasses pertinent topics such as the "life cycle of big data," "search tools for big data," and "characteristics of big data," as well as the "opportunities" and "challenges" associated with big data in the healthcare domain. This search query assists in identifying scholarly publications and research projects that explore different facets of the big data ecosystem in the healthcare industry and provide insights into the issues and constraints encountered by the existing system. The objective of the search query for RQ-2 was to ascertain the most advantageous domains for the utilization of big data analytics within the healthcare sector. This involves finding use cases where such analytics have been effectively implemented, evaluating the resulting achievements within each domain, and assessing the prospective advantages in terms of healthcare outcomes and cost savings. The inquiry integrates key phrases such as "promising application areas in healthcare," "utilization of big data analytics in the healthcare sector," and "healthcare implementations." These are combined with terms such as "healthcare outcomes" and "reduction of costs for healthcare." Moreover, this encompasses distinct domains of application such as "standardized documentation in healthcare," "analysis of multimodal data," "implementation of digital systems in healthcare," "utilization of blockchain technology in healthcare," "advancement of healthcare education," within the framework of big data in the healthcare sector. This search query aids in identifying pertinent scholarly publications.

The primary focus for the search query of RQ-3 is to find credible sources of healthcare data, prevalent modelling tools, approaches, and commercial solutions utilized in big data analytics for healthcare. Additionally, it seeks to explore the many factors associated with data governance and ethics in this domain. The search encompasses concepts such as "trustworthy data sources," "credible data sources," "tools for modelling," "techniques for modelling," "commercially available solutions," "analytics for large datasets," "management of data," and "ethical implications." By

TABLE 3. Research questions for this review, their motivations, and the scheme to get their answers.

Research Question	Motivation	Scheme to answer the Question
RQ-1	The motivation behind this research question is to overview the big data ecosystem in healthcare, their interconnections, and the challenges and constraints they pose is manifold. Despite the growing importance of big data in healthcare, it became evident that there was a dearth of comprehensive studies focusing on the entire ecosystem as a whole. In this study, we aim to cover the knowledge gap that we identified in the existing literature pertaining to the healthcare ecosystem. The objective of this research question is about how various components within the healthcare data ecosystem operate as a fundamental aspect of harnessing the full potential of healthcare data. Moreover, understanding the interactions between different components of the healthcare ecosystem is essential for optimizing data utilization, improving patient care, and streamlining healthcare processes to make data-driven decisions and leverage the vast amount of healthcare data effectively. Additionally, this research question aims to identify the challenges such as data privacy concerns, security vulnerabilities, data interoperability issues, and the need for maintaining data quality currently faced by the healthcare data ecosystem. By shedding light on these challenges, we can pave the way for the development of targeted solutions and strategies to address them, ultimately contributing to the enhancement of healthcare data management and, consequently, healthcare outcomes. In summary, this research question is driven by the desire to fill the knowledge gap.	We present a short comparison of Web2.0 and Web3.0 with reference to Big Data healthcare; then review the Big Data life cycle; further, we highlight the Big Data search tools; then Big Data characteristics; then opportunities, challenges and lastly briefly discuss Big Data in terms of healthcare. This RQ is answered in the corresponding section IV.
RQ-2	There are several strong reasons for investigating the most promising big data analytics applications in healthcare, identified from the existing literature body. First, big data analytics has the potential to bring a widely acceptable transformation for healthcare, as is becoming clearer throughout the healthcare sector and academia. The scope and variety of this potential, however, are largely unexplored. The goal of this research is to shed light on the most promising fields where big data analytics may produce meaningful advantages. Second, the strong desire to enhance healthcare outcomes through data-driven insights as big data analytics has the potential to improve patient care, treatment plans, and overall healthcare quality and to notice improvements by exploring successful implementations in a variety of application domains. Then we aim to find examples from the real world where big data analytics has given healthcare workers the knowledge and authority to make decisions that improve patient outcomes. Then, this study intends to promote innovation and broad technological adoption in the healthcare industry. Highlighting effective implementations can encourage healthcare companies to adopt data analytics and start their own revolutionary journeys. We list areas that have gained attention due to their potential to improve outcomes and reduce costs.	This question is being answered in section V with respective subsections starting from Healthcare standard documentation, multimodal analytics then the literature of data fusion; further presented current digital healthcare systems; additionally, we presented a block-chain, healthcare education perspectives, and lastly, we briefly added literature for [AI, ML, DL, NLP] with reference to Big Data healthcare.
RQ-3	The motivation behind this particular question is to find the most reliable and authentic sources of healthcare data, as well as commonly used modeling tools, techniques, and commercial solutions in big data analytics for healthcare to address the knowledge gap. In the evolving landscape of healthcare, where data-driven decisions can have life-changing implications, it is imperative to ensure that the data upon which these decisions are based should be authentic. By conducting a comprehensive study of trustworthy healthcare data sources, we aim to provide a valuable resource for researchers, healthcare providers, and policymakers seeking to access data that can be confidently relied upon. Furthermore, the choice of modeling tools and techniques is pivotal in the realm of healthcare analytics. Our desire is to offer guidance and insights into the most commonly used and effective tools and techniques in this domain. This knowledge can empower researchers and practitioners to employ advanced analytics methodologies that yield meaningful insights and drive innovation in healthcare practices. Additionally, healthcare organizations are increasingly turning to commercial solutions to streamline their data analytics processes. Understanding the landscape of available commercial solutions, along with their reliability and efficacy, is essential for making informed decisions about technology investments. This research is motivated by the goal of providing a robust evaluation of these solutions, enabling healthcare entities to choose the best-fit tools for their specific needs.	This RQ is answered in the corresponding section VI with a brief discussion and summarized tables. Initially, this question is answered with a wide variety of modeling tools, and then a detailed overview of datasets covering broader areas of healthcare public health datasets to specific disease datasets are discovered after extensive literature. At last, this question is backed by an overview of current big data analytics solutions in healthcare.
RQ4	The primary objective of this research question is to understand the constantly changing nature of big data which leaves open research areas in big data applications in healthcare, particularly with an emphasis on solutions that improve healthcare pertaining towards patient-centric application. An expansion on integration of big data and healthcare has improved patient care, streamlined operations, and lower costs and also has transformed the healthcare business over the past decade. The integration of both areas has raised questions about patient data privacy and security. This difficulty must be addressed to retain patient trust and ethical data processing. Data interoperability concerns continue to emphasize the need for standardized solutions that enable smooth data exchange across healthcare systems and devices for more holistic patient care. Patient privacy, security, computational scalability, efficiency, etc. are the most common issues being addressed since the inception of the idea of integration. In this study, we aim to discuss open problems and potential future research directions specific to the harnessing of big data applications in healthcare for the community. Further, we focus on presenting advanced topics that need to be addressed in the near future such as the open challenges ranging from multimodality, data mining, precision medicine, and pre-trained models for healthcare. These open research challenges aim to advance healthcare through big data analytics. Researchers can revolutionize healthcare for people, communities, and systems globally by tackling these concerns and investigating solutions.	Healthcare requires patient engagement and behavior change, and using big data to improve them is appealing. Researchers must develop new methods to engage patients in their care. To answer this RQ we prepared the most important key open challenges identified from a vast area of future dimensions and we provide a key guide for open research challenges that will serve as a base in the future for upcoming research. We answer this RQ in section VII with respective subsections.

integrating the specified keywords, the search results include scholarly articles and research studies that examine reliable sources of healthcare data, commonly used modelling tools and techniques, commercially available solutions in the field of big data analytics, as well as strategies to tackle data governance and ethical concerns within the healthcare domain. The search query for RQ-4 primarily aims to uncover open research difficulties that have been documented in the literature during the past years. Additionally, it seeks to investigate prospective solutions that might contribute to the advancement of healthcare. The search encompasses key terms such as ''big data healthcare open research challenges,'' ''big data healthcare future research directions,'' and ''advancement of healthcare.'' Furthermore, it incorporates potential avenues for future research within the realm of big data healthcare.

The word cloud, as illustrated in Figure 4, was utilized to condense and visually represent the predominant search phrases employed in this study for the purpose of identifying pertinent research. Therefore, it can be inferred that our search strings exhibit a strong correlation with the specific research inquiries and the studies that have been chosen.

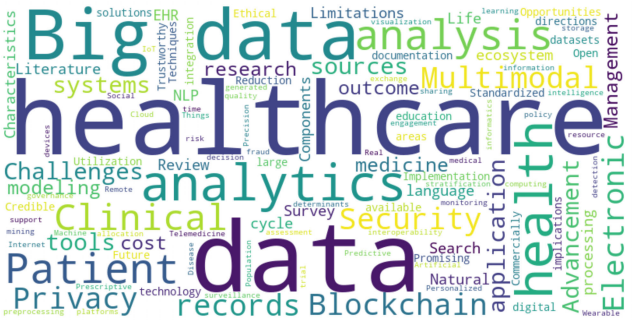

FIGURE 4. Word cloud of search terms used in SLR.

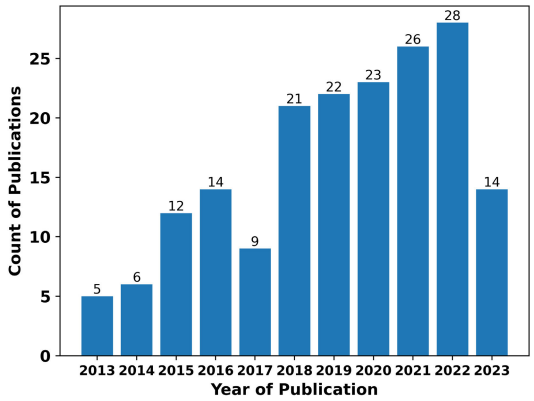

FIGURE 5. Classification of publications by year.

D. INCLUSION AND EXCLUSION CRITERIA

In the early stage of our research, we conducted a comprehensive search using specific keywords, which resulted in thousands of records. To ensure the relevance and quality of the studies that we analyzed, we implemented a detailed methodology that is depicted in Figure 3. Our methodology included several inclusion and exclusion criteria, such as selecting studies that were published within a specific time frame, written in the English language, and fully available to readers. All other papers that did not meet these criteria were excluded from our analysis. By following this rigorous methodology, we were able to focus on a specific subset of studies that met our pre-defined standards for inclusion in our research.

E. DATA EXTRACTION AND SYNTHESIS

The present study involved a comprehensive review of the selected research papers listed in the bibliography. We followed a systematic approach to extract and synthesize the relevant data, based on the key attributes specified in Table 4. These attributes included paper ID, study title, author names, publication date, open database access, publication source, research context, document type, the topic addressed, and citation count. We recorded the data in an Excel sheet and synthesized it in a way that enabled us to effectively manage and evaluate the research data.

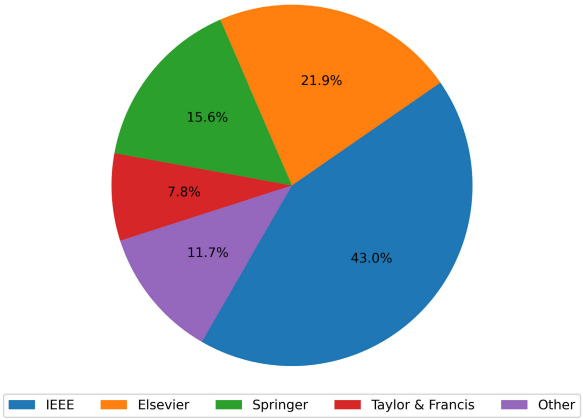

FIGURE 6. Distribution of publication by the publisher.

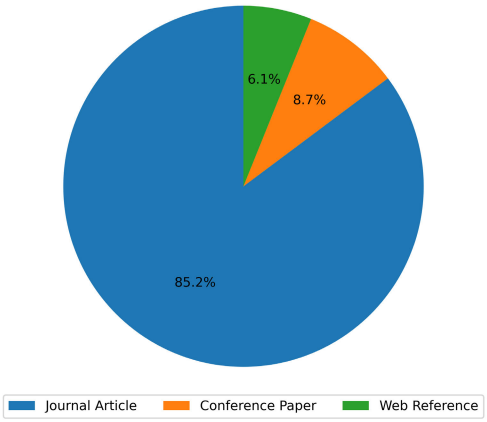

FIGURE 7. Distribution of papers by reference type.

To draw relevant conclusions from the data, we performed various analyses to extract insights such as distribution of publications by year, publisher, reference type, and research study type. Firstly, we present a classification of publications by year, as depicted in Figure. 5; the analysis of this chart shows the increasing pattern of selected studies for this survey. This also shows the diversification of publication years. Secondly, we examined the distribution of publications by their publisher, as shown in Figure. 6; this chart shows the diversification of our study selection process is strictly aligned with the inclusion and exclusion criterion as presented in Figure. 3; this also shows that we tried to represent the presence of all major publishers. Further, we present the distribution of papers by reference type, as shown in Figure. 7; the objective of this chart is to show the diversified inclusion of reference type (Journal Article, Conference Paper, etc.) included in this survey; this chart also highlights the importance of journal article category in our particular manuscript domain. Last but not least, we presented an overview of the distribution of study type, as shown in Figure. 8; this chart is a further addition to the previous chart. The analysis of this chart shows the growing interest in article types published within the scope of this

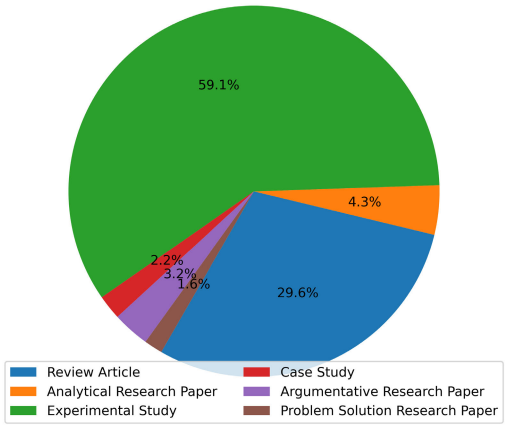

FIGURE 8. Distribution of references by research study type.

TABLE 4. Data extraction key attributes.

Attribute	Description
Study_ID	Uniquely identification of paper
Study_Title	Title of the study paper
Author	Author of study
Date	Date of publication of the paper
Database	Database provider
Publication source	Who is the publisher of the selected study paper?
The topic addressed	Is this study addressing the topic of study?
Citation count	Citation count of the particular study

manuscript the experimental papers and the review articles such as our survey. These analyses helped us to identify trends, patterns, and insights that are relevant to our research questions.

F. QUALITY EVALUATION QUESTION

In this subsection, we present four quality questions that we applied to evaluate the final selected papers with Quality Evaluation of Evidence (QEQ) scores.

- Does the article explicitly discuss the big data and data analysis methods used?
- Does the article discuss advantages and challenges related to the topic?
- Does the article present or discuss potential applications of the topic in healthcare?
- Are the outcomes presented in the article valid and aligned with the utilized methodology and topic of interest?

Table 5 summarizes the paper quality based on the designed QEQ. The score is calculated on the basis that: If the answer to a particular question turns out to be true, Y is counted and each Y is mapped to 2.5 points collectively all four Y are mapped to 10 points as the formula defined in Eq.1. If any of the selected studies scored between 6 to 10, we counted that particular study as the most relevant study, and the class category is set to be ''High''. For a paper to be in a ''Medium'' class, the particular paper must score between 3 to 5, and for a paper to be in a ''low-class'' the score should be at least 2.5. In our analyses, very few studies fell in the low category because of rigorous methodology; all papers had passed through our inclusion and exclusion criteria. All the mentioned reference studies are used in our manuscript except those studies that fall below ''low-class'' papers from our QEQ table. For example, a study that does not pass any of the quality questions is categorized as extremely low and directly excluded from consideration such studies are not recorded in this table.

$$TotalScore\_Yes = \left(\sum_{i=1}^{4} Y_{Qi}\right) \times 2.5 \tag{1}$$

where Y_Qi represents the sum of ''Yes'' answers (ranging from 1 to 4) when all responses are recorded as Yes.

$$TotalScore\_Other = 10 - \left(\sum_{i=1}^{4} N_{Qi}\right) \times 2.5 \tag{2}$$

where N_Qi represents the sum of ''No'' answers (ranging from 1 to 4) when recorded answers are a mix of Yes and No.

This QEQ table list is recommended to be considered as a subset of the full QEQ study. We applied a QEQ filter to 256 numbers of studies as mentioned in Figure. 3, while here we maintain a subset to contribute to effective SLR methodology and avoid multi-page tables to maintain simplicity and reduce the length of the manuscript. While our designed QEQ questions provide a useful framework for evaluating the quality of the papers, it's important to keep in mind that they are just one aspect of our comprehensive evaluation process. So it is suggested that other factors may influence the quality of the papers including but not limited to the study design, sample size, statistical analysis, potential biases, and the overall relevance and contribution of the research to the field.

G. SCHEME TO ANSWER RESEARCH QUESTIONS

This study aims to provide answers to the research questions developed based on an extensive review of existing surveys. Table 3 lists the proposed research questions, their motivation, and the direction of the answers provided in the study. Additionally, a research question-answer mapping Table 6 is included, which provides an overview of the references used for specific answers. Furthermore, the study includes a research gap Table 2, which compares various parameters used in the review with existing surveys. This approach provides a comprehensive analysis of the research questions and the existing literature, contributing to a better understanding of the research gap and providing insights for future research.

IV. ECOSYSTEM FOR BIG DATA EMPLOYED IN HEALTHCARE - ANSWER TO RQ-1

The big data ecosystem/platforms can potentially improve the applicability of clinical research studies in real-world scenarios, which traditionally have been hindered by the diversity of the populations being studied. In addition,

TABLE 5. Quality Evaluation Question Score - [Total Score is the sum of all Y and all N use equations 1 and 2 for calculation - High Class = 6 to 10, Medium Class = 3 to 5 and Low Class = 2.5] - Citation count is maintained at the time of developing this particular table (27th April 2023) - [YOP - Year of Publication, CC - Citation Count, Q -QEQ].

Study &YOP	CC	Q1	Q2	Q3	Q4	Total Score	Class	Study &YOP	CC	Q1	Q2	Q3	Q4	Total Score	Class
[24] & 2013	514	Y	Y	Y	Y	10	High	[9] & 2013	267	Y	Y	Y	Y	10	High
[25] & 2013	182	N	Y	Y	N	5	Medium	[20] & 2023	0	Y	N	N	N	2.5	Low
[26] & 2014	59	Y	Y	Y	N	7.5	High	[27] & 2014	419	Y	Y	Y	N	7.5	High
[10] & 2014	3271	Y	Y	N	Y	7.5	High	[28] & 2014	154	Y	Y	Y	N	7.5	High
[11] & 2015	731	Y	Y	Y	Y	10	High	[29] & 2015	588	Y	N	Y	N	5	Medium
[30] & 2015	62	Y	Y	Y	N	7.5	High	[31] & 2015	145	Y	Y	Y	Y	10	High
[32] & 2015	1038	Y	Y	Y	Y	10	High	[33] & 2015	802	Y	Y	Y	N	7.5	High
[34] & 2015	26	N	N	Y	N	2.5	Low	[35] & 2015	107	Y	Y	N	N	5	Medium
[36] & 2015	51	Y	Y	Y	N	7.5	High	[37] & 2015	112	Y	Y	Y	N	7.5	High
[38] & 2016	894	Y	N	N	Y	5	Medium	[39] & 2016	251	Y	Y	Y	N	7.5	High
[40] & 2016	833	Y	Y	Y	Y	10	High	[41] & 2016	358	Y	Y	Y	Y	10	High
[42] & 2016	5339	Y	Y	N	N	5	Medium	[12] & 2016	506	Y	Y	Y	Y	10	High
[43] & 2016	35	Y	Y	N	N	2.5	Low	[44] & 2016	137	Y	Y	N	N	5	Medium
[45] & 2016	42	Y	Y	N	Y	7.5	High	[46] & 2016	127	N	Y	Y	N	5	Medium
[47] & 2017	227	Y	Y	N	N	5	Medium	[48] & 2017	11	Y	Y	N	N	5	Medium
[49] & 2017	1010	Y	N	Y	Y	7.5	High	[50] & 2017	202	Y	N	Y	Y	7.5	High
[51] & 2017	53	Y	N	N	Y	10	Medium	[52] & 2017	67	Y	Y	Y	N	7.5	High
[53] & 2017	317	Y	Y	Y	Y	10	High	[54] & 2018	438	Y	Y	N	N	5	Medium
[55] & 2017	160	Y	Y	Y	Y	10	High	[56] & 2018	2032	Y	N	Y	N	5	Medium
[57] & 2018	23	Y	Y	N	N	7	Medium	[58] & 2018	24	Y	Y	N	N	5	Medium
[59] &2018	9	N	Y	N	N	2.5	Low	[13] & 2018	175	Y	Y	Y	Y	10	High
[60] & 2018	180	Y	Y	Y	Y	10	High	[61] & 2018	40	Y	Y	N	N	5	Medium
[62] & 2018	246	Y	Y	Y	N	7.5	High	[63] & 2018	24	Y	N	Y	N	5	Medium
[64] & 2018	77	Y	Y	N	N	5	Medium	[65] & 2018	21	Y	Y	N	Y	2.5	Low
[66] & 2018	225	Y	Y	Y	N	7.5	High	[67] & 2018	46	Y	Y	N	N	5	Medium
[68] & 2018	47	Y	Y	N	N	5	Medium	[69] & 2018	26	Y	Y	Y	Y	10	High
[70] & 2018	417	Y	Y	N	N	5	Medium	[71] & 2018	1318	Y	Y	Y	Y	10	High
[72] &2019	56	Y	Y	N	N	5	Medium	[73] & 2019	163	Y	Y	Y	N	7.5	High
[74] & 2019	769	Y	Y	Y	Y	10	High	[75] &2019	685	Y	Y	N	N	5	Medium
[15] & 2019	89	Y	Y	Y	Y	10	High	[76] & 2019	66	Y	Y	N	N	5	Medium
[77] & 2019	22	Y	N	Y	N	5	Medium	[78] &2019	204	Y	Y	Y	Y	10	High
[79] & 2019	236	Y	Y	Y	N	7.5	High	[80] & 2019	65	Y	Y	N	N	5	Medium
[81] & 2019	106	Y	Y	Y	N	7.5	High	[82] & 2019	39	Y	Y	Y	Y	10	High
[7] & 2020	118	Y	Y	Y	N	7.5	High	[83] & 2020	26	Y	N	Y	N	5	Medium
[84] & 2020	99	Y	N	Y	N	5	Medium	[85] & 2020	7	Y	N	N	N	2.5	Low
[86] & 2020	191	Y	Y	Y	N	10	High	[18] & 2020	113	Y	Y	Y	Y	10	High
[87] & 2020	39	Y	Y	Y	Y	10	High	[17] & 2020	26	Y	Y	Y	N	7.5	High
[88] & 2020	5	Y	Y	Y	N	7.5	High	[89] & 2021	11	Y	N	N	N	2.5	Low
[90] & 2021	2	Y	N	N	N	2.5	Low	[91] & 2021	64	Y	N	N	N	2.5	Low
[92] & 2021	3	Y	Y	N	N	5	Medium	[93] & 2021	206	Y	N	N	Y	5	Medium
[94] & 2021	111	Y	Y	Y	N	7.5	High	[95] & 2021	69	Y	Y	Y	N	7.5	High
[96] & 2021	12	Y	N	Y	N	5	Medium	[97] &2022	11	Y	Y	Y	Y	10	High
[98] & 2022	668	Y	Y	N	N	5	Medium	[16] & 2022	236	Y	N	Y	N	5	Medium
[99] & 2022	826	Y	N	Y	N	5	Medium	[100] & 2022	2	Y	N	N	Y	5	Medium
[101] & 2022	65	Y	Y	Y	N	7.5	High	[102] & 2022	826	Y	N	Y	N	5	Medium
[103] & 2022	0	Y	N	N	N	2.5	Low	[104] & 2022	1	Y	N	N	N	2.5	Low
[105] & 2022	0	Y	Y	Y	N	7.5	High	[19] & 2022	6	Y	Y	Y	Y	10	High
[17] & 2022	26	Y	Y	Y	N	7.5	High	[106] & 2022	2	Y	Y	Y	N	7.5	High
[107] & 2022	43	Y	Y	Y	N	7.5	High	[108] & 2022	85	Y	Y	Y	N	7.5	High
[109] & 2022	826	Y	N	Y	N	5	Medium	[16] & 2022	236	Y	N	Y	N	5	Medium
[110] & 2022	1	Y	Y	N	N	5	Medium	[111] & 2022	5	Y	N	Y	N	5	Medium
[112] &2022	21	Y	N	Y	N	5	Medium	[113] & 2022	2	Y	Y	Y	N	7.5	High
[102] & 2022	826	Y	N	Y	N	5	Medium	[114] & 2023	0	Y	Y	N	N	5	Medium
[115] & 2023	2	Y	Y	Y	Y	10	High	[116] & 2023	31	Y	N	Y	N	5	Medium

it provides an opportunity to perform patient stratification, which is necessary for successful and precise medical treatment [50].

The deployment of an ecosystem in the healthcare sector has a significant and transformative history, commonly referred to as Healthcare Informatics and Analytics (HCI&A)

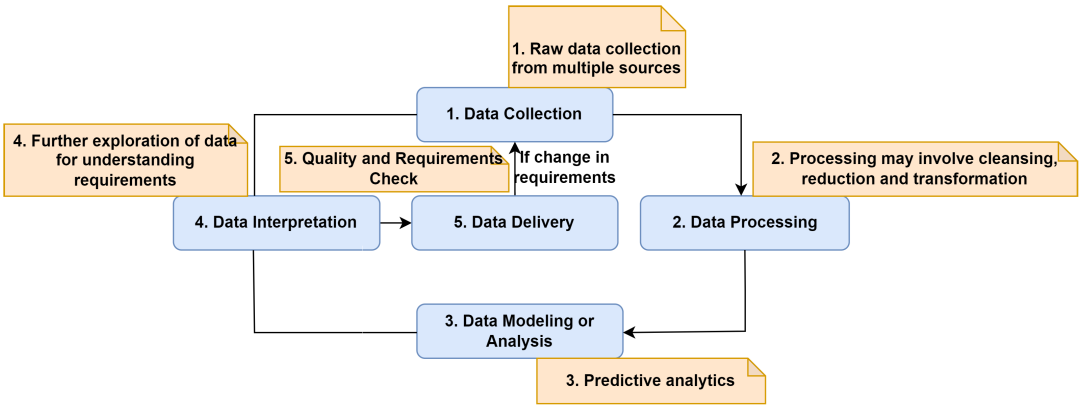

FIGURE 9. An illustration of a typical BDA healthcare life-cycle.

TABLE 6. Table of reference map for research questions.

Research Question	Referenced Articles
RQ1	[13]–[15], [18], [19], [25], [32], [34], [35], [37]–[39], [41], [43]–[48], [50], [51], [54], [55], [57]–[61], [64], [65], [68], [68], [71], [76], [78], [80], [87], [94], [101], [107], [108], [114], [115], [117]–[133]
RQ2	[7], [16], [18], [24], [26]–[28], [30], [33], [36], [38], [42], [49], [56], [62], [63], [66], [67], [69], [70], [73], [75], [77], [79], [83]–[86], [88]–[97], [99], [100], [102]–[106], [109], [111]–[113], [116], [134]–[147]
RQ3	[27], [35], [40], [52], [53], [63], [74], [77], [81], [82], [89], [98], [101], [105], [110], [148]–[159]
RQ4	[96], [107], [138], [144], [160]–[184]

[87]. HCI&A encompasses the integration of information technology and data analytics in healthcare to improve patient care, decision-making, and operational efficiency. HCI&A has evolved over time, with distinct stages known as HCI&A 1.0, HCI&A 2.0, and HCI&A 3.0. Each stage represents advancements in technology and approaches to healthcare informatics. HCI&A 1.0 focused on the implementation of electronic health records (EHRs) and basic data analysis. HCI&A 2.0 introduced more sophisticated analytics techniques, such as predictive modeling and machine learning, to extract insights from large healthcare datasets. HCI&A 3.0 aims to leverage emerging technologies, such as artificial intelligence and the Internet of Things, to enable more personalized and proactive healthcare [185].

These revolutions in HCI&A are closely intertwined with the evolution of the World Wide Web. The utilization of Web 2.0, characterized by user-generated content and social media platforms, has played a significant role in facilitating collaboration, information sharing, and patient engagement in healthcare. However, as technology continues to advance, the healthcare industry is transitioning toward the development and optimization of Web 3.0. Web 3.0, also known as the Semantic Web, emphasizes the use of linked data and semantic technologies to enhance interoperability, knowledge representation, and intelligent decision support in healthcare.

In this section, we discuss the big data ecosystem that has been deployed in healthcare. To create our proposed ecosystem, we have adopted the most relevant aspects from the literature and have developed an optimized version of these characteristics, which is depicted in Figure 1. The ecosystem comprises various components that are responsible for the efficient processing and management of big data in healthcare. By implementing this optimized big data ecosystem, healthcare organizations can improve their operations and enhance patient care.

A. HEALTHCARE - BIG DATA LIFE-CYCLE

The term life cycle itself portrays the visual picture of stages involved in a particular domain as depicted in Figure 9. Like all other domains, healthcare typically involves the following stages, typically considered in the healthcare data life-cycle; these steps may change as per requirements. The life-cycle [13], [15], [18], [19], [39], [50], [78], [115], [128] usually starts with the data collection stage, followed by processing to transform and clean the data. The data analysis then involves applying such as statistical or machine learning techniques to identify patterns or insights in the data. The next step, data interpretation, involves drawing conclusions and making decisions based on the analysis. Finally, the data delivery stage involves presenting the findings to the end-users or stakeholders in a useful and actionable format.

- **Data Collection:-** This stage involves collecting various (multimodal) data types, including patient data, medical records, laboratory data, and other information from relevant sources.
- **Data Storage:-** Once the data has been collected, it needs to be stored in a way that is easily accessible and searchable. This may involve the use of cloud storage or other forms of secure data storage.
- **Data Processing:-** The data processing involves using various tools and techniques to analyze the data and extract useful insights from it. Further, the common method includes machine learning algorithms, statistical analysis, or other advanced ways of data processing.

- **Data Analysis:-** Once the data has been processed, it can be analyzed to identify meaningful trends and patterns present in the data. This may further apply data visualization tools to help make sense of the data and identify important insights; various tools are listed in Table 12.
- **Data Interpretation:-** The last but not least stage of the big data life-cycle in healthcare is data interpretation. This involves taking the insights gained from the data analysis and using them to make informed decisions about patient care, treatment plans, and other healthcare decisions.

The big data life-cycle in healthcare is a complex process that requires careful planning and execution to ensure that the data is collected, stored, handled securely processed effectively, and analyzed in a way that provides meaningful insights and improves patient outcomes [35].

B. CHARACTERISTICS OF BIG DATA/HEALTH RECORDS

The term ''big data'' has gained widespread usage in recent years, as the volume, velocity, and variety of data being generated and collected has increased exponentially. This has led to a need for advanced analytical methods that can extract valuable insights from this data. Big Data Analytics (BDA) is a collection of techniques and technologies designed to address these challenges, including machine learning, data mining, and predictive analytics. These tools enable researchers and analysts to identify patterns and relationships in large and complex datasets and to use this information to make better decisions and predictions.

Defining big data precisely can be challenging, as it often depends on the perspective of domain experts. Rather than a specific definition, experts typically look at the ''V's'' of data, which represent the characteristics of data. The number of V's has increased over time, with different experts defining them in their own way. In the literature, the number of V's associated with big data has expanded from the fundamental 3-V's to 5-V's, 7-V's, 40-V's, and even 51-V's, as reported in scientific material [44], [76], [108].

In the context of healthcare, the literature commonly refers to six V's,

- Volume: Refers to the size of data, indicating the vast amount of information generated and collected in healthcare.
- Variety: Represents the complexity of data, encompassing different data types and formats, such as structured, unstructured, and semi-structured data.
- Velocity: Signifies the speed at which data is generated, transmitted, and processed, highlighting the real-time nature of healthcare data.
- Veracity: Refers to the quality and reliability of data, considering factors such as accuracy, consistency, and trustworthiness.
- Value: Represents the knowledge and insights that can be derived from data, emphasizing the potential benefits and actionable information that data analysis can provide in healthcare.
- Variability: Reflects the variability and dynamic nature of data, accounting for fluctuations and changes in data patterns and characteristics over time.

These six V's provide a framework for understanding the key dimensions of big data in the healthcare domain, enabling researchers and practitioners to manage and analyze data effectively.

C. BIG DATA SEARCH TOOLS

In the context of Big Data, traditional search engines cannot handle the vast amount of unstructured data. Thus, Big Data Search Tools are developed to enable efficient and effective searching and analysis of massive datasets. During literature work we found one of Github repository [118] that lists the two sources as Big Data Search Tools such as i) Apache Luncene and ii) Apache Solr as summarized in Table 7. Both of the tools are different in terms of their key applications. Solr has gained more popularity among developers and researchers due to its ease of use, scalability, and ability to handle large amounts of unstructured data. Solr has also been widely adopted in the industry, with companies such as Netflix, eBay, and Instagram using it for their search applications [117], [118]. Our review suggests that these tools can be vital in managing and searching Big Data in various domains, including healthcare.

D. THE ROLE OF BIG DATA IN HEALTHCARE

Enabling EHRs opens up exciting prospects for enhancing clinical decision-making infrastructure. There are, nevertheless, significant challenges to overcome. In this section, we discuss the four major components of big data in terms of healthcare data: volume (size), variety (diversity), variability (temporal resolution), and value (quality). Then we discuss the sequential flow of descriptive, predictive, and prescriptive analytics, and how each plays a key role in clinical decision-making.

Currently, conventional primary healthcare models are dependent on various disconnected systems and information sources, which are obsolete now. The new digital healthcare paradigm will move toward an inherent capability to ensure information is exchanged between systems in a way that is both seamless and secure. EHRs are massively heterogeneous and multimodal by nature. Clinical data must be preserved at all costs, as this is a key premise underlying all medical information management systems. Further, it is not only to preserve the data's originality but to keep it as secure, private, and de-identified as possible. Big Data is characterized by its size, speed, diversity, accuracy, and significance. Big Data in terms of healthcare is really big. In 2013 it was predicted globally that the healthcare data to be produced 153 exabytes; But in 2020, 2314 exabytes of data were reported by [123]. Moreover, the 11000% change in data volume is also reported by [123]. It can be concluded that the amount of information

TABLE 7. A comprehensive list of open source big data healthcare tools categorized by type.

Reference	Type	Tool Name	Key Application
[34], [48], [119], [120]	Collection	Flume and Sqoop	Flume is an open-source data collection and processing system designed to efficiently collect, aggregate, and move large amounts of log data from multiple sources to a centralized data store, such as Hadoop Distributed File System (HDFS), HBase, or Amazon S3. Sqoop simplifies the process of transferring data between Hadoop and external data storage systems
[121]	Storage	Hadoop	Hadoop is an open-source framework that is used for storing and processing big data in a distributed manner.
[114]	Processing	Hive and Oozie	Apache Hive provides a high-level language for querying and managing data that is stored in Hadoop clusters Apache Oozie is a workflow scheduler system for managing Apache Hadoop jobs
[117], [122]	Search	Solr and Apache LUCENE	Solr is a multilingual purpose search engine, it is used for Indexing and Searching LUCENE specifically used for Image Retrieval
[186]	Analytics	R and Mahout	R and Mahout are two popular tools for analytics in the Hadoop ecosystem
[187]	Visualization	Node.js	Web service for visualization using JavaScript
[188]	Streaming Processing	Storm, S4, Spark	Real-time data streaming
[189]	Management	Zookeeper and Ambari	Zookeeper is a distributed coordination service for managing large distributed systems Ambari is a web-based management tool for Hadoop clusters.

has doubled every year. Healthcare is a real big data sector [59], [124]. It is reported that 30% of the stored world data is health sector data [59].

The velocity indicates how rapidly information is being generated, stored, or transmitted. Every year, a patient receives approximately 80 megabytes of data in EHR [124]. By 2025 the growth of healthcare sector data will be around 36% of the global data [125], [126].

The value or quality of the data is determined by how well it can be used to generate and evaluate hypotheses. It's also important to know if the provided or collected data can help to predict what will happen in the future. If so, we can act early to make things better. Viability [107] is also a quality dimension that shows whether the data are useful for the use case. Because of the data, data mining and artificial intelligence, and all other sub-techniques including but not limited to machine learning, deep learning, and natural language processing can be effectively applied, allowing us to understand more about clinical decision-making systems.

Big data in terms of healthcare is a conceptual framework of artificial intelligence as a path through *descriptive, diagnostic, predictive, and prescriptive analytics.* Understanding historical data is the goal of descriptive analytics, which employ methods such as data aggregation, data mining, and user-friendly visualizations to get there. Reports that answer questions such as "How many patients were admitted to a hospital last year?" are typical examples of descriptive analytics. Within the last 30 days, how many patients did not survive? Or, how many people become infected while being treated? Descriptive analytics provides simple methods for summarizing data with histograms and graphs to display the attributes of data distributions. The connecting of datasets is typically necessary to acquire substantial insight and understanding for the purpose of optimizing healthcare delivery while reducing costs. To rephrase, it is preferable to combine facts from various sources. Simply, this means coordinating efforts throughout a hospital to share patient data. In so far as it is based on a single moment in time in the past, descriptive analytics is restricted in its potential to inform decision-making. It is helpful, but it may not be predictive of everything that happens.

On the other hand, diagnostic analytics aims to determine the reason behind a phenomenon by analyzing the collected data. Diagnostic analytics could include correlation techniques that find links between clinical variables, treatments, and drugs. While predictive analytics allows us to determine what will happen and how likely it will happen, we might want to know, for example, how likely a patient is to die, how long they will be in the hospital, or how likely they are to get an infection. Predictive analytics uses the data's past values to give useful information about important events that will happen in the future. Predictive analytics are in trend/demand because healthcare professionals believe in evidence-based systems to predict and avoid adverse effects. In addition, predictive analytics facilitate early detection, saving lives and improving patients' quality of life. Lastly, prescriptive analytics optimize decisions. They use all available information to make the best action decision. Predictive analytics help us evaluate clinical interventions and examine the system's usefulness. Furthermore, prescriptive analytics predicts what will happen and the reasoning behind why it will happen. Prescriptive analytics helps turn a prediction model into a decision model.

The availability of healthcare large data offers several benefits but also poses several significant challenges. The first of these is interoperability and then privacy and security. With such a diversified healthcare system, which comprises continuous data sources and stakeholders such as healthcare providers, physicians, government agencies, and wearable technology. It is necessary to implement a centralized data repository. Maintaining the high level of interoperability required for efficient information sharing at the right times is a significant challenge.

The lack of standards in the healthcare field exacerbates the situation even worse. Patient privacy and safety must be considered during the interoperability design phase. A lack of interoperability, for instance, could lead to medical blunders and put patients at high risk. Having timely access to data is

also crucial for ensuring patient safety. Patient data should be shared in real-time in response to a valid request, but care must be taken to protect patient confidentiality. This adds a new level of complexity to healthcare administration. One difficulty with big data in healthcare is quick changes in actual facts.

Big data in healthcare is an invaluable resource that can be described by its size, variety, speed, veracity, and value. Clinical decision support systems use the information in this data by following a path from descriptive analytics to predictive analytics to prescriptive analytics.

E. THE NEED OF BIG DATA ANALYTICS IN HEALTHCARE

The number of BDA applications in healthcare is gradually expanding due to the increasing volume of big data in this area. Big data in healthcare may arrive from various sources, including diverse and multi-spectral observations on patients, such as their demographics, treatment histories, and diagnostic results. Data can be structured (e.g., genotype, phenotype, or genomics data) or unstructured (e.g., a collection of observations) (e.g., clinical notes, prescriptions, or medical imaging). When it comes to implementing data in healthcare, it is frequently necessary to generate and gather high-quality real-time data. Decision-makers in the healthcare industry can take meaningful action due to significant insights gained from large amounts of information. The enormous rise in data acquired via EHRs, registries, or wearable sensors has brought a big data revolution to the health sector. This huge available data gives several benefits such as increased quality of life, disease diagnosis, treatment, and healthcare service delivery system. Big data generated in healthcare is massive, heterogeneous, and fast. In addition to non-uniform data, big data in healthcare requires real-time data analysis. Big data is evolving rapidly, and healthcare organizations are deploying technology to keep themselves updated.

According to Hardy Carter [51] and Dimitrov [38], the demand for big data in healthcare can be classified into potential advantages, as listed in IV-E. Real-time applications of big data in healthcare can also be grouped into three subcategories, namely: a) improving patient care, b) enhancing doctors' experience, and c) reducing organizational efforts.

- Predictive modeling to identify patient-centered conditions
- Early detection and prevention of diseases or medical conditions
- Extensively research and development to cure diseases
- Promotion and use of Electronic Health Records (EHRs)
- Patient engagement and empowerment through data-driven insights
- Predictive analytics to identify and mitigate risks for patients
- Alert generation for instant care
- Health data analysis for strategic planning and resource allocation
- Fraud reduction and data security enhancement

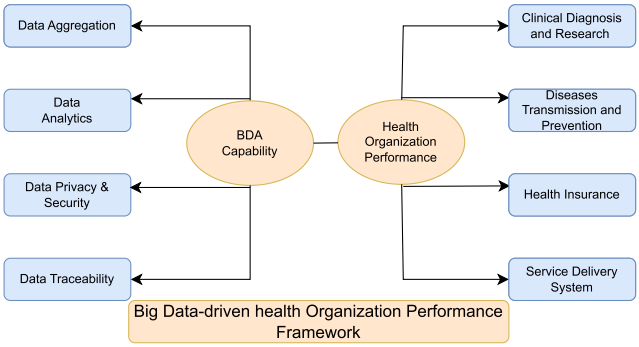

FIGURE 10. Big data-driven health organization performance framework.

- Reducing unnecessary hospital visits or emergency room visits
- Integration of medical imaging and other diagnostic tools for better healthcare outcomes
- Smart and better staff management
- Continued education and development opportunities for medical professionals
- Self-harm and suicide prevention using predictive analytics and intervention
- Help to develop new inventions
- Reducing administrative and managerial costs through data-driven decision-making.

The list of potential applications of Big Data Analytics in healthcare presented above is not exhaustive. However, it provides a comprehensive overview of some of the most important applications. Achieving these benefits will enable big data-driven health organizations to outperform their peers in their daily operations. Figure 10 depicts the Big Data-driven health organization performance framework, which includes sub-health-domains. It is noteworthy that the successful implementation of big data analytics in healthcare requires the collaboration of various stakeholders, including clinicians, data scientists, and policymakers. This collaboration enables the integration of data from multiple sources and the development of algorithms that can provide actionable insights. It is also important to ensure that ethical considerations are taken into account when using patient data for research purposes. This can be achieved through the establishment of clear guidelines and protocols for data sharing and informed consent [101].

F. OPPORTUNITIES FOR BDA IN HEALTHCARE

As per the discussion in the above section; We prepared a comprehensive list of possible tasks or Opportunities; these could be further summarized as per the following four categories.

We may consider any of the below-mentioned opportunities as BDA healthcare applications for future application of this review. BDA can also be used to analyze patient data, such as symptoms, medical history, and lab results, to help healthcare providers diagnose more accurately.

- Medical Diagnosis - Medical Diagnosis can be multi-modal in nature; such as medical images (X-rays, MRIs) or clinical raw text, time series, or tabular format data.
- Community Healthcare - BDA can improve community healthcare by analyzing population health data, identifying risks and trends, and developing targeted prevention and intervention programs. BDA can identify high-risk populations for diseases like diabetes and heart disease and develop outreach programs to promote healthy behaviors and prevent disease.
- Hospital Monitoring - BDA may be used to monitor and enhance hospital operations, including patient flow, resource allocation, and quality of treatment. For example, BDA may be used to measure patient wait times, detect bottlenecks in the treatment process, and optimize resource allocation to enhance efficiency and save costs. BDA may also be used to evaluate patient safety and quality of treatment by assessing patient data, such as medication mistakes and adverse events.
- Patient Care - BDA may be used to improve patient care by evaluating patient data, monitoring patient progress, and delivering individualized treatment suggestions. For example, BDA can be used to assess patient vital signs, examine pharmaceutical efficacy, and forecast patient outcomes. Moreover, BDA may be utilized to create individualized treatment plans predicated on patient data and medical history, allowing doctors to give more precise and efficient care.

G. CHALLENGES OF BDA IN HEALTHCARE

The benefits of big data analytics (BDA) in healthcare are significant, but implementing BDA poses various challenges. Our review of relevant literature [25], [32] has identified three main categories of challenges: data, process, and management challenges. To help visualize these challenges, we have created Figure 11. Note that the figure only lists a few challenges for each category, as the list of potential challenges is extensive. In this section, we briefly describe the most commonly reported challenges.

- **Data privacy and security:** Healthcare data is sensitive and contains personal information that must be protected to comply with regulations such as HIPAA. Implementing Big Data in healthcare requires a robust security infrastructure to protect patient information from unauthorized access or theft.
- **Data quality:** The quality of healthcare data [35] is crucial to ensure accurate analysis and predictions. However, healthcare data is often incomplete, inaccurate, or inconsistent due to various factors such as human error, outdated systems, or inadequate data management practices.
- **Data integration:** Healthcare data is often stored in multiple disparate systems, making it difficult to integrate and analyze effectively. Integrating data from different sources requires a standardized data format and a robust data integration infrastructure.
- **Resource Constraints:** Implementing BDA in healthcare requires significant resource investment, including hardware, software, and personnel. Lack of resources may hinder the implementation of BDA in healthcare.
- **Data governance:** Effective healthcare data governance is crucial to ensure compliance with regulations, maintain data quality, and protect patient privacy. This requires a clear definition of roles and responsibilities, policies and procedures for data management, and a framework for data sharing.
- **Skills and expertise:** Implementing Big Data in healthcare requires skills and expertise in various areas such as data analytics, data science, machine learning, and software development. Healthcare organizations may need to invest in training or hire new talent to build the necessary capabilities.
- **Cost:** Implementing Big Data Healthcare can be costly due to the need for infrastructure, hardware, software, and human resources. Healthcare organizations may need to invest significant resources to implement a robust Big Data infrastructure.
- **Resistance to change:** Healthcare is a highly regulated and conservative industry, which can lead to resistance to change. Implementing Big Data in healthcare requires a culture shift towards data-driven decision-making and a willingness to adopt new technologies and practices.
- **Ethical considerations:** The use of Big Data in healthcare presents ethical questions, such as the use of patient data for research or business, the possibility of discrimination or bias, and the need to tell patients how their data is being used.

Addressing these challenges requires a strategic approach that considers the unique characteristics of the healthcare industry and the specific needs of patients and providers. By addressing these challenges, Big Data can significantly impact healthcare, leading to better outcomes, improved efficiency, and reduced costs [18], [37], [43], [45], [46], [65], [68], [71]. Each of the above-mentioned challenges could be defined in a detailed manner but presently we only focus on security and privacy concerns as discussed in the following subsection. The list of challenges can not be finalized as we came across various interchangeable terms. Individual researchers [55], [57], [58], [64], [68], [80], [127] put their efforts to list few of them as we have also presented a very short but effective list of challenges being faced in BDA healthcare.

H. BIG DATA SECURITY AND PRIVACY IN HEALTHCARE

Security and privacy [60], [61] in the context of big data are crucial considerations. However, both terms are mistakenly treated as the same and refer to distinct concepts that are difficult to differentiate.

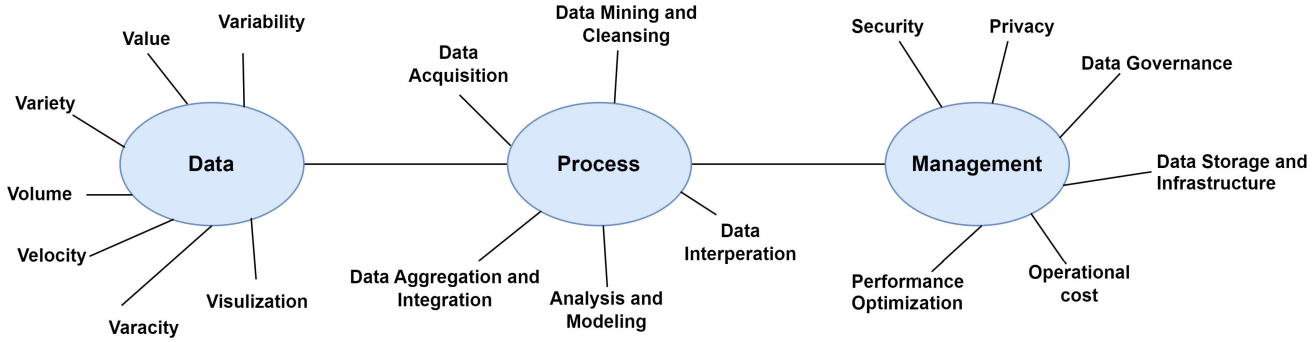

FIGURE 11. Potential classification of BDA challenges in healthcare.

Security is the confidentiality, integrity, and availability of data while privacy is the appropriate use of user's information. For security, various techniques such as Encryption, Firewalls, etc. are used to prevent data compromise from technology or vulnerabilities in an organization's network. To maintain privacy organization can't sell its patients', and users' information to a third party without the user's prior consent. Security may provide confidentiality or protect an enterprise or agency while privacy concerns patient's right to safeguard their information from any other parties. Security offers the ability to be confident that decisions are respected. While privacy is the ability to decide what information an individual goes and where to. At last, security focuses on data protection, while privacy concerns the appropriate use of user's information [47], [54], [129].

Security is often defined as the prevention of illegal access, various definitions also include the preservation of data integrity and availability, among other things. It is primarily concerned with protecting data against malicious attempts and stealing data for financial gain. Although security is critical for data protection, it is insufficient when securing personal information. Further, we may have to deal with the below-mentioned sub-areas for better understanding and implementation and leave them uncovered as it is out of our coverage.

- Big data security life cycle [54], [61]
- Technologies in use [130], [131]

Privacy is frequently characterized as protecting sensitive information, such as personally identifiable health care information, from being disclosed to unauthorized parties. In particular, it focuses on the use and control of an individual's personal data, including the development of rules and the establishment of authorization criteria to guarantee that personal information about patients is gathered, disseminated, and handled appropriately. Below are the main sub-domains for further exploration and we left this part and did not shed it as it is out of the domain of study.

- Data protection laws [55], [61]
- Privacy-preserving methods in big data [68], [132]
  - -- De-identification [18]
  - -- HybrEx [41], [54]
  - -- Identity-based anonymization [133]

V. BIG DATA APPLICATIONS IN HEALTHCARE - ANSWER TO RQ2

The healthcare industry's sources of Big Data include hospital records, patient medical records, test results, and Internet of Things (IoT) devices. Biomedical research generates Big Data that is used in public healthcare. By integrating biological and healthcare data, modern healthcare organizations can modify medical therapy and even personalize medicine, as noted by Dash et al. [74]. Healthcare has become a vital component of people's lives, resulting in an explosion of medical big data. Healthcare practitioners are now utilizing IoT-based wearable technology to expedite diagnosis and treatment. The Internet has recently connected billions of sensors, devices, and automobiles [38]. Remote patient monitoring is one such technique used currently in inpatient treatment. Despite the benefits of these technologies, they also raise significant concerns regarding the privacy and security of data during transit and logging. The delay in treatment could risk the patient's life.

A. HEALTH STANDARDS DOCUMENTATIONS

The development of the health platform has taken a significant amount of time. The Clinical Documentation Architecture (CDA R1) was first defined in May 2005 [134], and it became the American National Standards Institute (ANSI) approved HL7 standard [135], [136], which became the specification for the Reference Information Model (RIM) [190]. Even while it has been disseminated worldwide, its implementation is still not as widespread as it should be. The Continuity of Care Document (CCD) is an HL7 CDA implementation of the Continuity of Care Record (CCR). A summary of the patient's health state, including issues, drugs, and allergies, is included in the CCR data set. This summary includes fundamental information regarding the patient's care plan, documentation, and health insurance [62], [137]. Further, from the literature, it is observed that a recent development and speedy migration from HL7 document structure to FHIR. FHIR is a standard for electronic healthcare data exchange proposed by HL7. FHIR is HL7's latest and most popular healthcare data-sharing standard. It integrates HL7's v2 and v3 but also offers more contemporary and flexible interoperability. Further, it uses

HL7 communications protocols including HL7 Version 2 and HL7 Version 3, and current web technologies like RESTful APIs and JSON for organized data exchange. Healthcare providers, suppliers, standards development groups, and individual contributors collaborated on FHIR. HL7 encouraged conversations, consensus-building, and field testing to ensure FHIR met stakeholder interoperability needs. Healthcare community input to provide a more current and adaptable healthcare interoperability standard [191], [192].

B. MULTIMODAL BIG DATA ANALYTICS

The term ''multimodality'' refers to the process of utilizing a wide range of data types along with various modes of representation. Data in healthcare is multimodal by nature and it is becoming more multimodal day by day. Emerging technologies allow people to use different ways to interact with a system and combine different types of information simultaneously. Multimodal data with the help of AI tries to understand and get insights from different types of data parameters by making connections between them [193]. In fields such as biology, medicine, and health, it can help analyze connections between different biological processes, health indicators, and outcomes. It can also be used to create models for understanding and explaining these relationships [90], [104], [113].

In the field of high-performance computational sciences such as big data analytics and processing, multimodality is relatively a new concept that aims to integrate multiple data streams in various formats such as text-image-video and audio to enhance the precision of information extraction and inference, reduce bias, and generate an overall better representation of the physical, medical, or societal processes that are described by the data. Incorporating multimodal into the processing of multidimensional and multimodal data sets in mission-critical domains such as health and medicine can help to design better decision support systems; inherently better health analytics, improve prediction, diagnosis, risk factors, and patient follow-ups. These systems are to be used by health professionals and policymakers.

Multimodal data encompasses information derived from multiple sources or modalities, which unveil essential characteristics of real-world domains, including clinical applications. Currently in the clinical domain, disease severity or disease diagnosis and mortality prediction-based machine learning models require multimodal data to achieve better results than the conventional approaches. Missing data is commonly reported in multimodal data [86], [113]. As different types of examinations are conducted for individual patients and missing data may arise due to mishandling/data corruption errors from several components such as demographics data collection, lab test data, clinical notes, etc [113].

Recently, COVID-19 became a reason for adopting a multimodal approach because from the literature it has been noticed that COVID-19 data was generated in a multimodal nature. This increased the demand for such tools and techniques to predict, prevent, and manage diseases at a large scale from a single patient to the whole sample [84], [90]. The extensive use of multimodal approaches also has been reported in several studies in other areas of bio-medicine and health, such as chronic disease surveillance, screening and assessing child mental health, oncology, emotion detection, ophthalmology, and detecting dementia [138]. The article by Baltrusaitis et al. [56] provides a comprehensive review of recent developments in multimodal machine learning.

MAET – Mask Adherence Estimation Tool, an application-based study presented by Gupta and Srivastava [104]; MAET is a robust system to detect the pattern of public mask-wearing using pre-trained model YOLOv5 and integrates YOLOv5 with explainability to help the user understand at an individual and aggregate level. One more study [100] presented applied research for predicting ICU-admission ratio using a factor graph-based model. Another study [89] presented a semantic network analysis of pandemic patients' vaccine text dataset of Reddit. Zadorozhny et al. [111] suggested a set of practical evaluations and tasks to consider when selecting the best Detection of Out-of-Distribution (OOD) samples for a particular medical dataset. During the pandemic, users shared their stories, and experiences, and governments used to convey pre-cautionary messages on social media channels including but not limited to Reddit, Twitter, and Facebook. This rich information became a useful source for researchers to collect and analyze multimodal data. And Rohan Bhambhoria et.al presented a naïve NER-named entity extraction-based paper for clinical insights on covid-19 Twitter data.

In study [139], authors introduced a Python-based library known as PyHealth. It is a complete Python healthcare AI toolkit developed for ML researchers and healthcare professionals. PyHealth accepts a wide range of healthcare data, including longitudinal EHRs, continuous signals (ECG, EEG), and clinical notes (to be added), and supports deep learning and other advanced machine learning algorithms. PyHealth has five key benefits. First, predictive health algorithms such as XGBoost and auto-encoders are included, as well as current deep learning architectures such as convolutional and adversarial models. Second, PyHealth has broad coverage with models for sequence, visual data, physiological signals, and unstructured text data. Third, PyHealth provides a consistent API, thorough documentation, and interactive examples for all methods, making complicated deep learning models simple to use. Fourth, most PyHealth models have cross-platform unit testing with continuous integration, code coverage, and maintainability checks. In addition, PyTorch supports fast GPU computing for deep learning models, enabling parallelization in select modules (data preprocessing). The PyHealth library comprises a collection of 30 AI-based models. For a comprehensive list of these healthcare AI models available in PyHealth, interested readers are encouraged to refer to Table-1 in [139].

In a recent study Joshi et al. [106] state that data sharing and collaborative model training are promising ways to improve the quality of healthcare models. However, it is usually difficult to implement such settings in practice due to data privacy concerns and relative regulations such as the GDPR and HIPAA.

Shah et al. [66] presents a tutorial study on big data and predictive analytics. They presented four major barriers to useful risk prediction:

- Data quality and heterogeneity
- User trust, transparency, and commercial interests
- Statistical prediction
- Thoughtful identification of risk-sensitive decisions

A significant amount of potential exists for big data and predictive analytics to promote better and more efficient treatment, and there have been important recent developments, particularly in the field of image analytics such as below mentioned sub-domains [66].

- Clinical diagnosis and research
- Disease transmission and prevention
- General Healthcare
- Health insurance
- Service delivery system

C. THE CONCEPT OF DATA FUSION IN MULTIMODAL DATA

Data Fusion can be considered a study of data sets from different sources communicating with each other [33], [95]. Further, this study suggests that data fusion improves the performance of a particular framework/methodology or algorithm if considered in data analysis. The concept of Fusion is generally classified into two types: model-agnostic approaches and model-based approaches. The latter is further classified into three sub-types; i) Early ii) Late and iii) Hybrid followed by the data fusion keyword. Data fusion, which uses ML and DL techniques to combine data from different sources, is becoming increasingly important in medicine. Data fusion is widely used in the research community as a proper method for multimodal data analysis [86]. Since the inception of the fusion concept several methods [33], [56], [96], [137], [161], [193], [194] have been proposed to deal with the fusion of multiple data types for such we prepared a comprehensive taxonomy view for fusion handling techniques as presented in Figure 12.

The integration of multiple sources of data is a challenging task. However, data fusion's techniques and levels aim to provide data integration services for multimodal data where a single modality does not work. Data fusion is also challenged by noisy and irrelevant data that could lead to weak models and degraded performance [95]. Furthermore, data fusion steps including combining and normalizing data require high computational power, which is severely challenging for multi-modalities data fusion [88]. Last but not least challenge with data fusion is that no ''off the shelf'' technique is available that could always work for any type of data combination and could not guarantee enhanced results compared to a single modality. Nevertheless, algorithms such as GLRM-generalised low-rank modelling could be considered to combine different types of data and to develop better prediction models.

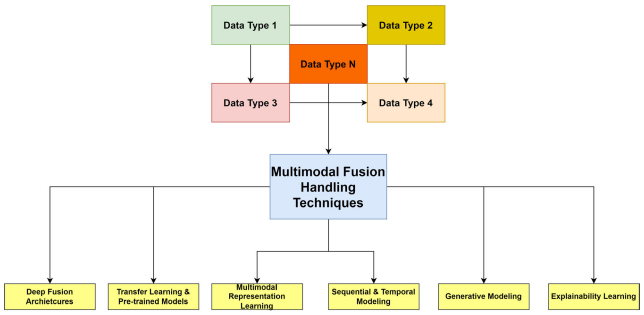

FIGURE 12. Multimodal fusion handling techniques taxonomy from literature.

D. CURRENT DIGITAL HEALTHCARE SYSTEM

The term ''digital health'' is understood as advanced analytics based on multi-modal data. It tries to maximize the use of IoT-based sensors to enable clinicians to access the right information at the right time. With digital health systems, there are also ways to collaborate with specialists from across borders. The implementation of digital healthcare systems has revolutionized the healthcare industry by automating routine laboratory work and essential procedures, thereby enabling clinicians to allocate their time and attention to critical cases. The integration of AI and DL platforms within these systems further enhances their capabilities, allowing them to perform necessary actions and support clinicians in delivering efficient and effective diagnoses and treatments for patients. Furthermore, the digital system also helps automate billing and further documentation work. Last but not least digital healthcare systems mean providing care to a single patient while also providing care to thousands of patients all at the same time [63].

The need and increase in demand for digital/internet-based healthcare systems is rapidly growing. It is anticipated by the end of 2050, older people aged around 60 will reach 200 million and 80% [140] of them will be from developing countries, and for them, a healthcare system is a major concern.

The authors [137] propose the Tianxia120 digital medical health system for ''one-step service'' to both patients and hospitals. The system can rigorously promote the change of service status between doctors and patients from ''passive mode'' to ''proactive mode'' and realize online service that is similar to offline medical treatment scenarios. There are separate terminals for patients and doctors. Further authors claim that this system is full-function as well as rich in size of data and major security concerns are already tackled.

Davide Ferrari et al. [103] presented a review-based study for data-driven and AI-based clinical practices. They have thoroughly reviewed a couple of recent studies and

concluded a list of recurrent research issues including but not limited to i) Data Imbalance ii) Data Inconsistency iii) Data Sparsity followed by case three real-time case studies (''My Smart Age with HIV'', ''Covid–19, predicting respiratory failure'' and ''Covid–19, predicting oxygen therapy states'') to support the challenge list.

In current healthcare innovation, the four P's concept is gaining popularity. P4 - Preventive, predictive, personalized (individually tailored), and participatory; cannot only ensure people's lives much better but also save a lot of money and make healthcare more efficient. This study focused on age-related disorders and investigated the potential of a data-driven method to forecast the wellness states of aging persons, as opposed to the knowledge-driven approach that depends on easy-to-interpret measures routinely supplied by clinical specialists. The results show that the data-driven method is better at making predictions. We also show that a post hoc inference procedure can be used to explain the predictive models in a way that makes sense and opens the door to new kinds of personally tailored and preventative care [85], [96].

E. BIG DATA AND MDL FOR HEALTHCARE

Analysis of big data by MDL - Machine or Deep Learning offers considerable benefits in evaluating a large and complex set of healthcare corpus [70], [141]. However, before moving further MDL poses several challenges that need consideration as mentioned above IV-G. One of the advantages of MDL in healthcare is its flexibility and scalability compared to traditional bio-statistical methods. It can be used for various tasks, including risk stratification, diagnosis and classification, and survival predictions. MDL can also analyze diverse data types, including demographic data, laboratory findings, imaging data, and doctors' free-text notes, and incorporate them into predictions for disease risk, diagnosis, prognosis, and appropriate treatments. However, the application of MDL in healthcare also presents unique challenges including but not limited to data pre-processing, model training, and refinement of the system with respect to the actual clinical problem are crucial. Additionally, ethical considerations, such as medico-legal implications, doctors' understanding of machine/deep learning tools, and data privacy and security, must be considered [49], [56], [94], [97]. While reviewing an immense and complicated set of healthcare corpus, conducting an analysis of big data using MDL - Machine or Deep Learning offers a number of advantages that are worth considering [70], [141]. However, before going any further, MDL presents a number of difficulties that need to be taken into consideration, as was mentioned above in section IV-G. In comparison to more conventional approaches to bio-statistics, the flexibility and scalability of MDL make it an attractive choice for use in the healthcare industry. It can be utilized for a variety of purposes, including the stratification of risks, the diagnosis and classification of conditions, and the forecasting of survival times. MDL is also capable of analyzing many sorts of data, like as demographic information, laboratory findings, imaging data, and free-text notes written by medical professionals, and incorporating the results of these analyses into predictions on disease risk, diagnosis, prognosis, and the most relevant treatments. However, the application of MDL in the healthcare industry also presents a number of one-of-a-kind challenges. These challenges include but are not limited to, the pre-processing of data, the training of models, and the refinement of the system in relation to the actual clinical problem. Additionally, it is necessary to take into account ethical concerns, such as medico-legal implications, doctors' understanding of machine learning and deep learning tools, as well as data privacy and security [49], [56], [94], [97].

F. NATURAL LANGUAGE PROCESSING IN HEALTHCARE

The study [91], proposed a model namely MedCAT – An open-source toolkit for annotation of medical concepts that is capable of self-supervise machine-learning algorithm for concepts extraction using any of the standard concept vocabulary UMLS and/or SNIMED-CT; MedCAT also provides a customizable information extraction interface. MedCAT achieved an improved F-score in comparison to the available benchmark (F1:0.448–0.738 vs 0.429–0.650). Furthermore, MedCAT is an open-source Named Entity Recognition + Linking (NER+L) and contextualization library. MedCAT is based on CogStack [142]; it is an application framework that extracts data from unstructured data. CogStack ecosystem integration makes MedCAT easy to deploy in health systems. The annotation tool, MedCATtrainer [143] lets clinicians take a glance at, change, and improve the extracted concepts through a web interface made for training MedCAT information extraction pipelines.

The authors in [93] introduce a quick and precise fully automated way to find COVID-19 in a patient's chest CT scan also termed HRCT lung scans. They have introduced their own set of CT scan images of 48,260 from 282 healthy people and 15,589 images from 95 people with COVID-19 infections. Also, they have proposed a naïve image processing algorithm to quickly analyze the status of lungs to discard non-suspicious images from the complete input dataset; this helps to reduce preprocessing time and minimize the false detection ratio. Further, this study, combined the ResNet50V2 model with a new feature pyramid network optimized for classification challenges, allowing the model to explore images at varying resolutions without losing information on fine details. They claim that they are the first to evaluate their naïve algorithm on (Xception and ResNet50V2); this improves classification performance significantly because COVID-19 infections come in numerous sizes, including microscopic ones. With single image classification, this approach resulted in 98.49% precision and in a real-time system correctly classified 234 out of 245 input images. Few of the studies from the literature targeting Multimodal data using NLP are listed in a tabular form 8. Further, from an applications perspective to resources, we have presented a summary in Table 8 for NLP in Healthcare.

TABLE 8. A comprehensive list of literature targeting NLP multimodal data.

Study and YOP	Brief Summary	Dataset and Achieved results	Core application area	Future work/gape analysis
[69]	In this study, authors presented an experimental study to reduce errors while predicting possibilities of "Intravenous Thrombolytic Therapy" in stroke patients using task-based electronic medical records with the help of natural language processing. In addition to being a major contributor to global mortality and disability, stroke is also a major loss on healthcare budgets. The Ischemic strokes account for 75% to 90% of all strokes. MetaMap, an NLP tool developed by the National Library of Medicine was used. Furthermore, the study presents three phases of experiments as follows Phase I: extraction of CUIs from IVT eligibility criteria Phase II: identification of CUIs from EMRs Phase III: a task-specific EMR interface for assessing IVT eligibility criteria Interested readers are suggested to the referenced study for details.	The dataset was collected from "Ditmanson Medical Foundation Chiayi Christian Hospital" and was altered to preserve patient privacy. Their approach detected IVT with micro-averaged precision, recalls, and F1 values of 0.998, 0.812, and 0.895 and document-level measures of 1, 0.972, and 0.986. They concluded that the task-specific interface was more accurate at assessing IVT eligibility (91% vs. 80%, p = 0.016) than the present interface. The interfaces almost took statically the same time in minutes (2.46 min and 1.70 min with p = 0.754).	Predicting the existence of IVT in Stroke Patients.	Future studies should consider combining information from structured data and incorporating feedback from domain experts to iteratively optimize the information's relevance, enabling physicians to improve acute stroke care.
[144]	In this study, Alsentzer et al. pre-trained and publicly released various clinically-focused BERT models, some of which are trained exclusively on the clinical text and others tweaked over BioBERT. On the contrary, they find strong evidence that their clinical embeddings outperform both general domain and BioBERT-specific embeddings on non-de-ID tasks, and further added that the use of clinical-note-type specific corpora might generate additional selective performance improvements. They presented a unique and valuable work. Authors believe that all clinical NLP researchers should have access to these embeddings, as future researchers can achieve more efficient performance without requiring the massive amounts of computational resources needed to train models over the MIMIC corpus.	In this study, the MIMIC-III 1.4 version dataset was considered and 2 million clinical notes were utilized. In comparison to BioBERT or standard BERT, clinically fine-tuned BioBERT performs significantly better on three out of the five tasks (MedNLI, i2b2 2010, and i2b2 2012). Remarkably, clinical BERT produces an entirely new state of the art on MedNLI, with a performance of 82.7% accuracy, in contrast to the previous state of the art, which produced 73.5% accuracy.	They intend to demonstrate that utilization of domain-specific datasets/corpus increases the model's performance. In short, their application is BERT model performance enhancement.	Limitations of this study are: -Low architecture (computational power) -MIMIC only contains critical care unit notes from one hospital I-e. BIDMC. Using clinical notes from different and several institutions would impact performance by not being biased toward a single entity.
[77]	In this study, the authors examine the effectiveness of novel variants of LSTM, such as AWD-LSTM, on MIMIC-III discharge notes. Using this model, they conclude that one can achieve decent results for the proposed hypothesis, such as extracting diagnosis/procedure codes from clinical notes. The model uses deep learning and natural language processing techniques to automatically assign ICD-9 codes to clinical notes. Experiments have shown that deploying AWD-LSTM would therefore serve as a benchmark for further research in identifying diagnoses, procedures, and treatments utilizing a single model all at once. Additionally, with the automation of medical coding, it is possible to save time and reduce costs arising from manual coding errors.	MIMIC-III dataset's clinical notes were examined. Authors utilized the state-of-the-art deep learning algorithm ULMFiT on the publicly available open clinical note dataset I.e. MIMIC III with 1.2M notes to select the top-10 and top-50 diagnosis and procedure codes. Their models predicted the top-10 diagnoses and procedures with 80.3% and 80.5% accuracy, and the top 50 ICD-9 codes with 70.7% and 63.9%.	In this study, authors intended to evaluate their novel model to allocate ICD-9 code automatically to the clinical text to save cost and time and reduce human error.	One of the difficulties they encountered was a constraint of resources to support the execution of high-end activities.
[79]	In this study, Si et al. present an analysis based on a new approach to enhance context extraction from clinical notes using natural language processing. Then the proposed model was compared to state-of-the-art available word embedding models such as (Word2Vec GloVe and fastText).	They utilized four corpus including (i2b2 2010, i2b2 2012, SemEval 2014 & SemEval 2015).Their proposed approach outperforms in contrast to the state-of-the-art methods with the results (f1: 90.25%, 93.18%, 80.74%, and 81.65%). Further, they conclude that utilizing state-of-the-art models for context extraction seems valuable.	This research aimed to compare and contrast several word embedding techniques and examine their performance on four clinical concept extraction tasks.	They used a limited set of corpora for comparisons.

G. HEALTH EDUCATION PROMOTION

Health education [27] is gaining importance and is being considered a crucial topic in healthcare discussions. A gamification-based health education promotion study presented by Hsu et al. shortly named it KABAN. The study focused on the health literacy and knowledge of older adults through game-based learning. In a study conducted by pre-trained instructors and based on instructors' feedback, KABAN strongly supported the proposed idea that older adults' health literacy can be enhanced through effective gamification learning designs. They discussed the reasoning behind selecting their age group i.e. elder people because from the existing studies [105] they found that this age group of people around the globe are less educated than the current eras' young and teenagers. Older people are motivated and interested in learning more about their health through a game-based health intervention [83]. In [105], the authors also talked about how important instructors were to the success of the study and how their feedback on the design of the intervention was helpful. The shortcoming of this study is the selective age group and selective people from socially active communities. A good understanding of health will make it easier for older people to live in a way that is good for their health, and initiatives such as KABAN can help people learn about health. The authors of this study look for future interventions in this area of healthcare.

Healthcare organizations are searching for appropriate technology that will simplify resources to improve the patient experience and the organization's overall performance. Authors from the studies [16] and [18] suggest that healthcare can be thought of as a system with three basic parts: *the patient, the provider, and the system.* Core medical care service providers, such as physicians, nurses, technicians, and hospital administrators, are included in this category. Critical services that are related to medical care services, such as medical research and health insurance, as well as recipients of medical care services, such as patients and the general public.

H. BLOCK-CHAIN AND HEALTHCARE

In study [75], authors proposed using blockchain to secure healthcare large data administration and analysis. Blockchain technology is prohibitively expensive for most resource-constrained IoT devices destined for smart cities, necessitating substantial bandwidth and computational power. Using blockchain with IoT devices presents several challenges. To address these issues, we present a novel architecture of modified blockchain models suitable for IoT devices. Our model's extra privacy and security features are based on advanced cryptographic primitives. These technologies use a blockchain-based network to make IoT data and transactions more secure and anonymous.

In a recent study [102] authors discussed the potential challenges associated with personal health records and the vulnerability of centralized healthcare systems. To secure the data and solve the existing challenges they proposed a blockchain-based architecture that allows patients to manage their health information securely further they claim that their novel architecture has the capability to handle issues and provides better privacy and security to patients' records. Another study [112] proposed Blockchain-based privacy-preserving for healthcare data in the cloud. They discussed the potential benefits of cloud-based electronic healthcare systems. We also listed several software solutions targeting cloud-based healthcare in Table 14. Furthermore, blockchain technology can give structure and security to healthcare data, as discussed in another study [109], which also examines the difficulties of implementing such information in a web-based context. Last but not least this study [99], presented a comprehensive overview of the benefits of integration IoT and block-chain in healthcare applications, they claim their survey will serve as the baseline for future researchers targeting IoT and Block-chain in healthcare.

I. ELECTRONIC HEALTH RECORD (EHR)

In the current times, EHR electronic health record is gaining attention in both (public, and private) sectors of hospitals, clinic, and medical service centers. EHR seems to be an additional component to deal with by doctors, physicians, assistants, and other medical staff, etc. Before EHR, doctors used to write clinical notes, and maintain records manually [7], [67] and difficulties have been reported in the literature while using EHR systems by medical consultants. Researchers have conducted many studies to highlight how medical consultants' interactions with EHR systems may affect patient communication, for example, using a keyboard, or mouse, and gazing at the computer [67]. Further, studies have also examined the implications of EHR use on physician-patient communication and the possible implications for the quality of health care [28], [30]. One study [28] has claimed that patient-centered interaction suffers because the physician's focus is diverted to the electronic health record (EHR) rather than on the patient. About half a decade ago, one of the studies was presented [36] LAB-IN-A-BOX; a naive framework for tracking activities during physician-patient interaction (the system was semi-automatic) and gained attention at that time and authors claimed that their approach has the potential to uncover important insights.

Electronic health records contain brief information about patients that can be followed over time, including the patient's medical and medication history, symptoms, complaints, therapy, procedures and tests, final diagnosis, discharge meds, and treatment notes or referral notes. It provides experts with a large amount of data as a review or a key to the start in case a new consultant takes over the case. Data maintenance and collection for future work on data is a new challenge for data centers of hospitals as every second a new record is being inserted. In the domain of healthcare, EHR data plays a vital role in the development of artificial intelligence, machine learning, and big data analytical systems. Much work has been proposed on prediction, analysis, and natural language inference. Further, it has many open challenges to be tackled. Some of the open sources of data collection present a wider scope for future work such as MIMIC – Medical Information Mart for Intensive Care dataset; The MIMIC database is the largest Electronic Health Record (EHR) database that is freely available to the public and may be used to test various machine learning methods. EHR produces data in two/three forms such as structured, unstructured, and semi-structured, such as Lab results, doctor medications, and clinical notes are examples [42], [77]. Despite these developments, access to medical data to improve patient care remains a significant barrier [92].

J. PATIENT CENTRIC HEALTHCARE

After a thorough review, we here mention Patient-Centric care; it is one of the progressing areas [116], [145] that is being considered as a novelty solution in combo with other fields of healthcare such as personalized medicine solutions or maybe personalized clinical prediction, etc. Patient-centric solutions have the potential to significantly improve healthcare outcomes by focusing on the needs and preferences of individual patients. These solutions can help forecast disease outbreaks, prevent diseases, improve patient outcomes, and reduce healthcare costs. By developing clinical prediction models tailored to patients' specific needs, healthcare providers can improve the accuracy of diagnoses and treatments, ultimately leading to better patient outcomes. Patient-centric solutions also have the potential to enhance patient engagement and satisfaction by involving patients in their own care and providing them with personalized treatment plans [195]. Here we summarized the existing Patient Centric solution or framework as presented in Table 10.

VI. HEALTHCARE DATASETS, MODELING TOOLS AND TECHNIQUES - ANSWER TO RQ3

The most reliable and authentic sources of healthcare data include electronic health records (EHRs), claims and billing data, health registries and surveillance systems,

TABLE 9. A summary of NLP in healthcare - from applications to resources.

Levels of NLP	Applicatiosn of NLP	NLP techniques	NLP systems	Challenges of NLP	NLP resources
-Phonological Analysis -Morphological Analysis -Lexical Analysis -Syntactic Analysis -Semantic Analysis -Pragmatic Analysis	-Information Extraction -Information Retrieval -Question and answering -User Interfaces -Document Categorization -Machine Translation -Text Summarization	-Symbolic/Logical Approaches -Statistical Approaches -Connectionist Approach -Hybrid Approach	-Medical Language Extraction and Encoding System (Medlee) -Clinical Text Analysis and Knowledge Extraction System (Ctakes) -Medical Literature Analysis and Retrieval System Online (Medline) -Metamap -Gene Tuc -Genia Corpus	-Rapid Growth of Incompatible Vocabularies in the Healthcare Domain -Negation and Uncertainty in Clinical Texts -Presence of Spelling Errors	-UMLS -Systemized Nomenclature of Medicine-Clinical Terms (SNOMED-CT) -Medical Subject Heading -Logical Observation Identifier Names and Code

TABLE 10. Notable list of patient-centric healthcare frameworks.

Reference	Framework	Data	Key Application
[24]	A patient-centric personalized healthcare framework	EMR- Electronic Medical Records	This work establishes the basis for a personalized healthcare approach utilizing Big Data, focusing on improving patient-centered outcomes, achieving meaningful use, and reducing hospital readmissions. The practicality of this approach is demonstrated through data-driven applications.
[26]	A framework for u-healthcare system	motion data and vital signs data	Key application of this study is a personalized healthcare system based on vital sign processing.
[31]	BDCaM-Context aware cloud-based framework for personalized patient car	Physiological data such as BP and HR	This study introduces a cloud-based big data framework aimed at improving personalized patient care through context-aware monitoring. By leveraging the power of big data, this framework aims to enhance patient outcomes and reduce the risk of hospital readmissions.
[28]	PCP - Primary Care Provider	EHR	This study proposed a novel model; PCP - Primary Care Provider. The experiments were conducted on a small population and results concluded that those participants gain more attention from PCPs when PCPs focused on the EHR. Further, they added, that how healthcare providers interact with electronic health records (EHRs) during medical encounters can impact patient-centered communication.
[73]	N/A	N/A	In this study authors present a review of the evolution of data-driven methods that offer the possibility to address Precision Medicine.
[145]	N/A	EHR	This study proposed meta-algorithm modeling solution for Patient-Centric applications. Further, this study evaluates applied data science techniques to Patient-Centric applications.
[195]	PCA Framework for EHR	EHR	This study utilizes blockchain technology, with a primary focus on patient-centricity, involving the implementation of pre-transaction signature verification mechanisms. As per the author's conclusion, these mechanisms play a crucial role in facilitating the execution of smart contracts pertaining to electronic medical data.
[196]	PCH	EMR	This study proposes the Patient-Centric Healthcare architecture (PCH), a safe and efficient Block-chain, Cloud, and IoT architecture for healthcare system interoperability.
[146]	N/A	N/A	The authors in this study conduct a comprehensive review to explain how academic research could be integrated into business intelligence solutions for patient-centric applications. They reviewed the period between 2000–2016 and concluded that business intelligence applications that include patient-centredness have grown since 2010 and focus on organization, humanism, and patient-centric conditions.
[147]	Blockchain-based Novel Architectural Patient-Centric Framework	N/A	A Block-chain, a decentralized technology, has the potential to address many of the issues associated with traditional electronic health records (EHRs). And authors in this article proposed a novel patient-centric blockchain-based EHR decentralized healthcare management system to solve existing issues.

health surveys, and wearable devices and remote monitoring. These sources provide valuable insights into patient health information, healthcare utilization, and population-level data. In terms of modeling tools, techniques, and commercial solutions in big data analytics for healthcare, machine learning and predictive modeling, NLP, data visualization and dashboards, and commercial analytics platforms are commonly used. These approaches enable the analysis of healthcare data, prediction of patient outcomes, extraction of information from unstructured data, and presentation of data in a user-friendly format. To address data governance and ethical considerations, strategies such as ensuring data privacy and security, obtaining informed consent, mitigating biases, and implementing ethical review and oversight processes are essential. In this section, we will discuss modelling tools and techniques, datasets, and solutions used in healthcare within the context of big data.

A. MODELING TOOLS AND TECHNIQUES

Every year, hundreds of prediction models [81], [101] are published in scientific publications, many of which use datasets too small for the total number of participants or events. Riley et al. [98] addressed in their study to propose a new methodology for sample size calculations for new experiments. In this article, the authors demonstrate how to calculate the sample size needed to build a clinical prediction model. There have been several studies in the past that used various modelling techniques in the domain of healthcare.

Jayanthi et al. [52] conducted a comprehensive survey of predictive modelling tools for diabetes prediction, which

we have summarized as general healthcare predictive tools collection and prepared a Table 11 after going through several studies such as in this study [110], authors used MTL RNN model for predictive modelling and found that the multitask models using MTL and RNNs outperformed single-task models in terms of individual-level predictions. Another study [77] used NN as predictive modelling for the top 10% of the diagnosis from the raw clinical text dataset. This study [148] employed a genetic algorithm to schedule the medical treatments using LSWT-GA which adopts a survival analysis strategy using heuristic knowledge to predict the effective schedule. Study [89] presents a detailed discussion on the mental health impacts of COVID-19 using Reddit Dataset by modelling NLP and computing technologies. The table is structured as a predictive method type and the name of the particular method.

In Table 12, we provide a comprehensive overview of frequently utilized big data technologies. Nevertheless, it is crucial to acknowledge that the primary objective of the table is to furnish a concise depiction of each item along with their respective benefits. The further detailed analysis and investigation of comparisons between these tools in terms of core services, architectural level, and outcomes have not been addressed in our review, however, we provide a concise description and key factors of each tool and their official source. Additionally, we would like to strengthen more on big data tools and techniques by summarizing them into multiple categories such as i) Distributed Computing Frameworks, such as Apache Spark and Hadoop, and ii) Streaming platforms, like Apache Kafka. iii) Machine Learning libraries, like Apache Mahout and TensorFlow provide algorithms and tools for training models, and predictive performance used for data-driven decisions. iv) Data visualization tools such as Tableau and Power BI help to present complex data in an easily understandable format. and v) No SQL databases, MongoDB, and Casandra are powerful and commonly useful tools for storing and managing unstructured and semi-structured data. By utilizing widely used big data tools and technologies, companies and researchers/academic practitioners could harness the full potential of these mentioned tools to get useful insights that facilitate the user requirements and scientific reasoning respectively.

The Multitask Learning (MTL) model is proposed to estimate bladder pressure with the assistance of time series data in this paper [110]. When it comes to modeling population-level time series data, MTL may be able to achieve a higher level of accuracy than ordinary neural networks. Taking advantage of the many types of data that are present in the population, can be accomplished by separating the prediction process for each individual participant in the population into their own individual task. They employ this innovative technology to forecast bladder pressure and then bladder contractions based on an external urethral sphincter electrocardiograph (EUS EMG) signal. The EUS EMG measures the muscle that controls the urethral sphincter.

TABLE 11. List of most repetitive predictive modelling tools in healthcare.

Predictive Method Type	Name of Method
Artificial intelligence models (These models are relatively new and are widely used)	• Neural networks • Genetic algorithm • Principal component analysis • Fuzzy logic • Nearest neighbor pairing • Rule induction • Simulated annealing • Conjugate gradient • Computing Methodologies
Statistical models (These methods rely heavily on past data)	• Fuzzy logic • Linear regression • Logistic regression • Anova • Time series • Trees • Non-linear regression • Survival analysis

They came to the conclusion that the multitasking models are superior to the single-tasking models when it comes to making predictions about individuals. The MTL RNN model performed significantly better than the other models when it came to predicting intra- and inter-individual differences in bladder contraction.

The healthcare industry has become a promising area for data mining and machine learning due to the availability of large volumes of data. However, the lack of publicly available benchmark datasets poses a significant challenge in quantifying progress in machine learning for healthcare research. This issue has led to the development of various initiatives to facilitate the sharing and access to healthcare data, such as the Medical Information Mart for Intensive Care (MIMIC) and the National Institutes of Health (NIH) National Library of Medicine's open-access database, PubMed Central (PMC). To solve this issue, Harutyunyan et al. [155] present four clinical prediction standards based on the "MIMIC-III database". These include predicting death, length of stay, recognizing physiologic decline, and phenotype classification.

Applications in clinical healthcare, natural language processing, speech recognition, and computer vision can all benefit from deep learning models (also known as deep neural networks). Few studies have compared the performance of deep learning models with current machine learning models and prognostic scoring systems using publicly available healthcare datasets. This is because few studies have used deep learning models. When it comes to determining mortality, length of stay, and ICD-9 code group, the author of this study [156] investigates how well Deep Learning models, ensembles of machine learning models (the Super Learner approach), SAPS II and SOFA scores perform. The MIMIC-III (v1.4) dataset, which is available to the public, was used for the benchmarking tasks. This dataset includes all of the patients who were hospitalized in an intensive care unit at Beth Israel Deaconess Medical Center between the years 2001 and 2012. Overall, deep learning models perform

TABLE 12. A comprehensive list of commonly employed big data analytics tools.

Tool	Description	Advantages	Link
Apache Hadoop	An open-source software framework that utilizes clustered file systems to handle big data using the MapReduce programming model.	Quick access to data due to the Hadoop Distributed File (HDFS) System. Highly scalable and cost-effective.	http://hadoop.apache.org/
Apache Spark	Apache Spark is an open-source, powerful data processing and analytics engine. Spark provides an interface for complete programming clusters with implicit data parallelism and fault tolerance and for interacting with other Spark components.	Quicker batch and stream processing. Ease of transformations. High compatibility with various tools.	http://spark.apache.org/
Apache Storm	An open-source data computation system with real-time computation abilities.	- Can be used with numerous programming languages. - Simple implementation - Consistent at scalability. - Fault-tolerant.	http://storm.apache.org/
Apache Kafka	An open-source distributed real-time data streaming platform designed to handle high throughput and is fault-tolerant. Kafka is	- Distributed Architecture - Fault tolerance - Easily integrable - Highly reliable	https://kafka.apache.org/
Apache Mahout	An open-source project used for the implementation of scalable machine learning algorithms such as classification, clustering, etc.	It is compatible with Hadoop libraries to effectively scale the cluster. - Provides ready-to-use frameworks for data analysis on big data sets. - Analyzes big data sets faster and more effectively.	http://mahout.apache.org/

better than any other strategy, particularly when utilizing 'raw' clinical time series data as input attributes.

B. BIG DATA DATASETS

Obtaining authentic and recent datasets is a critical task in the healthcare domain. In the literature, we reviewed several challenges regarding the datasets such as few challenges are listed below. While these listed challenges are very few among the open challenges for releasing a safe and reliable dataset.

- Dataset authentic source
- Dataset access and availability
- Dataset quality and completeness
- Dataset bias
- Dataset guides
- Dataset formats availability
- Dataset privacy and security
- Dataset volume and scalability
- Dataset heterogeneity

To address these challenges, we compiled a list of authentic sources (to the best of our knowledge) that list healthcare domain data, which can be found in Table 13. While there are numerous data sources available, we have focused on the most frequent and vital datasets that have been reported in the literature. These listed authentic sources also adhere to rules and regulations guided by HIPAA [106].

After conducting an extensive review of the literature [27], [35], [63], [101], [105], we have found that there has been an increasing focus on exploring the healthcare domain in the areas of system infrastructure and operation, quality of healthcare data, digital health education, and medical image analysis. These areas have emerged as key categories where most of the healthcare datasets can be classified, and they represent important research areas for the development and implementation of AI, ML, and DL in healthcare.

C. CURRENT BIG DATA ANALYTICS HEALTHCARE SOLUTIONS

In the age of digitization, the volume of data and research publications is growing at an unprecedented rate. Consequently, new big data analytics solutions are being proposed almost every day [74], [82]. In this subsection, we present a table of current solutions from the literature to shed light on recent developments. Table 14 provides a list of current healthcare solutions deployed on a large and medium scale [53], with a focus on AI, ML, DL, and NLP-based solutions relevant to the application areas we discussed. Although many other solutions exist, we have customized the list to fit our specific domain.

D. THE BIG DATA ADVANTAGE IN HEALTHCARE - USE CASES

With the digitization of Electronic Medical Records (EMR), Electronic Health Records (EHR), medical imaging, laboratory results, insurance data, and prescriptions, healthcare has generated a massive amount of data known as Big Data. Analysis of this Big Data can potentially improve the quality of medical and healthcare services by providing meaningful insights that help in informed decision-making, disease surveillance, and other healthcare and medical services. This can benefit patients, physicians, healthcare organizations, pharmaceutical companies, policymakers, and other stakeholders. Big Data applications can include individual and population health surveillance, predicting health issues, calculating medical complications and risks associated with a patient, analyzing suitable treatments, and evaluating the effectiveness of current treatment strategies [149]. Big Data can inform patients about their current and future health states, empowering them to make better-informed decisions. Integrating Big Data and healthcare makes it possible to scale the quality and accountability of health services, which offers numerous benefits, including improving the accuracy, timeliness, and effectiveness of healthcare services [150]. The benefits of using big data in healthcare are numerous and significant. One of the most important benefits is improved patient outcomes. With the ability to collect and analyze vast amounts of patient data, healthcare providers can identify patterns and trends in patient care and adjust treatment plans accordingly. This can result in more accurate diagnoses, better treatment outcomes, and improved patient satisfaction.

TABLE 13. A comprehensive list of healthcare dataset sources from existing literature and other resources.

Category	Datasets	Source Link
General Category	• HealthData • Big Cities Health Inventory Data Platform • Chronic Disease Data • Human Mortality Database • Mental Disorders Datasets • MHealth Dataset • Medicare Provider Utilization and Payment Data • LASA-Life Science Database Archive • WHO datasets	• https://healthdata.gov • https://bigcitieshealthdata.org • https://www.cdc.gov • https://mortality.org • https://ourworldindata.org/mental-health [157] • https://www.mhealthgroup.org/datasets.html https://archive.ics.uci.edu/ml/datasets/MHEALTH/ • https://data.cms.gov • https://www.nasa.gov • https://www.who.int/data/sets
Electronic Health Records (EHRs)	• MIMIC-III • eICU Collaborative Research Database • PhysioNet Archive (lists many EHR datasets) • Optum EHR • NHS Digital	• https://mimic.mit.edu/ and [42] • https://eicu-crd.mit.edu/ and [158] • https://physionet.org/ • [159] • https://digital.nhs.uk/
Medical Imaging	• ImageCLEF • RSNA Pneumonia Detection Dataset • ChestX-ray8 • Open Access Series of Imaging Studies (OASIS) • OpenfMRI • Alzheimer's Disease Neuroimaging Initiative (ADNI)	• https://www.imageclef.org/ • https://www.rsna.org/education/ • https://arxiv.org/abs/1705.02315 • https://www.oasis-brains.org/ • Not available • https://adni.loni.usc.edu
Genomics	• The Cancer Genome Atlas (TCGA) • The Human Genome Project • 1000 Genomes Project • GEO Datasets	• https://www.cancer.gov/ccg/research/genome-sequencing/tcga/ • https://www.genome.gov/human-genome-project • https://www.broadinstitute.org/projects/1000-genomes • https://www.ncbi.nlm.nih.gov/gds
Pharmaceuticals	• DrugBank • ChEMBL or ChEMBLdb • PubChem	• https://go.drugbank.com/ • https://www.ebi.ac.uk/chembl/ • https://pubchem.ncbi.nlm.nih.gov
Mobile Health (mHealth)	• Activity Recognition from Single Chest-Mounted Accelerometer • UCI Smartwatch and Smartphone-Based Activity Recognition Dataset	• https://archive.ics.uci.edu/ml/datasets/ • https://archive.ics.uci.edu/ml/datasets
Social Media	• Twitter Health Surveillance Dataset • Reddit Health Dataset • i2b2 Obesity Challenge Dataset	• https://github.com/DataTalks-ML/Twitter-Health-Surveillance • [197] • https://www.i2b2.org/NLP/Obesity
Cancer	• SEER cancer incidence • BROAD Institute Cancer Program Datasets • CT Medical Images	• https://seer.cancer.gov/ • https://www.broadinstitute.org/datasets • Multiple sources
COVID-19	• COVID-19 Open Research Dataset • COVID-19 Radiology Dataset	• https://innovation.mit.edu/cord19/ • Multiple sources
Hospital	• Medicare Hospital Quality • Healthcare Cost and Utilization Project (HCUP)	• https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits • https://www.ahrq.gov/data/hcup/index.html
Multimodal	• "The Cancer Genome Atlas Program" (TCGA) - Genomics and Imaging	• https://www.cancer.gov/ccg/research/genome-sequencing/tcga

TABLE 14. A comprehensive list of current healthcare big data analytics deployed solutions.

Healthcare Deployed Solution	Key Application	Source Link
AICURE	AiCure is a New York-based artificial startup company funded by venture capitalists and the National Institutes of Health. It improves clinical trials and medication adherence with the help of machine learning and computer vision. Their tagline is "see, hear, and understand" to know how patients respond to treatment.	https://aicure.com/
Roam Analytics	Roam Analytics combines AI and machine learning to evaluate patient data and deliver real-time insights to healthcare providers. This Silicon Valley business is building healthcare software using a lot of data. It can be integrated into EHRs and other healthcare data sources to provide a complete patient health history. Data analytics should tailor patient care and improve healthcare quality.	https://roamanalytics.com/
DrChrono	DrChrono, a cloud-based EHR platform for medical offices, manages patient data, billing, and clinical workflows. AI and machine learning algorithms manage appointment reminders and prescription refills and deliver tailored healthcare advice on the platform.	https://www.drchrono.com/
Fathom Health	Fathom Health uses AI to detect medical claims fraud and inaccuracies. Machine learning aids health insurers in patient care and cost considerations. Their platform analyzes healthcare data to find cost savings and high-quality care.	https://www.fathomhealth.com/
QVENTUS	Qventus provides a predictive platform specifically designed for hospitals. By analyzing data and using predictive analysis, the platform provides insights that allow hospitals to take proactive measures to manage patient flow more efficiently and optimize resource allocation.	https://qventus.com/
Clinic On Go	Manages patient information and appointments	http://www.merlinoapps.com/
Doctor Buddy	It a one-in-all application that manages and stores both patient's health and medical records	https://docbuddy.com/
HealthTouch	Records and tracks key health statistics	https://healthtouch.net/
PlushCare	With PlushCare, you can get prescriptions and treatment for various ongoing and non-emergency conditions. Choose an appointment time, plug-in any insurance information, and get connected to a doctor	https://plushcare.com/
Doctor on Demand	Doctor on Demand service aims to serve patients on demand.	https://doctorondemand.com/
Isabel	A clinical decision support tool that provides healthcare professionals access to an online system that assists them to make an accurate diagnosis quickly. It consists of over 6000 diseases and conditions	https://www.isabelhealthcare.com/
Medisafe Meds	Provides constant tracking of your health progress regarding adding medications, getting reminders for taking and receiving pills	https://www.mymedisafe.com/
AmWell	The app analytics company App Annie has recently named AmWell - American Well as the world's most popular consumer telehealth app. It connects the patients and doctors by way of a remote connection	https://patients.amwell.com/
Practo	A comprehensive health application for scheduling doctor visits at clinics and hospitals, ordering medications, setting medication reminders, consulting doctors online, managing digital health data, and reading health recommendations	https://www.practo.com/
Portea	Assists patients in obtaining consultation and treatment from specialists/physicians. Additionally, it enables the delivery of medications online and the acquisition of medical opinions with hospital-quality nursing care at home	https://www.portea.com/
Livongo	A digital health platform that helps manage chronic conditions such as diabetes	https://www.livongo.com/
Proteus	Digital Health Develops ingestible sensors and digital health tools for medication management	https://www.labcenter.com/
Zocdoc	An online platform that allows patients to find and book appointments with healthcare providers	https://www.zocdoc.com/
Babylon Health	Health Provides AI-powered healthcare chatbots and telemedicine services	https://www.babylonhealth.com/
Suki.AI	Voice-enabled digital assistant for physicians to simplify clinical documentation	https://www.suki.ai/
2nd.MD	Offers second opinion consultations with top medical specialists	https://www.2nd.md/
Biofourmis	Develops AI-powered wearable devices to monitor and manage chronic conditions	https://www.biofourmis.com/
Health Catalyst	Provides healthcare analytics and data warehousing solutions to improve patient outcomes	https://www.healthcatalyst.com/
Verily	Uses data analytics and machine learning to develop healthcare solutions and clinical research	https://verily.com/
PathAI	Uses AI to analyze medical images and improve diagnostic accuracy for diseases such as cancer	https://www.pathai.com/
Komodo Health	Provides real-time insights and analytics for healthcare organizations to improve patient care	https://www.komodohealth.com/
AITIA	Uses machine learning to analyze patient data and develop personalized treatment plans	https://www.aitiabio.com/
Doc.ai	A decentralized AI platform for personalized health insights	https://doc.ai/
Augmedix	Provides virtual medical scribe services to help physicians with clinical documentation	https://augmedix.com/
K Health	Provides AI-powered symptom checker and telemedicine services	https://www.health.com/
Buoy Health	Provides an AI-powered digital health assistant for personalized health advice and triage	https://www.buoyhealth.com/
Aiva Health	Provides voice-powered patient engagement and remote monitoring solutions	https://www.aivahealth.com/
HealthTensor	Uses AI to automate medical documentation and improve clinical decision-making	https://withregard.com/
Eko Health	Develops AI-powered digital stethoscopes and ECG devices for remote patient monitoring	https://www.ekohealth.com/
Philips Healthcare	provides various healthcare technologies and services, including medical imaging, patient monitoring, and telehealth solutions.	https://www.patientcare.philips.com/
Veeva Systems	Provides cloud-based software solutions for the life sciences industry, including CRM, content management, and data analytics tools.	https://www.veeva.com/cn/
Siemens Healthineers	Develops and provides medical imaging and laboratory diagnostics equipment and healthcare IT solutions for hospitals and healthcare organizations.	https://www.siemens-healthineers.com/
NextGen Healthcare	Develops and provides EHR and practice management software for healthcare providers and population health and analytics tools.	https://www.nextgen.com/
Epic Systems	Provider of EHR and clinical management software systems for hospitals and healthcare organizations.	https://www.epic.com/
Athenahealth	Offers a suite of cloud-based software and services for medical practices, including EHR, practice management, and patient engagement tools.	https://www.athenahealth.com/
GE Healthcare	offers various healthcare technologies and services, including medical imaging, monitoring and diagnostic equipment, and data analytics solutions.	https://www.gehealthcare.com/
Allscripts	Provider of electronic health record (EHR) technology and practice management software for healthcare organizations.	https://www.allscripts.com/
Cerner	Develops and provides health information technology solutions, including EHR systems, for healthcare providers and patients.	https://www.cerner.com/

Some of the most common use cases for big data in healthcare include,

- Reducing healthcare cost
- Reducing hospital re-admissions
- Optimized workforce and workflows
- Real-time alerting
- Analysing Electronic Health Records (EHRs)
- Control data for public health research
- Efficient medical practices
- Efficient strategic planning
- Improving safety practices
- Better patient engagement
- Preventing unnecessary hospital and ER visits

One of the health systems tried various technologies with various vendors to reduce their debt and accurately predict the propensity to pay [151]. Healthcare.AI is one of the prominent solutions that provide a range of services for healthcare systems. These services include prediction models for previous payment behavior, payment balance, credit scores, and previous interactions, among others. Unlike other companies that rely on a single feature, Healthcare.AI offers a multi-faceted approach, enabling better analysis of healthcare data. The system has been found to be effective in increasing revenue, with one anonymous company achieving revenue of $2M after implementing Healthcare.AI. There are several success stories associated with Healthcare.AI, making it a smart and reliable artificial intelligence-based healthcare system. Relevant use cases and success stories have been documented in the literature [152], [153].

1) $2M revenue increase with healthcare.AI multi-feature predictive model
2) Better strategic planning with Healthcare.AI
3) Improved resource optimization
4) Early detection of acute myocardial infarction (AMI) mortality
5) Use of digital tools boosting the efficiency of African healthcare systems
6) AI helping to identify preventable health emergency [154]

The utilization of big data has the potential to transform the healthcare industry and create new opportunities for healthcare providers to improve patient outcomes, optimize operations, and reduce costs. Based on the research conducted by Groves et al. [40], it can be concluded that big data has opened up new pathways in healthcare, namely: i) Right Living, ii) Right Care, iii) Right Provider, iv) Right Value, and v) Right Innovation. These pathways represent a plethora of use cases that can be achieved through the application of big data. The emergence of these pathways demonstrates that big data is not just a buzzword, but rather a critical component in the advancement of healthcare.

VII. OPEN RESEARCH CHALLENGES - ANSWER TO RQ4

This section presents an overview of existing research gaps from a comprehensive standpoint, intending to provide future researchers with a valuable starting point for new investigations. Additionally, our objective is to elucidate the open research challenges that have emerged in recent years, with a focus on addressing the necessary solutions for the progression of healthcare.

A. MULTIMODALITY IN HEALTHCARE

Since the emergence of various types of data, the concept of multimodality has garnered significant attention [160]. Multimodal data offers a complementary and comprehensive source of information that cannot be adequately captured by a single modality alone. The utilization of multiple modalities has shown promising results in tasks such as natural language understanding, computer vision, audio processing, sentiment analysis, machine translation, and more [161]. The fusion of diverse modalities holds the potential for improved performance, robustness, and contextual understanding in numerous applications, including healthcare, multimedia analysis, autonomous driving, virtual reality, and human-computer interaction.

While the concept of multimodality has been explored from various perspectives, the application of multimodal approaches specifically within the healthcare domain [96], [138] is an important area for future research. In particular, addressing multimodality and multitasking, along with handling the challenges associated with Multimodal Imbalance data, requires attention. It is worth highlighting the issue of imbalanced data [162] within the realm of multimodal healthcare, as it remains a task that lacks standardized implementation. Additionally, ethical considerations represent another crucial aspect underlying every healthcare application or solution, warranting further exploration as part of the open research challenges in this field.

B. DATA MINING IN HEALTHCARE

One promising area as an open challenge in Healthcare is Data Mining. Data mining is a potentially fruitful topic that remains fraught with difficulties in the healthcare realm [107], [163]. Integrating and analyzing data from several disparate sources is a substantial obstacle in the field of healthcare data mining as we discussed in the multimodality challenge. The creation of scalable and effective algorithms for the processing of large-scale healthcare datasets is another barrier that must be overcome. The amount, velocity, and diversity of data in the healthcare industry are all continuing to expand at a rapid rate, which creates issues for processing and scalability. In order to effectively process and evaluate healthcare data in a timely way, data mining algorithms need to be able to effectively manage large amounts of data and make use of distributed and parallel computing frameworks.

Authors in their studies [164], [165] identified issues including assessing diagnostic and treatment record similarity in the domain of data mining. They [164] emphasize the need for similarity metrics to examine diagnostic and therapeutic data. This entails assessing data type, granularity, and patient record features. Next, they extract typical

diagnostic and treatment patterns from EMRs. The authors explain how to extract patterns from clustering findings. Clustering analysis helps find common patterns and trends in patients' diagnostic records. They also extract common treatment patterns from clustering data to identify recurring treatment techniques. The next step is forecasting typical diagnostic patterns. Data mining and prediction models are used to forecast diagnostic trends based on patient data. Healthcare providers can forecast patient diagnoses using past data and patterns. Additionally, they evaluate and prescribe usual therapy regimens. Using usual patterns, the authors evaluate the effectiveness and appropriateness of different treatment strategies. They also investigate the possibility of prescribing treatment strategies to doctors based on patient features and historical data.

Another important difficulty in the field of healthcare data mining involves the interpretability and explainability of the models [166], [167]. It is necessary to have models that are both transparent and interpretable in order to make important decisions in healthcare. These models must also be able to give explanations that are easy to grasp for any forecasts or recommendations they make. Increasing trust in the healthcare system, lowering barriers to clinical adoption, and enhancing decision-making are all possible outcomes of developing interpretable data mining models.

C. PRECISION MEDICINE

Precision medicine also known as personalized medicine as we discussed in subsection V-J (patient-centric healthcare application), precision medicine has been recorded as an open challenge for more than a century [168]. Precision medicine's fundamental principle involves customizing healthcare interventions based on an individual's genetic, behavioral, and environmental characteristics is not new, but it remains a problem and an ongoing study subject.

The 2015 US Precision Medicine Initiative [169] popularized ''precision medicine''. This effort promoted precision medicine via research, technology, and data exchange. Since then, genetics and other aspects of healthcare have been better understood.

Precision medicine further can be understood as an application of computational predictive modelling as defined in Table 11 where researchers aim to develop a predictive software for medicine such as in [170], authors put their efforts for prediction for COVID-19 medicine on BRICS countries as a case study using deep learning. However, practical precision medicine implementation remains difficult. This has numerous causes such as we summarized in the studies [171], [172], [173], [174]:

- Precision medicine requires the integration and analysis of genetic, clinical, lifestyle, and environmental data. Integrating and harmonizing numerous data sources and building effective analytical procedures to gain insights are difficult.
- Understanding the genetics of illnesses and treatment response has advanced, but applying these discoveries to clinical practice is difficult. Validation, standardization, and standards for genetic and molecular interpretation and clinical use are needed.
- Precision medicine uses personal and genetic data. Ensuring patient privacy, and informed permission, and resolving ethical and legal data sharing and usage problems are crucial.
- Implementing precision medicine across varied populations and healthcare settings raises concerns about access, cost, and healthcare inequities.
- Healthcare practitioners need the right skills to incorporate precision medicine into clinical practice. Training programs and educational activities must educate healthcare practitioners about sophisticated technology and individualized methods.

Precision medicine has immense potential to improve healthcare, but it takes research, cooperation, and innovation to make it widely available. The discipline is evolving to overcome these limitations and fully utilize precision medicine to improve patient outcomes.

D. ETHICAL CONSIDERATIONS AND BIAS MITIGATION

The analysis of multimodal healthcare data raises ethical concerns, particularly in relation to bias and discrimination. Biases can arise due to unbalanced data, under-representation of certain data categories, and biases introduced during data collection and labeling processes. Consequently, it is crucial for researchers to proactively address these ethical concerns and develop tools that can detect and mitigate biases in multimodal healthcare analytics, thereby promoting fair and equitable outcomes.

To ensure the integrity and fairness of multimodal healthcare data analysis, researchers must focus on several key areas. Firstly, it is essential to employ robust methodologies to identify biases present in the data, critically evaluating the data collection process and implementing appropriate measures to address and mitigate biases. Additionally, researchers should strive for transparency and explainability in their analyses, documenting data sources, preprocessing techniques, and modelling decisions, and providing clear explanations for the outputs of algorithms. By prioritizing ethical considerations and employing bias detection and mitigation techniques, researchers can contribute to the development of unbiased multimodal healthcare analytics, fostering trust and promoting equitable outcomes for all individuals involved.

E. LIMITATION OF PRE-TRAINED MODELS FOR MULTIMODAL HEALTHCARE

The adoption of NLP applications across various technological domains has witnessed significant growth, and the emergence of pre-trained models for healthcare [175] represents a recent and highly trending topic within the

TABLE 15. A comprehensive list of pre-trained models for healthcare.

Pre-Trained Model	Key Application	Reference
BioBERT:	A pre-trained biomedical language representation model based on the BERT architecture, designed to understand and process text data in the biomedical domain, such as electronic health records, medical literature, and clinical notes.	[144], [177]
MIMIC-CXR:	A pre-trained model for chest X-ray analysis, trained on a large dataset from the Medical Information Mart for Intensive Care (MIMIC) database, allowing for the detection of abnormalities and diseases.	[178]
X-Net	A pre-trained deep learning model focused on chest X-ray interpretation but more specifically designed to detect common thoracic diseases. X-Net is specifically designed for small data. It is not just a model but is also considered as an architecture on which several other applications are being developed.	[179], [180]
U-Net:	U-Net is one of many popular pre-trained models for semantic segmentation, particularly in medical imaging applications like identifying and segmenting organs, tumors, or lesions.	[181]
VQA-Med:	Visual Question Answering in the Medical domain (VQA-Med) is a pre-trained model that combines visual and textual information to answer questions related to medical images. It has been also applied in healthcare for tasks like answering clinical questions based on radio-logical or histopathology images.	[182]

healthcare sector. The concept of pre-trained or internet-trained models is derived from transfer learning [176]. While several pre-trained models have been developed for either text or image-based data, there is a noticeable absence of pre-trained models that cater specifically to the current needs of multimodal healthcare data. To emphasize the significance of these models, we have compiled a comprehensive list of existing pre-trained healthcare models. However, it is evident that there exists a considerable gap for future researchers to develop pre-trained models that can effectively handle multiple types of healthcare data. Although Table 15 showcases an example of the top five models observed in this domain, it is important to note that numerous other models based on standard architectures have been designed and developed.

Addressing the challenge of developing pre-trained models for multimodal healthcare represents one of several open research challenges that warrant attention in the future. By advancing the field of pre-trained models, researchers can greatly contribute to the effective analysis and utilization of multimodal healthcare data, thereby enhancing healthcare outcomes and driving innovation in the domain.

F. EXPLORATION ON BIG DATA ECOSYSTEM

The exploration of the Big Data Ecosystem stands as a pressing challenge that necessitates the attention of future researchers [183], [184]. In our investigation, we have specifically addressed this challenge through our first research question in section IV, where we have devised an elaborate framework for data-driven health organizations. This framework encompasses an examination of performance metrics, characteristics, search tools, data quality, and privacy challenges, as well as the demands and opportunities within the healthcare domain. However, we acknowledge that this research challenge merits further exploration in the future, primarily due to the increasing complexity of the ecosystem over time and the ongoing advancements in the field. It is imperative to delve deeper into this subject to enhance our understanding and uncover novel insights for the advancement of healthcare.

G. OTHER OPEN RESEARCH CHALLENGES

While conducting our review, it became evident that there are numerous additional research challenges within the field of healthcare analytics. We have compiled a list of these unexplored challenges, which can serve as valuable avenues for future researchers to explore and expand upon, advancing knowledge and innovations in the field. The following list highlights some of the most pressing and in-demand issues that require attention:

- *Trustworthy Healthcare:* trustworthy healthcare comprises various dimensions for healthcare delivery that improve confidence, and ethical behavior among patients, healthcare professionals, and other relevant parties. While trustworthy healthcare also has several challenges that need to be addressed in today's complex healthcare landscape. Challenges such as maintaining data privacy and security, addressing biases and inequalities, shared decision-making, promoting patient engagement, and ensuring transparency and accountability. Each challenge can also be considered with and without the incorporation of trustworthy healthcare.
- Scalability and Computational Efficiency: Developing scalable and computationally efficient algorithms and architectures to handle the increasing volume and complexity of healthcare data.
- *Data Privacy-Preserving Techniques:* Designing techniques and frameworks that protect the privacy and confidentiality of sensitive healthcare data during analysis and sharing.
- *Explainability and Interpretability:* Ensuring transparency and interpretability of analytics models to provide clinicians and stakeholders with understandable insights and justifications.
- *Predictive Analytics and Early Detection:* Leveraging advanced analytics to predict and detect healthcare events, diseases, or conditions at an early stage for timely intervention and improved outcomes.
- *Bias and Fairness in Analytics:* Addressing biases and ensuring fairness in healthcare analytics to avoid discriminatory outcomes and ensure equitable healthcare delivery.
- Validation and Generalization of Models: Validating and generalizing analytics models across diverse healthcare

settings to ensure their effectiveness and reliability in real-world applications.
- *Real-world Data Challenges:* Overcoming challenges related to the quality, heterogeneity, and integration of real-world healthcare data from multiple sources.
- *Adoption of Healthcare Solutions by Healthcare Professionals:* Exploring factors influencing the adoption and integration of healthcare analytics solutions by healthcare professionals, promoting their effective utilization in clinical practice.
- *Data-driven Clinical Guidelines and Protocols:* Developing data-driven approaches to inform the creation and update of clinical guidelines and protocols, ensuring evidence-based and personalized healthcare decision-making.

These open research challenges offer promising opportunities for future investigations, where researchers can contribute to the advancement of healthcare analytics and pave the way for improved healthcare delivery and patient outcomes.

VIII. DISCUSSION AND IMPLICATIONS OF RQ'S

In this section, we present a summary of the study results, highlighting the implications of each research question addressed in the survey.

The implication of our designed Research questions (RQs) contributes to several aspects of research for Big Data Healthcare. First, each research question focuses on the systematic literature review study, providing a clear guide for examining specific topics in a detailed manner. Our RQs also align with the research framework, methodology, and analysis methods ensuring a cohesive and rigorous study design. Moreover, by addressing knowledge gaps and adding existing theories or frameworks, our RQs establish the relevance and importance of our research. The results obtained from answering these RQs enable researchers to evaluate and understand study data, facilitating the development of relevant conclusions. Additionally, the Implications of our RQs extend to generating new knowledge and insights, contributing to the expansion of understanding in the field. Overall, our RQs shape the entire research process, encompassing the emphasis and organization of the study as well as the discoveries and contributions made.

The goal of **RQ-1** is to address the gap observed in the previous studies by synthesizing the extensive data ecosystem, explicitly focusing on every single component including but not limited to the healthcare life-cycle, further characteristics of big data, the search tools, additionally the role, and the need of big data in healthcare. The opportunities and challenges of BDA in healthcare. Each component is well described above in a particular section or subsection. We developed and presented a typical life-cycle in Figure 9 deployed in healthcare. Further, we developed and delivered a potential classification of BDA challenges in healthcare in Figure 11. Lastly, literature helped us design an extensive data-driven health organization framework depicted in Figure 10.

RQ-2 and RQ-3 aim to address the existing gap by highlighting the potential of promising application areas of BDA in healthcare and the authentic source of data, tools, and techniques, respectively.

To further elaborate on the discussion of **RQ-2**, after covering the history of health standard documentation, we also delved into the most promising application areas for BDA in healthcare, such as multimodal data analysis and fusion. Multimodal data analysis combines data from multiple sources, such as medical imaging, electronic health records, and genomics data, to better understand a patient's health status. The fusion concept refers to integrating data from different modalities, such as combining imaging and genomics data to improve diagnostic accuracy and treatment decisions.

We also discussed the benefits of natural language processing (NLP) in healthcare, such as extracting valuable information from unstructured clinical notes and text-based sources, enabling more accurate diagnosis and treatment decisions. Furthermore, we delved into the application of electronic health records (EHRs), which have become a critical data source for BDA in healthcare. We highlighted the potential benefits of using EHR data, such as improved patient outcomes, reduced costs, and increased efficiency of healthcare delivery.

Moving on to **RQ-3**, we discussed the different data sources that can be used for BDA in healthcare, such as clinical data from EHRs, medical imaging data, and sensor data from wearable devices. We also highlighted the different tools and techniques that can be used for BDA in healthcare, such as machine learning, data mining, and predictive analytics.

In addition to addressing these research questions, our survey comprehensively examined the open research challenges associated with BDA in healthcare. These challenges encompassed aspects such as data quality, privacy concerns, the need for interoperability and standardization, as well as the scarcity of skilled professionals. Furthermore, we discussed the potential opportunities and benefits of applying BDA in healthcare, including improved patient outcomes, personalized medicine, and the potential for cost savings.

The response to **Research Question 4 (RQ4)** brings attention to many ongoing research challenges within the realm of healthcare analytics. These problems cover wider areas such as multimodality, ethical considerations, bias mitigation, limitation of pre-trained healthcare models, the need for exploration of the Big Data Ecosystem, and other pertinent aspects. These listed challenges present significant opportunities for future scholars to investigate and make meaningful contributions to the progress of healthcare.

By summarizing the implications of each research question, this study provides valuable insights into the implications of BDA in the healthcare domain. These findings serve as a bridge for the above-listed contributions and a foundation

TABLE 16. Comprehensive citation analysis of our study - ["This table only mentions citation used in the main text, while there are certain references inside tables also."].

RQ1	RQ2	RQ3	RQ4
• Big Data Life-cycle [13], [15], [18], [19], [35], [39], [50], [78], [115], [128] • Characteristics of Big Data [44], [76], [108]. • Big Data Search Tools [117], [118], [123] • The Role of Big Data in Healthcare [59], [107], [124]–[126] • The Need of Big Data Analytics in Healthcare [38], [51], [101] • Opportunities for BDA in Healthcare • Challenges of BDA in healthcare [18], [25], [32], [35], [37], [43], [45], [46], [55], [57], [58], [64], [65], [68], [71], [80], [127] • Big Data Security and Privacy in Healthcare [18], [47], [54], [55], [60], [61], [68], [129]–[133]	• Health Standards Documentations [62], [134]–[137], [190]–[192] • Multimodal Big Data Analytics [56], [84], [86], [89], [90], [90], [100], [100], [104], [104], [111], [113], [113], [113], [138], [139], [193] • The concept of Data Fusion in Multimodal Data [33], [86], [88], [95] • Current Digital Healthcare System [63], [96], [103], [137], [140] • Big Data and MDL for Healthcare [49], [49], [56], [56], [70], [70], [94], [94], [97], [97], [141] • NLP in Healthcare [69], [77], [79], [91], [93], [142]–[144] • Health Education Promotion [16], [18], [27], [83], [105] • Block-Chain and Healthcare [75], [99], [102], [109], [112] • Electronic Health Record [7], [28], [30], [36], [42], [67], [77], [92] • Patient Centric Healthcare [24], [26], [31], [73], [116], [145]–[147], [195], [195], [196]	• Modeling Tools and Techniques [52], [77], [81], [89], [98], [101], [110], [148] • Big Data Datasets (particular references can be referred from the respective table) • Current Big Data Analytics Healthcare Solutions [53], [74], [82] • The Big Data Use Cases in Healthcare [40], [149]–[154]	• Multimodality in Healthcare [96], [138], [160]–[162] • Data Mining in Healthcare [107], [163]–[167] • Precision Medicine [168]–[174] • Ethical Consideration and Bias • Limitation of Pre-Trained Models in Healthcare [144], [175]–[182] • Exploration of Big Data Ecosystem [183], [184] • Several other Open Research Challenges

for future research endeavors, fostering the advancement of knowledge and innovations in this field.

A. RQ'S FINDINGS AND TAKEAWAY

In this particular subsection, we provide statistics on the key findings and takeaways from this systematic literature review.

Table 16 presents a statistical analysis of references used in our study. It indicates a substantial volume of scholarly study pertaining to several facets of big data in the healthcare domain. We present this table for the ease of future researchers; they can easily go through from cited articles with respect to the sub-domain. The allocation of references among the research inquiries underscores the extensive range of this discipline and the varied domains of exploration. Concluding the research, RQ1, which centers around the life cycle, features, and search tools of Big Data, exhibits a considerable volume of references. This highlights the significance of comprehending the fundamental elements of Big Data and the requisite instruments for proficiently handling and scrutinizing healthcare data. While RQ2 investigates the many uses of big data in the healthcare sector, it has a substantial volume of references. This justifies the increasing inclination to utilize big data in order to enhance healthcare procedures and achieve better outcomes. The references encompass a diverse array of applications, including health standards documentation, multimodal data analytics, natural language processing in healthcare, integration of blockchain technology, electronic health records, and patient-centric healthcare. On the other hand, RQ3) examines the technologies, datasets, and analytics employed in the domain of big data healthcare, and exhibits a moderate quantity of references in our study and this may be due to the limitation of our employed search strategy or bias towards the keywords. The smaller number of citations also entails that there is a substantial gap which can be further filled by future researchers. This statement implies that although there is a continuous investigation in this field, further investigation and advancement of modelling tools and methodologies, datasets, and analytics are still required to fully exploit the potential of Big Data in the healthcare sector. Lastly, RQ4 has a comparatively received average number of references in relation to the remaining research inquiries, since it explores the realm of emerging trends and difficulties in the context of big data healthcare. This suggests that further investigation is required in areas such as multimodality in healthcare, data mining, precision medicine, ethical issues and prejudice, limits of pre-trained models, the study of the big data ecosystem, and open research problems. Finally, we want to refer to the Yearly Google Trend plot as shown in Figure 13 of the decade (2013 to 2023) for the "Big Data and Healthcare" search. This plot shows the great interest and evolving popularity of the particular topic over the years. We derived this real-time data from Google Trend web analytics. This plot is combined and plotted in a year-wise pattern using a monthly trend. The lowest sum of monthly trend is observed for the year 2013, while the peak was observed in 2022 with a value of 892 and for the data of 9 months of this current year till the date of compiling this manuscript Sep 2023 the monthly value sum up to 621 normalized searches indicating sustained level of interest and suggested the further exploration for the mentioned limitations in the subsequent subsection.

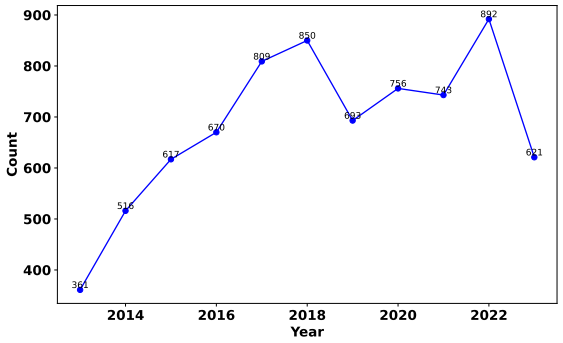

FIGURE 13. Yearly Google trend of the decade for "big data healthcare" (2013 - 2023).

In summary, the takeaway from this quantitative analysis of references is to serve as an indicator of the dynamic research environment within the field of Big Data in Healthcare. This underpins the significance of comprehending the underlying principles of Big Data, investigating its wide-ranging applications, creating efficient tools and analytics, and tackling the rising trends and difficulties within this domain. Ongoing research and collaborative efforts in these domains will play a significant role in harnessing the whole potential of Big Data for enhancing healthcare practices and optimizing patient outcomes.

B. LIMITATIONS OF OUR STUDY

We acknowledge that our study does not review the healthcare pivotal components such as vital signs, diseases, patient-specific problems, and hospital manageability. We feel that there is a need for a review paper in the future that covers Big Data Healthcare perspectives with respect to the mentioned areas. Further, we acknowledge that this study lacks non-academic credible sources such as industry reports, government agencies reports regarding statistics of healthcare, and published reports. We only cover scholarly articles published between the last decade (2013 to 2023). Future researchers are suggested to tackle these considerations for a new orientation of the study.

IX. CONCLUSION

The use of big data analytics has become increasingly relevant in healthcare over the past decade, as it offers the potential to revolutionize how medical professionals deliver care and manage health systems. Our comprehensive review study, which covered a wide range of published articles from 2013 to 2023, aimed to investigate the applications, implications, and impacts of big data frameworks in healthcare. Through this research, we identified novel research questions and conducted a thorough review to shed light on this important area of study.

Our findings demonstrate that the large-scale and complex nature of healthcare data presents significant challenges to big data analytics in healthcare. The data is often high-dimensional, noisy, and unstructured, which can make it difficult to draw meaningful conclusions. To overcome these challenges, it is necessary to develop reliable and trustworthy big data healthcare frameworks that prioritize patient privacy and data security. Furthermore, our study has highlighted the need to optimize big data frameworks to enhance patient outcomes, reduce costs, and improve overall quality of life. While there are still challenges to overcome, our study has provided valuable insights into the applications and implications of big data in healthcare. We believe that our research can guide healthcare professionals and researchers in developing effective and efficient big data frameworks that leverage the full potential of this technology. Additionally, our study has identified several opportunities for future research in this field, which could lead to further advancements and improvements in healthcare delivery and management.

DECLARATION OF INTEREST

The authors of this manuscript declare no conflict of interest.

REFERENCES

[1] I. Ahmad, Z. Asghar, T. Kumar, G. Li, A. Manzoor, K. Mikhaylov, S. A. Shah, M. Höyhtyä, J. Reponen, J. Huusko, and E. Harjula, "Emerging technologies for next generation remote health care and assisted living," *IEEE Access*, vol. 10, pp. 56094–56132, 2022.

[2] J. Santos-Pereira, L. Gruenwald, and J. Bernardino, "Top data mining tools for the healthcare industry," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 34, no. 8, pp. 4968–4982, Sep. 2022.

[3] H. Tao, M. Z. A. Bhuiyan, M. A. Rahman, G. Wang, T. Wang, M. M. Ahmed, and J. Li, "Economic perspective analysis of protecting big data security and privacy," *Future Gener. Comput. Syst.*, vol. 98, pp. 660–671, Sep. 2019.

[4] L. Nurgalieva, D. O'Callaghan, and G. Doherty, "Security and privacy of mHealth applications: A scoping review," *IEEE Access*, vol. 8, pp. 104247–104268, 2020.

[5] H. F. Ahmad, W. Rafique, R. U. Rasool, A. Alhumam, Z. Anwar, and J. Qadir, "Leveraging 6G, extended reality, and IoT big data analytics for healthcare: A review," *Comput. Sci. Rev.*, vol. 48, May 2023, Art. no. 100558.

[6] M. U. Hassan, M. H. Rehmani, and J. Chen, "Differential privacy techniques for cyber physical systems: A survey," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 1, pp. 746–789, 1st Quart., 2020.

[7] R. Agrawal and S. Prabakaran, "Big data in digital healthcare: Lessons learnt and recommendations for general practice," *Heredity*, vol. 124, no. 4, pp. 525–534, Apr. 2020.

[8] S. M. Shah and R. A. Khan, "Secondary use of electronic health record: Opportunities and challenges," *IEEE Access*, vol. 8, pp. 136947–136965, 2020.

[9] R. Nambiar, R. Bhardwaj, A. Sethi, and R. Vargheese, "A look at challenges and opportunities of big data analytics in healthcare," in *Proc. IEEE Int. Conf. Big Data*, Oct. 2013, pp. 17–22.

[10] 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.

[11] 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.

[12] 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. S31559.

[13] M. Islam, M. Hasan, X. Wang, H. Germack, and M. Noor-E-Alam, "A systematic review on healthcare analytics: Application and theoretical perspective of data mining," *Healthcare*, vol. 6, no. 2, p. 54, May 2018.

[14] 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.

[15] 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.

[16] A. Tandon, A. Dhir, A. K. M. N. Islam, and M. Mäntymäki, "Blockchain in healthcare: A systematic literature review, synthesizing framework and future research agenda," *Comput. Ind.*, vol. 122, Nov. 2020, Art. no. 103290.

[17] S. Imran, T. Mahmood, A. Morshed, and T. Sellis, "Big data analytics in healthcare—A systematic literature review and roadmap for practical implementation," *IEEE/CAA J. Autom. Sinica*, vol. 8, no. 1, pp. 1–22, Jan. 2021.

[18] 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.

[19] A. C. Ikegwu, H. F. Nweke, C. V. Anikwe, U. R. Alo, and O. R. Okonkwo, "Big data analytics for data-driven industry: A review of data sources, tools, challenges, solutions, and research directions," *Cluster Comput.*, vol. 25, no. 5, pp. 3343–3387, Oct. 2022.

[20] H. Zhang, S. Lee, Y. Lu, X. Yu, and H. Lu, "A survey on big data technologies and their applications to the metaverse: Past, current and future," *Mathematics*, vol. 11, no. 1, p. 96, Dec. 2022.

[21] D. Moher, A. Liberati, J. Tetzlaff, and D. G. Altman, "Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement," *Ann. Internal Med.*, vol. 151, no. 4, p. 264, Aug. 2009.

[22] M. J. Page et al., "PRISMA 2020 explanation and elaboration: Updated guidance and exemplars for reporting systematic reviews," *bmj*, vol. 372, 2021.

[23] B. Kitchenham, O. P. Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, "Systematic literature reviews in software engineering—A systematic literature review," *Inf. Softw. Technol.*, vol. 51, no. 1, pp. 7–15, 2009.

[24] N. V. Chawla and D. A. Davis, "Bringing big data to personalized healthcare: A patient-centered framework," *J. Gen. Internal Med.*, vol. 28, pp. 660–665, Sep. 2013.

[25] A. A. Tole, "Big data challenges," *Database Syst. J.*, vol. 4, no. 3, pp. 31–40, 2013.

[26] T. W. Kim, K. H. Park, S. H. Yi, and H. C. Kim, "A big data framework for u-healthcare systems utilizing vital signs," in *Proc. Int. Symp. Comput., Consum. Control*, Jun. 2014, pp. 494–497.

[27] D. Lupton, "Health promotion in the digital era: A critical commentary," *Health Promotion Int.*, vol. 30, no. 1, pp. 174–183, Mar. 2015.

[28] R. L. Street, L. Liu, N. J. Farber, Y. Chen, A. Calvitti, D. Zuest, M. T. Gabuzda, K. Bell, B. Gray, S. Rick, S. Ashfaq, and Z. Agha, "Provider interaction with the electronic health record: The effects on patient-centered communication in medical encounters," *Patient Educ. Counseling*, vol. 96, no. 3, pp. 315–319, Sep. 2014.

[29] A. Belle, R. Thiagarajan, S. Soroushmehr, F. Navidi, D. A. Beard, and K. Najarian, "Big data analytics in healthcare," *BioMed Res. Int.*, vol. 2015, Jul. 2015, Art. no. 370194.

[30] N. J. Farber, L. Liu, Y. Chen, A. Calvitti, R. L. Street, D. Zuest, K. Bell, M. Gabuzda, B. Gray, and S. Ashfaq, "EHR use and patient satisfaction: What we learned," *J. Family Pract.*, vol. 64, no. 11, pp. 687–696, 2015.

[31] A. R. M. Forkan, I. Khalil, A. Ibaida, and Z. Tari, "BDCaM: Big data for context-aware monitoring—A personalized knowledge discovery framework for assisted healthcare," *IEEE Trans. Cloud Comput.*, vol. 5, no. 4, pp. 628–641, Oct. 2017.

[32] X. Jin, B. W. Wah, X. Cheng, and Y. Wang, "Significance and challenges of big data research," *Big Data Res.*, vol. 2, no. 2, pp. 59–64, Jun. 2015.

[33] D. Lahat, T. Adali, and C. Jutten, "Multimodal data fusion: An overview of methods, challenges, and prospects," *Proc. IEEE*, vol. 103, no. 9, pp. 1449–1477, Sep. 2015.

[34] K.-S. Noh and D.-S. Lee, "Bigdata platform design and implementation model," *Indian J. Sci. Technol.*, vol. 8, no. 18, pp. 1–8, 2015.

[35] S. R. Sukumar, R. Natarajan, and R. K. Ferrell, "Quality of big data in health care," *Int. J. Health Care Quality Assurance*, vol. 28, no. 6, pp. 621–634, Jul. 2015.

[36] N. Weibel, S. Rick, C. Emmenegger, S. Ashfaq, A. Calvitti, and Z. Agha, "LAB-IN-A-BOX: Semi-automatic tracking of activity in the medical office," *Pers. Ubiquitous Comput.*, vol. 19, no. 2, pp. 317–334, Feb. 2015.

[37] 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.

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

[39] R. Fang, S. Pouyanfar, Y. Yang, S.-C. Chen, and S. S. Iyengar, "Computational health informatics in the big data age: A survey," *ACM Comput. Surv.*, vol. 49, no. 1, pp. 1–36, Mar. 2017.

[40] P. Groves, B. Kayyali, D. Knott, and S. Van Kuiken, "The 'big data' revolution in healthcare: Accelerating value and innovation," Center US Health Syst. Reform Bus. Technol. Office, 2016.

[41] P. Jain, M. Gyanchandani, and N. Khare, "Big data privacy: A technological perspective and review," *J. Big Data*, vol. 3, no. 1, pp. 1–25, Dec. 2016.

[42] A. E. W. Johnson, T. J. Pollard, L. Shen, L.-W.-H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "MIMIC-III, a freely accessible critical care database," *Scientific Data*, vol. 3, no. 1, pp. 1–9, May 2016.

[43] 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, 2016.

[44] R. Patgiri and A. Ahmed, "Big data: The V's of the game changer paradigm," in *Proc. IEEE 18th Int. Conf. High Perform. Comput. Commun., IEEE 14th Int. Conf. Smart City, IEEE 2nd Int. Conf. Data Sci. Syst. (HPCC/SmartCity/DSS)*, Dec. 2016, pp. 17–24.

[45] M. G. Kahn and C. Weng, "Clinical research informatics for big data and precision medicine," *Yearbook Med. Informat.*, vol. 25, no. 1, pp. 211–218, Aug. 2016.

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

[47] K. Abouelmehdi, A. Beni-Hssane, H. Khaloufi, and M. Saadi, "Big data security and privacy in healthcare: A review," *Proc. Comput. Sci.*, vol. 113, pp. 73–80, Jan. 2017.

[48] R. K. Behera, A. K. Sahoo, and C. Pradhan, "Big data analytics in real time—Technical challenges and its solutions," in *Proc. Int. Conf. Inf. Technol. (ICIT)*, Dec. 2017, pp. 30–35.

[49] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, "Disease prediction by machine learning over big data from healthcare communities," *IEEE Access*, vol. 5, pp. 8869–8879, 2017.

[50] P. Grover and A. K. Kar, "Big data analytics: A review on theoretical contributions and tools used in literature," *Global J. Flexible Syst. Manag.*, vol. 18, no. 3, pp. 203–229, Sep. 2017.

[51] C. Hardy, "Empathizing with patients: The role of interaction and narratives in providing better patient care," *Med., Health Care Philosophy*, vol. 20, no. 2, pp. 237–248, Jun. 2017.

[52] N. Jayanthi, B. V. Babu, and N. S. Rao, "Survey on clinical prediction models for diabetes prediction," *J. Big Data*, vol. 4, no. 1, pp. 1–15, Dec. 2017.

[53] Z. Lv, H. Song, P. Basanta-Val, A. Steed, and M. Jo, "Next-generation big data analytics: State of the art, challenges, and future research topics," *IEEE Trans. Ind. Informat.*, vol. 13, no. 4, pp. 1891–1899, Aug. 2017.

[54] K. Abouelmehdi, A. Beni-Hessane, and H. Khaloufi, "Big healthcare data: Preserving security and privacy," *J. Big Data*, vol. 5, no. 1, pp. 1–18, Dec. 2018.

[55] D. Vatsalan, Z. Sehili, P. Christen, and E. Rahm, "Privacy-preserving record linkage for big data: Current approaches and research challenges," in *Handbook of Big Data Technologies*. Berlin, Germany: Springer, 2017, pp. 851–895.

[56] T. Baltrusaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019.

[57] Y. Canbay, Y. Vural, and S. Sagiroglu, "Privacy preserving big data publishing," in *Proc. Int. Congr. Big Data, Deep Learn. Fighting Cyber Terrorism (IBIGDELFT)*, Dec. 2018, pp. 24–29.

[58] W. Dai, S. Wang, H. Xiong, and X. Jiang, "Privacy preserving federated big data analysis," in *Guide to Big Data Applications*. Berlin, Germany: Springer, 2018, pp. 49–82.

[59] D. Faggella. *Where Healthcare's Big Data Actually Comes From*. Accessed: 2018. [Online]. Available: https://www.techemergence.com/where-healthcares-bigdata-actually-comes-from

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

[61] H. Khaloufi, K. Abouelmehdi, A. Beni-hssane, and M. Saadi, "Security model for big healthcare data lifecycle," *Proc. Comput. Sci.*, vol. 141, pp. 294–301, Jan. 2018.

[62] C. S. Kruse, A. Stein, H. Thomas, and H. Kaur, "The use of electronic health records to support population health: A systematic review of the literature," *J. Med. Syst.*, vol. 42, no. 11, pp. 1–16, Nov. 2018.

[63] M. Mazzanti, E. Shirka, H. Gjergo, and E. Hasimi, "Imaging, health record, and artificial intelligence: Hype or hope?" *Current Cardiol. Rep.*, vol. 20, no. 6, pp. 1–9, Jun. 2018.

[64] P. R. M. Rao, S. M. Krishna, and A. P. S. Kumar, "Privacy preservation techniques in big data analytics: A survey," *J. Big Data*, vol. 5, no. 1, pp. 1–12, Dec. 2018.

[65] 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.

[66] N. D. Shah, E. W. Steyerberg, and D. M. Kent, "Big data and predictive analytics: Recalibrating expectations," *J. Amer. Med. Assoc.*, vol. 320, no. 1, pp. 27–28, Jul. 2018.

[67] R. L. Street, L. Liu, N. J. Farber, Y. Chen, A. Calvitti, N. Weibel, M. T. Gabuzda, K. Bell, B. Gray, S. Rick, S. Ashfaq, and Z. Agha, "Keystrokes, mouse clicks, and gazing at the computer: How physician interaction with the EHR affects patient participation," *J. Gen. Internal Med.*, vol. 33, no. 4, pp. 423–428, Apr. 2018.

[68] Z. Sun, K. D. Strang, and F. Pambel, "Privacy and security in the big data paradigm," *J. Comput. Inf. Syst.*, vol. 60, no. 2, pp. 146–155, 2018.

[69] S.-F. Sung, K. Chen, D. P. Wu, L.-C. Hung, Y.-H. Su, and Y.-H. Hu, "Applying natural language processing techniques to develop a task-specific EMR interface for timely stroke thrombolysis: A feasibility study," *Int. J. Med. Informat.*, vol. 112, pp. 149–157, Apr. 2018.

[70] M. Wainberg, D. Merico, A. Delong, and B. J. Frey, "Deep learning in biomedicine," *Nat. Biotechnol.*, vol. 36, no. 9, pp. 829–838, Oct. 2018.

[71] Y. Wang, L. Kung, and T. A. Byrd, "Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations," *Technol. Forecasting Social Change*, vol. 126, pp. 3–13, Jan. 2018.

[72] M. A. P. Chamikara, P. Bertok, D. Liu, S. Camtepe, and I. Khalil, "An efficient and scalable privacy preserving algorithm for big data and data streams," *Comput. Secur.*, vol. 87, Nov. 2019, Art. no. 101570.

[73] D. Cirillo and A. Valencia, "Big data analytics for personalized medicine," *Current Opinion Biotechnol.*, vol. 58, pp. 161–167, Aug. 2019.

[74] S. Dash, S. K. Shakyawar, M. Sharma, and S. Kaushik, "Big data in healthcare: Management, analysis and future prospects," *J. Big Data*, vol. 6, no. 1, pp. 1–25, Dec. 2019.

[75] A. Dwivedi, G. Srivastava, S. Dhar, and R. Singh, "A decentralized privacy-preserving healthcare blockchain for IoT," *Sensors*, vol. 19, no. 2, p. 326, Jan. 2019.

[76] N. Khan, A. Naim, M. R. Hussain, Q. N. Naveed, N. Ahmad, and S. Qamar, "The 51 V's of big data: Survey, technologies, characteristics, opportunities, issues and challenges," in *Proc. Int. Conf. Omni-Layer Intell. Syst.*, May 2019, pp. 19–24.

[77] S. Nuthakki, S. Neela, J. W. Gichoya, and S. Purkayastha, "Natural language processing of MIMIC-III clinical notes for identifying diagnosis and procedures with neural networks," 2019, *arXiv:1912.12397*.

[78] 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.

[79] Y. Si, J. Wang, H. Xu, and K. Roberts, "Enhancing clinical concept extraction with contextual embeddings," *J. Amer. Med. Inform. Assoc.*, vol. 26, no. 11, pp. 1297–1304, Nov. 2019.

[80] H.-Y. Tran and J. Hu, "Privacy-preserving big data analytics a comprehensive survey," *J. Parallel Distrib. Comput.*, vol. 134, pp. 207–218, Dec. 2019.

[81] B. Van Calster, L. Wynants, D. Timmerman, E. W. Steyerberg, and G. S. Collins, "Predictive analytics in health care: How can we know it works?" *J. Amer. Med. Inform. Assoc.*, vol. 26, no. 12, pp. 1651–1654, Dec. 2019.

[82] L. Wang and C. A. Alexander, "Big data analytics in healthcare systems," *Int. J. Math., Eng. Manag. Sci.*, vol. 4, no. 1, p. 17, 2019.

[83] Y. An, "Designing effective gamified learning experiences," *Int. J. Technol. Educ.*, vol. 3, no. 2, pp. 62–69, Feb. 2020.

[84] J. Chen and K. C. See, "Artificial intelligence for COVID-19: Rapid review," *J. Med. Internet Res.*, vol. 22, no. 10, Oct. 2020, Art. no. e21476.

[85] D. Ferrari, G. Guaraldi, F. Mandreoli, R. Martoglia, J. Milić, and P. Missier, "Data-driven vs knowledge-driven inference of health outcomes in the ageing population: A case study," in *Proc. Workshops 23rd Int. Conf. Extending Database Technol./23rd Int. Conf. Database Theory*, vol. 2578, 2020, pp. 1–6.

[86] S.-C. Huang, A. Pareek, S. Seyyedi, I. Banerjee, and M. P. Lungren, "Fusion of medical imaging and electronic health records using deep learning: A systematic review and implementation guidelines," *Npj Digit. Med.*, vol. 3, no. 1, pp. 1–9, Oct. 2020.

[87] M. I. Pramanik, R. Y. Lau, M. A. K. Azad, M. S. Hossain, M. K. H. Chowdhury, and B. Karmaker, "Healthcare informatics and analytics in big data," *Exp. Syst. Appl.*, vol. 152, Aug. 2020, Art. no. 113388.

[88] Y.-D. Zhang, Z. Dong, S.-H. Wang, X. Yu, X. Yao, Q. Zhou, H. Hu, M. Li, C. Jiménez-Mesa, J. Ramirez, F. J. Martinez, and J. M. Gorriz, "Advances in multimodal data fusion in neuroimaging: Overview, challenges, and novel orientation," *Inf. Fusion*, vol. 64, pp. 149–187, Dec. 2020.

[89] L. Biester, K. Matton, J. Rajendran, E. M. Provost, and R. Mihalcea, "Understanding the impact of COVID-19 on online mental health forums," *ACM Trans. Manag. Inf. Syst.*, vol. 12, no. 4, pp. 1–28, Dec. 2021.

[90] W. S. Brakefield, N. Ammar, and A. Shaban-Nejad, "UPHO: Leveraging an explainable multimodal big data analytics framework for COVID-19 surveillance and research," in *Proc. IEEE Int. Conf. Big Data*, Dec. 2021, pp. 5854–5858.

[91] Z. Kraljevic, T. Searle, A. Shek, L. Roguski, K. Noor, D. Bean, A. Mascio, L. Zhu, A. A. Folarin, A. Roberts, R. Bendayan, M. P. Richardson, R. Stewart, A. D. Shah, W. K. Wong, Z. Ibrahim, J. T. Teo, and R. J. B. Dobson, "Multi-domain clinical natural language processing with MedCAT: The medical concept annotation toolkit," *Artif. Intell. Med.*, vol. 117, Jul. 2021, Art. no. 102083.

[92] G. Kusuma, A. Kurniati, C. D. McInerney, M. Hall, C. P. Gale, and O. Johnson, "Process mining of disease trajectories in MIMIC-III: A case study," in *Proc. Int. Conf. Process Mining*. Cham, Switzerland: Springer, 2021, pp. 305–316.

[93] M. Rahimzadeh, A. Attar, and S. M. Sakhaei, "A fully automated deep learning-based network for detecting COVID-19 from a new and large lung CT scan dataset," *Biomed. Signal Process. Control*, vol. 68, Jul. 2021, Art. no. 102588.

[94] M. M. Rathore, S. A. Shah, D. Shukla, E. Bentafat, and S. Bakiras, "The role of AI, machine learning, and big data in digital twinning: A systematic literature review, challenges, and opportunities," *IEEE Access*, vol. 9, pp. 32030–32052, 2021.

[95] S. Wang, M. E. Celebi, Y.-D. Zhang, X. Yu, S. Lu, X. Yao, Q. Zhou, M.-G. Miguel, Y. Tian, J. M. Gorriz, and I. Tyukin, "Advances in data preprocessing for biomedical data fusion: An overview of the methods, challenges, and prospects," *Inf. Fusion*, vol. 76, pp. 376–421, Dec. 2021.

[96] P. Yang et al., "Multimodal wearable intelligence for dementia care in healthcare 4.0: A survey," *Inf. Syst. Frontiers*, 2021, doi: 10.1007/s10796-021-10163-3.

[97] S. Amal, L. Safarnejad, J. A. Omiye, I. Ghanzouri, J. H. Cabot, and E. G. Ross, "Use of multi-modal data and machine learning to improve cardiovascular disease care," *Frontiers Cardiovascular Med.*, vol. 9, Apr. 2022, Art. no. 840262.

[98] R. D. Riley, J. Ensor, K. I. E. Snell, F. E. Harrell, G. P. Martin, J. B. Reitsma, K. G. M. Moons, G. Collins, and M. van Smeden, "Calculating the sample size required for developing a clinical prediction model," *BMJ*, vol. 368, p. m441, Mar. 2020.

[99] K. Azbeg, O. Ouchetto, S. J. Andaloussi, and L. Fetjah, "A taxonomic review of the use of IoT and blockchain in healthcare applications," *IRBM*, vol. 43, no. 5, pp. 511–519, Oct. 2022.

[100] Y. Cao, P. Cao, H. Chen, K. M. Kochendorfer, A. B. Trotter, W. L. Galanter, P. M. Arnold, and R. K. Iyer, "Predicting ICU admissions for hospitalized COVID-19 patients with a factor graph-based model," in *Multimodal AI in Healthcare: A Paradigm Shift in Health Intelligence*. Berlin, Germany: Springer, 2022, pp. 245–256.

[101] A. A. H. de Hond et al., "Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: A scoping review," *NPJ Digit. Med.*, vol. 5, nos. 1–13, p. 2, 2022.
[102] E. Elgamal, W. Medhat, M. A. Elfatah, and N. Abdelbaki, "Blockchain in healthcare for achieving patients' privacy," in *Proc. 20th Learn. Technol. Conf.*, Jan. 2023, pp. 59–64.
[103] D. Ferrari, F. Mandreoli, F. Motta, and P. Missier, "Data-driven, AI-based clinical practice: Experiences, challenges, and research directions," in *Proc. 30th Italian Symp. Adv. Database Syst. (SEBD 2022)*. Newcastle Upon Tyne, U.K.: Newcastle Univ., 2022, pp. 1–12.
[104] A. Gupta and B. Srivastava, "A robust system to detect and explain public mask wearing behavior," in *Multimodal AI in Healthcare: A Paradigm Shift in Health Intelligence*. Berlin, Germany: Springer, 2022, pp. 155–169.
[105] H.-I. Hsu, C.-C. Liu, S. F. Yang, and H.-C. Chen, "A health promotion program for older adults (KABAN!): Effects on health literacy, quality of life, and emotions," *Educ. Gerontol.*, vol. 49, no. 8, pp. 639–656, Aug. 2023.
[106] S. Joshi, M. Sharma, R. P. Das, J. Rosak-Szyrocka, J. Zywiołek, K. Muduli, and M. Prasad, "Modeling conceptual framework for implementing barriers of AI in public healthcare for improving operational excellence: Experiences from developing countries," *Sustainability*, vol. 14, no. 18, p. 11698, 2022.
[107] M. Karatas, L. Eriskin, M. Deveci, D. Pamucar, and H. Garg, "Big data for healthcare Industry 4.0: Applications, challenges and future perspectives," *Exp. Syst. Appl.*, vol. 200, Aug. 2022, Art. no. 116912.
[108] M. Naeem, T. Jamal, J. Diaz-Martinez, S. A. Butt, N. Montesano, M. I. Tariq, E. D.-L. Hoz-Franco, and E. De-La-Hoz-Valdiris, "Trends and future perspective challenges in big data," in *Advances in Intelligent Data Analysis and Applications*. Arad, Romania: Springer, 2022, pp. 309–325.
[109] S. Srivastava, M. Pant, S. K. Jauhar, and A. K. Nagar, "Analyzing the prospects of blockchain in healthcare industry," *Comput. Math. Methods Med.*, vol. 2022, pp. 1–24, Dec. 2022.
[110] S. Warrier, E. M. Rutter, and K. B. Flores, "Multitask neural networks for predicting bladder pressure with time series data," *Biomed. Signal Process. Control*, vol. 72, Feb. 2022, Art. no. 103298.
[111] K. Zadorozhny, P. Thoral, P. Elbers, and G. Ciná, "Out-of-distribution detection for medical applications: Guidelines for practical evaluation," in *Multimodal AI in Healthcare: A Paradigm Shift in Health Intelligence*. Berlin, Germany: Springer, 2022, pp. 137–153.
[112] G. Zhang, Z. Yang, and W. Liu, "Blockchain-based privacy preserving e-health system for healthcare data in cloud," *Comput. Netw.*, vol. 203, Feb. 2022, Art. no. 108586.
[113] C. Zhang, X. Chu, L. Ma, Y. Zhu, Y. Wang, J. Wang, and J. Zhao, "M3Care: Learning with missing modalities in multimodal healthcare data," in *Proc. 28th ACM SIGKDD Conf. Knowl. Discovery Data Mining*, 2022, pp. 2418–2428.
[114] U. R. Hodeghatta and U. Nayak, "Big data analytics and future trends," in *Practical Business Analytics Using R and Python: Solve Business Problems Using a Data-Driven Approach*. Berlin, Germany: Springer, 2023, pp. 601–612.
[115] D. Khanna, N. Jindal, H. Singh, and P. S. Rana, "Applications and challenges in healthcare big data: A strategic review," *Current Med. Imag. Rev.*, vol. 19, no. 1, pp. 27–36, Jan. 2023.
[116] A. Panesar, "Future of precision healthcare," in *Precision Health and Artificial Intelligence: With Privacy, Ethics, Bias, Health Equity, Best Practices, and Case Studies*. Berlin, Germany: Springer, 2023, pp. 105–120.
[117] T. Grainger and T. Potter, *Solr in Action*. Shelter Island, NY, USA: Manning, 2014.
[118] Github. *Big Data Search Tools*. Accessed: Apr. 27, 2023. [Online]. Available: https://github.com/apache/lucene-solr
[119] *Sqoop*. Accessed: Apr. 27, 2023. [Online]. Available: http://www.sqoop.apache.org/
[120] *Flume*. Accessed: Apr. 27, 2023. [Online]. Available: https://www.flume.apache.org/
[121] *Hadoop*. Accessed: Apr. 27, 2023. [Online]. Available: http://www.hadoop.apache.org/
[122] S. Chantamunee, C. C. Fung, K. W. Wong, and C. Dumkeaw, "Knowledge discovery from Thai research articles by solr-based faceted search," in *Recent Advances in Information and Communication Technology*. Berlin, Germany: Springer, 2019, pp. 337–346.
[123] Statista. *Global Healthcare Data Volume*. Accessed: Apr. 27, 2023. [Online]. Available: https://www.statista.com/
[124] Frontiersin. *Better Patient Outcomes Through Mining of Biomedical Big Data*. Accessed: Apr. 27, 2023. [Online]. Available: https://www.frontiersin.org/
[125] RBCCM. *The Healthcare Data Explosion*. Accessed: Apr. 27, 2023. [Online]. Available: https://www.rbccm.com/
[126] Forbes. *The Skyrocketing Volume of Healthcare Data Makes Privacy Imperative*. Accessed: Apr. 27, 2023. [Online]. Available: https://www.forbes.com/
[127] B. B. Mehta and U. P. Rao, "Privacy preserving big data publishing: A scalable k-anonymization approach using MapReduce," *IET Softw.*, vol. 11, no. 5, pp. 271–276, 2017.
[128] J. B. Awotunde, A. E. Adeniyi, R. O. Ogundokun, G. J. Ajamu, and P. O. Adebayo, "MIoT-based big data analytics architecture, opportunities and challenges for enhanced telemedicine systems," in *Enhanced Telemedicine and e-Health: Advanced IoT Enabled Soft Computing Framework*, G. Marques, A. K. Bhoi, I. de la Torre Díez, and B. Garcia-Zapirain, Eds. Cham, Switzerland: Springer, 2021, pp. 199–220, doi: 10.1007/978-3-030-70111-6_10.
[129] R. Chauhan, H. Kaur, and V. Chang, "An optimized integrated framework of big data analytics managing security and privacy in healthcare data," *Wireless Pers. Commun.*, vol. 117, pp. 87–108, Feb. 2020.
[130] Z. Lv and L. Qiao, "Analysis of healthcare big data," *Future Gener. Comput. Syst.*, vol. 109, pp. 103–110, Aug. 2020.
[131] Y. Zhang, M. Qiu, C.-W. Tsai, M. M. Hassan, and A. Alamri, "Health-CPS: Healthcare cyber-physical system assisted by cloud and big data," *IEEE Syst. J.*, vol. 11, no. 1, pp. 88–95, Mar. 2017.
[132] Y. Yang, X. Zheng, W. Guo, X. Liu, and V. Chang, "Privacy-preserving fusion of IoT and big data for e-health," *Future Gener. Comput. Syst.*, vol. 86, pp. 1437–1455, Sep. 2018.
[133] A. Shahid, T.-A.-N. Nguyen, and M.-T. Kechadi, "Big data warehouse for healthcare-sensitive data applications," *Sensors*, vol. 21, no. 7, p. 2353, Mar. 2021.
[134] HL7. *Hl7 Documentation*. Accessed: Apr. 27, 2023. [Online]. Available: https://www.hl7.org/
[135] R. H. Dolin, L. Alschuler, S. Boyer, C. Beebe, F. M. Behlen, P. V. Biron, and A. Shabo, "HL7 clinical document architecture, release 2," *J. Amer. Med. Inform. Assoc.*, vol. 13, no. 1, pp. 30–39, Jan. 2006.
[136] H. Ji, S. Kim, S. Yi, H. Hwang, J.-W. Kim, and S. Yoo, "Converting clinical document architecture documents to the common data model for incorporating health information exchange data in observational health studies: CDA to CDM," *J. Biomed. Informat.*, vol. 107, Jul. 2020, Art. no. 103459.
[137] Z. Jiao, Y. Xiao, Y. Jin, X. Chen, and X. Huang, "Tianxia120: A multimodal medical data collection bioinformatic system for proactive health management in Internet of Medical Things," *J. Healthcare Eng.*, vol. 2020, pp. 1–11, Sep. 2020.
[138] A. Shaban-Nejad, M. Michalowski, and S. Bianco, "Multimodal artificial intelligence: Next wave of innovation in healthcare and medicine," in *Multimodal AI in Healthcare: A Paradigm Shift in Health Intelligence*. Berlin, Germany: Springer, 2022, pp. 1–9.
[139] Y. Zhao, Z. Qiao, C. Xiao, L. Glass, and J. Sun, "PyHealth: A Python library for health predictive models," 2021, *arXiv:2101.04209*.
[140] G. W. Leeson, "The growth, ageing and urbanisation of our world," *J. Population Ageing*, vol. 11, no. 2, pp. 107–115, Jun. 2018.
[141] K. Y. Ngiam and I. W. Khor, "Big data and machine learning algorithms for health-care delivery," *Lancet Oncol.*, vol. 20, no. 5, pp. 262–273, May 2019.
[142] R. Jackson, I. Kartoglu, C. Stringer, G. Gorrell, A. Roberts, X. Song, H. Wu, A. Agrawal, K. Lui, T. Groza, D. Lewsley, D. Northwood, A. Folarin, R. Stewart, and R. Dobson, "CogStack—Experiences of deploying integrated information retrieval and extraction services in a large national health service foundation trust hospital," *BMC Med. Informat. Decis. Making*, vol. 18, no. 1, pp. 1–13, Dec. 2018.
[143] T. Searle, Z. Kraljevic, R. Bendayan, D. Bean, and R. Dobson, "MedCATTrainer: A biomedical free text annotation interface with active learning and research use case specific customisation," 2019, *arXiv:1907.07322*.

[144] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. B. A. McDermott, "Publicly available clinical BERT embeddings," 2019, *arXiv:1904.03323*.

[145] M. Spruit and M. Lytras, "Applied data science in patient-centric healthcare: Adaptive analytic systems for empowering physicians and patients," *Telematics Inform.*, vol. 35, no. 4, pp. 643–653, 2018.

[146] W. Zheng, Y.-C.-J. Wu, and L. Chen, "Business intelligence for patient-centeredness: A systematic review," *Telematics Informat.*, vol. 35, no. 4, pp. 665–676, Jul. 2018.

[147] A. P. Singh, N. R. Pradhan, A. K. Luhach, S. Agnihotri, N. Z. Jhanjhi, S. Verma, U. Ghosh, and D. S. Roy, "A novel patient-centric architectural framework for blockchain-enabled healthcare applications," *IEEE Trans. Ind. Informat.*, vol. 17, no. 8, pp. 5779–5789, Aug. 2021.

[148] M. Squires, X. Tao, S. Elangovan, R. Gururajan, X. Zhou, and U. R. Acharya, "A novel genetic algorithm based system for the scheduling of medical treatments," *Exp. Syst. Appl.*, vol. 195, Jun. 2022, Art. no. 116464.

[149] M. Shahin, S. A. Peious, R. Sharma, M. Kaushik, S. B. Yahia, S. A. Shah, and D. Draheim, "Big data analytics in association rule mining: A systematic literature review," in *Proc. 3rd Int. Conf. Big Data Eng. Technol. (BDET)*, Jan. 2021, pp. 40–49.

[150] S. Sakr and A. Elgammal, "Towards a comprehensive data analytics framework for smart healthcare services," *Big Data Res.*, vol. 4, pp. 44–58, Jun. 2016.

[151] A. Chandra, B. Handel, and J. Schwartzstein, "Behavioral economics and health-care markets," in *Handbook of Behavioral Economics: Applications and Foundations 1*, vol. 2. Amsterdam, The Netherlands: Elsevier, 2019, pp. 459–502.

[152] Healthcare.AI(a). *Innovative Use Cases by Healthcare.AI*. Accessed: Apr. 27, 2023. [Online]. Available: https://healthcare.ai/level/use-cases/

[153] Healthcare.AI(b). *H20 Driverless AI—Healthcare*. Accessed: Apr. 27, 2023. [Online]. Available: https://h2o.ai/solutions/industry/health/

[154] Mckinsey. *Innovative Use Cases by Healthcare.AI*. Accessed: Apr. 27, 2023. [Online]. Available: https://www.mckinsey.com/industries/healthcare/our-insights

[155] H. Harutyunyan, H. Khachatrian, D. C. Kale, G. Ver Steeg, and A. Galstyan, "Multitask learning and benchmarking with clinical time series data," *Sci. Data*, vol. 6, no. 1, pp. 1–18, Jun. 2019.

[156] S. Purushotham, C. Meng, Z. Che, and Y. Liu, "Benchmarking deep learning models on large healthcare datasets," *J. Biomed. Informat.*, vol. 83, pp. 112–134, Jul. 2018.

[157] H. Cai, Z. Yuan, Y. Gao, S. Sun, N. Li, F. Tian, H. Xiao, J. Li, Z. Yang, X. Li, and Q. Zhao, "A multi-modal open dataset for mental-disorder analysis," *Sci. Data*, vol. 9, no. 1, p. 178, Apr. 2022.

[158] T. J. Pollard, A. E. W. Johnson, J. D. Raffa, L. A. Celi, R. G. Mark, and O. Badawi, "The eICU collaborative research database, a freely available multi-center database for critical care research," *Sci. Data*, vol. 5, no. 1, pp. 1–13, Sep. 2018.

[159] A. Dispenzieri, J. Zonder, J. Hoffman, S. W. Wong, M. Liedtke, R. Abonour, A. D'Souza, C. Lee, S. Cote, R. Potluri, E. Ammann, N. Tran, A. Lam, and S. Nair, "Real-world treatment patterns, costs, and outcomes in patients with AL amyloidosis: Analysis of the optum EHR and commercial claims databases," *Amyloid*, vol. 30, no. 2, pp. 161–168, Apr. 2023.

[160] C. Jewitt, J. Bezemer, and K. O'Halloran, *Introducing Multimodality*. Evanston, IL, USA: Routledge, 2016.

[161] K. Bayoudh, R. Knani, F. Hamdaoui, and A. Mtibaa, "A survey on deep multimodal learning for computer vision: Advances, trends, applications, and datasets," *Vis. Comput.*, vol. 38, no. 8, pp. 2939–2970, 2021.

[162] M. S. Santos, P. H. Abreu, N. Japkowicz, A. Fernández, and J. Santos, "A unifying view of class overlap and imbalance: Key concepts, multi-view panorama, and open avenues for research," *Inf. Fusion*, vol. 89, pp. 228–253, Jan. 2023.

[163] M. Varshney, B. Bhushan, and A. B. Haque, "Big data analytics and data mining for healthcare informatics (HCI)," in *Multimedia Technologies in the Internet of Things Environment*, vol. 3. Berlin, Germany: Springer, 2022, pp. 167–195.

[164] C. Guo and J. Chen, "Big data analytics in healthcare," in *Knowledge Technology and Systems: Toward Establishing Knowledge Systems Science*. Berlin, Germany: Springer, 2023, pp. 27–70.

[165] N. Jothi and W. Husain, "Data mining in healthcare—A review," *Proc. Comput. Sci.*, vol. 72, pp. 306–313, Jan. 2015.

[166] A. Vellido, "The importance of interpretability and visualization in machine learning for applications in medicine and health care," *Neural Comput. Appl.*, vol. 32, no. 24, pp. 18069–18083, Dec. 2020.

[167] P. J. G. Lisboa, S. Saralajew, A. Vellido, R. Fernández-Domenech, and T. Villmann, "The coming of age of interpretable and explainable machine learning models," *Neurocomputing*, vol. 535, pp. 25–39, May 2023.

[168] R. Hodson, "Precision medicine," *Nature*, vol. 537, no. 7619, p. 49, 2016.

[169] F. S. Collins and H. Varmus, "A new initiative on precision medicine," *New England J. Med.*, vol. 372, no. 9, pp. 793–795, Feb. 2015.

[170] Q. Zhao and Z. Zheng, "Computational and mathematical methods in medicine prediction of COVID-19 in BRICS countries: An integrated deep learning model of CEEMDAN-R-ILSTM-Elman," *Comput. Math. Methods Med.*, vol. 2022, pp. 1–34, Apr. 2022.

[171] T. Hulsen, S. S. Jamuar, A. R. Moody, J. H. Karnes, O. Varga, S. Hedensted, R. Spreafico, D. A. Hafler, and E. F. McKinney, "From big data to precision medicine," *Frontiers Med.*, vol. 6, p. 34, Mar. 2019.

[172] B. Koopman, T. Wright, N. Omer, V. McCabe, and G. Zuccon, "Precision medicine search for paediatric oncology," in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2021, pp. 2536–2540.

[173] E. Wong, N. Bertin, M. Hebrard, R. Tirado-Magallanes, C. Bellis, W. K. Lim, C. Y. Chua, P. M. L. Tong, R. Chua, and K. Mak, "The Singapore national precision medicine strategy," *Nature Genet.*, vol. 55, no. 2, pp. 178–186, 2023.

[174] E. O. Aboagye, T. D. Barwick, and U. Haberkorn, "Radiotheranostics in oncology: Making precision medicine possible," *CA, A Cancer J. Clinicians*, vol. 73, no. 3, pp. 255–274, May 2023.

[175] D. M. Korngiebel and S. D. Mooney, "Considering the possibilities and pitfalls of generative pre-trained transformer 3 (GPT-3) in healthcare delivery," *npj Digit. Med.*, vol. 4, no. 1, p. 93, Jun. 2021.

[176] H. Wang, J. Li, H. Wu, E. Hovy, and Y. Sun, "Pre-trained language models and their applications," *Engineering*, vol. 25, pp. 51–65, Jun. 2023.

[177] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, Feb. 2020.

[178] A. E. W. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-Y. Deng, R. G. Mark, and S. Horng, "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports," *Sci. Data*, vol. 6, no. 1, p. 317, Dec. 2019.

[179] J. Bullock, C. Cuesta-Lazaro, and A. Quera-Bofarull, "XNet: A convolutional neural network (CNN) implementation for medical X-ray image segmentation suitable for small datasets," in *Proc. SPIE*, vol. 10953, pp. 453–463, Mar. 2019.

[180] J. Lakshmi, "Deep learning on medical image analysis on COVID-19 X-ray dataset using an X-Net architecture," in *Deep Learning for Medical Applications With Unique Data*. Amsterdam, The Netherlands: Elsevier, 2022, pp. 71–106.

[181] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI*. Munich, Germany: Springer, 2015, pp. 234–241.

[182] A. B. Abacha, S. A. Hasan, V. V. Datla, J. Liu, D. Demner-Fushman, and H. Müller, "VQA-MED: Overview of the medical visual question answering task at ImageCLEF 2019," in *Proc. CLEF*, 2019, vol. 2, no. 6, pp. 1–11.

[183] S. I. H. Shah, V. Peristeras, and I. Magnisalis, "Government big data ecosystem: Definitions, types of data, actors, and roles and the impact in public administrations," *J. Data Inf. Qual.*, vol. 13, no. 2, pp. 1–25, Jun. 2021.

[184] M. Rashid, H. Singh, V. Goyal, S. A. Parah, and A. R. Wani, "Big data based hybrid machine learning model for improving performance of medical Internet of Things data in healthcare systems," in *Healthcare Paradigms in the Internet of Things Ecosystem*. Amsterdam, The Netherlands: Elsevier, 2021, pp. 47–62.

[185] M. K. Hassan, A. I. El Desouky, S. M. Elghamrawy, and A. M. Sarhan, "Big data challenges and opportunities in healthcare informatics and smart hospitals," in *Security in Smart Cities: Models, Applications, and Challenges*, A. E. Hassanien, M. Elhoseny, S. H. Ahmed, and A. K. Singh, Eds. Cham, Switzerland: Springer, 2019, pp. 3–26, doi: 10.1007/978-3-030-01560-2_1.

[186] M. Manoochehri, *Data Just Right: Introduction to Large-Scale Data & Analytics*. Reading, MA, USA: Addison-Wesley, 2013.
[187] D. Wood, M. King, D. Landis, W. Courtney, R. Wang, R. Kelly, J. A. Turner, and V. D. Calhoun, "Harnessing modern web application technology to create intuitive and efficient data visualization and sharing tools," *Frontiers Neuroinform.*, vol. 8, p. 71, Aug. 2014.
[188] S. Chintapalli, D. Dagit, B. Evans, R. Farivar, T. Graves, M. Holderbaugh, Z. Liu, K. Nusbaum, K. Patil, B. J. Peng, and P. Poulosky, "Benchmarking streaming computation engines: Storm, flink and spark streaming," in *Proc. IEEE Int. Parallel Distrib. Process. Symp. Workshops (IPDPSW)*, May 2016, pp. 1789–1792.
[189] F. Junqueira and B. Reed, *ZooKeeper: Distributed Process Coordination*. Sebastopol, CA, USA: O'Reilly Media, 2013.
[190] K. Bointner and G. Duftschmid, "HL7 template model and EN/ISO 13606 archetype object model—A comparison," *Stud. health Technol. Informat.*, vol. 150, p. 249, Jan. 2009.
[191] P. L. Elkin and S. H. Brown, "Fast healthcare interoperability resources (FHIR)," in *Terminology, Ontology Arond Their Implementations*. Berlin, Germany: Springer, 2023, pp. 511–565.
[192] I. Bossenko, K. Linna, G. Piho, and P. Ross, "Migration from HL7 clinical document architecture (CDA) to fast health interoperability resources (FHIR) in infectious disease information system of Estonia," in *Proc. 38th ACM/SIGAPP Symp. Appl. Comput.*, Mar. 2023, pp. 882–885.
[193] A. Sufyan, M. Imran, S. A. Shah, H. Shahwani, and A. A. Wadood, "A novel multimodality anatomical image fusion method based on contrast and structure extraction," *Int. J. Imag. Syst. Technol.*, vol. 32, no. 1, pp. 324–342, Jan. 2022.
[194] W. C. Sleeman, R. Kapoor, and P. Ghosh, "Multimodal classification: Current landscape, taxonomy and future directions," *ACM Comput. Surv.*, vol. 55, no. 7, pp. 1–31, Jul. 2023.
[195] A. G. Chandini and P. I. Basarkod, "Patient centric pre-transaction signature verification assisted smart contract based blockchain for electronic healthcare records," *J. Ambient Intell. Humanized Comput.*, vol. 14, no. 4, pp. 4221–4235, Apr. 2023.
[196] A. N. Gohar, S. A. Abdelmawgoud, and M. S. Farhan, "A patient-centric healthcare framework reference architecture for better semantic interoperability based on blockchain, cloud, and IoT," *IEEE Access*, vol. 10, pp. 92137–92157, 2022.
[197] D. M. Low, L. Rumker, T. Talkar, J. Torous, G. Cecchi, and S. S. Ghosh, "Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on Reddit during COVID-19: Observational study," *J. Med. Internet Res.*, vol. 22, no. 10, Oct. 2020, Art. no. e22635.

AWAIS AHMED (Member, IEEE) received the bachelor's and master's degrees in computer science from FAST-NUCES, in 2016 and 2019, respectively. He is currently pursuing the Ph.D. degree with the University of Electronic Science and Technology of China (UESTC). He is on study leave from his position as a Lecturer with the Department of Computer Science, Muhammad Ali Jinnah University, Karachi. He was a Research Associate and an Instructor with FAST-NUCES. His research interests include big data healthcare, multimodal data, data science, machine learning, natural language processing, network security, and analytics.

RUI XI (Member, IEEE) received the M.S. and Ph.D. degrees in computer science from the University of Electronic Science Technology of China (UESTC), Chengdu, China, in 2014 and 2019, respectively. From July 2019 to July 2022, he was a Postdoctoral Researcher with Tsinghua University. He has been an Assistant Researcher with the School of Computer Science and Engineering, UESTC, since August 2022. His main research interests include the Internet of Things, wireless sensing, time-series analysis, and AI.

MENGSHU HOU (Member, IEEE) received the Ph.D. degree in computer application from the University of Electronic Science and Technology of China, Chengdu, in 2005. He is currently a Professor with the School of Computer Science and Engineering, University of Electronic Science and Technology of China. His research interests include distributed computing, databases, big data, and NLP.

SYED ATTIQUE SHAH (Senior Member, IEEE) received the Ph.D. degree from the Institute of Informatics, Istanbul Technical University, Istanbul, Turkey. During the Ph.D. degree, he studied as a Visiting Scholar with The University of Tokyo, Japan, National Chiao Tung University, Taiwan, and the Tallinn University of Technology, Estonia, where he completed the major content of his thesis. He was an Associate Professor and the Chairperson with the Department of Computer Science, BUITEMS, Quetta, Pakistan. He was also engaged as a Lecturer with the Data Systems Group, Institute of Computer Science, University of Tartu, Estonia. He is currently a Lecturer in smart computer systems with the School of Computing and Digital Technology, Birmingham City University, U.K. His research interests include big data analytics, the Internet of Things, machine learning, network security, and information management.

SUFIAN HAMEED (Member, IEEE) received the Ph.D. degree in networks and information security from the University of Gottingen, Germany. He is currently an Associate Professor with the Department of Computer Science, National University of Computer and Emerging Sciences (NUCES), Pakistan. He also leads IT Security Laboratories, NUCES. The research laboratory studies and teaches security problems and solutions for different types of information and communication paradigms. His research interests include network security, web security, mobile security and secure architectures, and protocols for the cloud and the IoT.