Revista de Informática Teórica e Aplicada - RITA - ISSN 2175-2745
Vol. 32, Num. 2 (2025) 83-93

RESEARCH

Big Data in Healthcare: Possibilities and Challenges - A Systematic Literature Review

Big Data na Area da Saúde: Possibilidades e Desafios - Uma Revisão Sistemática da Literatura

Luís M. Barata[1 2]*, João Louro[1], João Afonso[1]

Abstract: In today's world, the extensive use of Big Data emerges as a fundamental component in various sectors, and healthcare is no exception. This article explores the importance of using this data for the continuous improvement of patient care, providing them with a better quality of life. However, for this to happen, several challenges must be overcome, especially in data processing and analysis, and in this sector, it is crucial that their accuracy and integrity are not compromised, as they are related to life and death. In this study, we present a survey of technologies, possibilities, and challenges that arise in this field through a systematic review of various articles with cutting-edge approaches to the subject. Consequently, the advantages of using Big Data in healthcare will be highlighted, as well as the critical need to overcome the inherent challenges to achieve efficient and ethical implementation.

Keywords: Big Data — Health — Opportunities — Challenges

Resumo: No mundo atual, a utilização extensiva de *Big Data* surge como uma componente fundamental em vários setores, e a saúde não é exceção. Este artigo explora a importância da utilização desses dados para a melhoria contínua do atendimento ao paciente, proporcionando-lhe melhor qualidade de vida. No entanto, para que isso aconteça, vários desafios devem ser superados, especialmente no processamento e análise de dados, e neste setor é crucial que a sua precisão e integridade não sejam comprometidas, pois estão relacionados com a vida e a morte. Neste estudo, apresentamos um levantamento de tecnologias, possibilidades e desafios que surgem neste campo por meio de uma revisão sistemática de diversos artigos com abordagens de ponta sobre o tema. Consequentemente, serão destacadas as vantagens da utilização de *Big Data* na área da saúde, bem como a necessidade crítica de superar os desafios inerentes para alcançar uma implementação eficiente e ética.

Palavras-Chave: *Big Data* — Saúde — Oportunidades — Desafios

[1] *Instituto Politécnico de Castelo Branco (IPCB), Portugal*
[2] *Instituto de Telecomunicações (IT), Portugal*
***Corresponding author**: luis.barata@ipcb.pt

DOI: http://dx.doi.org/10.22456/2175-2745.143528 • **Received:** 24/10/2024 • **Accepted:** 15/03/2025

1. Introduction

The era of the digital revolution marks the replacement of analog technology with digital electronics. This era began in the 1980s and continues today [1]. This shift, along with the invention of the World Wide Web, has resulted in an exponentially growing amount of data year after year due to the popularization of mobile devices and browsers. Consequently, in the early 2000s, companies began to race towards new strategies for analyzing large amounts of unstructured information. This is because data is the most valuable resource for a company, and proper analysis of this data can reveal hidden patterns, enabling decision-making in various sectors [2].

This innovation occurred in 2006 with the emergence of the Hadoop platform, which allowed Big Data applications to run on a clustered platform [3]. This open-source software has continued to grow and is currently one of the most widely used in the Big Data field due to its ability to store and process large volumes of information at an extremely fast rate. As it is utilized by major companies such as Uber, Netflix, and X (formerly known as Twitter), it is important to explain how this platform functions and its significance for the future of healthcare. This framework has three main components [4]: Hadoop Distributed File System (HDFS) (storage unit), MapReduce (processing unit), and YARN (resource management unit).

HDFS is a cornerstone of large-scale data processing systems due to its ability to provide reliable, scalable, and efficient storage. As illustrated in Figure 1, HDFS consists of two primary components: the NameNode and the DataNodes [5]. The NameNode functions as the central organizer or "librarian," storing information (metadata) about how data is distributed, replicated, and whether each DataNode is operational. The DataNodes act as the "shelves," holding actual data blocks and handling read and write requests. By continuously monitoring the health of each DataNode, the NameNode can automatically adapt the cluster in real time, reallocating or decommissioning nodes to maintain reliability and performance. This architecture allows HDFS to scale transparently—adding more DataNodes is like adding additional shelves to accommodate an ever-growing library. In the event of a DataNode failure, redundant data copies ensure uninterrupted access, further demonstrating HDFS's resilience. Consequently, these design choices make HDFS especially adept at handling massive volumes of information by providing high throughput, fault tolerance, and flexibility for modern big data applications.

Figure 2 illustrates the functioning of MapReduce. Here, the input is divided into multiple parts for more efficient processing. Next, the key-value pairs are sorted, and finally, in the aggregation phase, the final output is obtained at the DataNodes and sent to the NameNodes [7].

YARN is the resource management layer, responsible for its efficient and fair allocation, ensuring that each application obtains the necessary resources without affecting other applications [9]. This process is illustrated in Figure 3.

As mentioned, the era of the digital revolution is still ongoing, meaning that evolution continues daily, and it is necessary to keep up with this progress in all areas of our daily lives. In the healthcare sector, there is an increasing awareness among the population about its importance, which implies that the amount of data on this topic is also growing, especially considering the rise of IoT devices. Big Data in healthcare lies in the ability to analyze large volumes of data by cross-referencing information to anticipate diseases and personalize treatments [11].

Figure 4 illustrates how data related to this topic is growing almost exponentially, suggesting the imminent need for the integration of these data from various sources.

However, the challenge lies in selecting relevant data and securely managing this type of sensitive information, ensuring patient privacy and data integrity. This study proposes a comprehensive review of the possibilities offered by Big Data in improving patient care, highlighting innovations and critical challenges faced by healthcare professionals and researchers. By outlining a solid methodology, clear objectives, and inclusion criteria, we aim to identify the most effective approaches for massive data analysis in the healthcare context.

The remaining sections of the article are structured as follows. In Section 2, the methodology, study objectives, and inclusion criteria used in the research are defined. The results of this systematic review are presented in Section 3 and discussed in Section 4. The document concludes in Section 5 with answers to the research questions and a brief summary of the key points addressed.

2. Research Metodology

This chapter aims to thoroughly analyze scientific studies that contain relevant information on the topic of our article (Big Data in Healthcare). By using well-defined criteria, we will address pertinent research questions on this subject. The investigation was conducted according to the PRISMA methodology - Preferred Reporting Items for Systematic Reviews and Meta-Analyses [12].

2.1 Research Questions

This systematic review is based on the following questions:

(RQ1) In what aspects can Big Data transform the healthcare sector?

(RQ2) What are the challenges that arise in the use of Big Data?

(RQ3) What methodologies are currently used to apply Big Data in healthcare?

2.2 Inclusion Criteria

The study analyzing the possibilities and challenges of applying Big Data techniques in healthcare was conducted using the following inclusion criteria: (1) Studies that identify advantages or challenges; (2) Studies that present concrete methods used in healthcare; (3) Studies written in English; (4) Studies with full-text availability; and (5) Studies published between 2015 and 2024, to capture the most recent trends and applications.

2.3 Research Strategy

The search terms for this systematic review — ”Big Data” AND ”Healthcare” AND ”Methodologies” — were selected based on key concepts central to the research questions. The authors identified these terms through an iterative process involving exploratory searches. The goal was to capture relevant studies without being overly broad or narrow. Boolean operators like AND were used to ensure that the search retrieved studies addressing all three concepts. The search string was tested in IEEE Xplore and ACM Digital Library, two platforms highly regarded in the fields of computing and technology, to confirm its effectiveness. Each study was then independently analyzed by the authors for relevant approaches and perspectives.

2.4 Extraction of Study Characteristics

Data were extracted from the identified studies using a predefined format: Reference, Year, Application Area, and Methodology suggested in the article. Table 1 presents the respective extracted information.

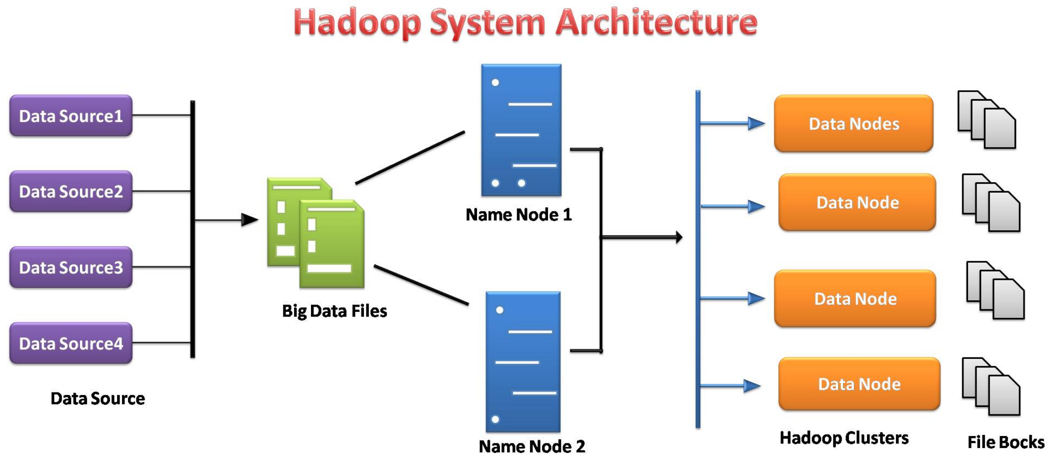

Figure 1. Hadoop Distributed File System [6]

3. Results

The Figure 5 illustrates the article selection process, where initially 75 publications were obtained from the two databases used in this systematic review IEEE and ACM. After removing duplicates (2), the titles and abstracts of each article were analyzed, and 45 were excluded because they were not directly related to the domain of our proposed article. Next, after analyzing the introduction and conclusion of the studies, 7 more were removed as they did not suggest methods or approaches. Finally, 6 studies were excluded because they did not differentiate from the others considering the inclusion criteria.

The 15 selected articles contain various approaches to the use of Big Data in healthcare and highlight numerous challenges and possibilities that will be discussed in the Discussion section. In terms of methodologies and algorithms used, the majority of the studies refer to the use of machine learning models in conjunction with the Hadoop platform and the MongoDB database. Other articles emphasize the amount of data generated, mainly by IoT devices, and the importance of keeping these data secure and easily accessible through Cloud Computing.

4. Discussion

The authors of the study [21] applied a practical approach to identify the benefits that the use of Big Data in healthcare could bring to the population. Initially, some challenges are highlighted for applying a functional methodology, namely the acquisition, processing, and analysis of data, as these come from various sources and consequently are in different formats, resulting in huge and complex datasets. The suggested methodology is carried out using a cardiovascular disease dataset composed of 70,000 rows and 12 attributes, requiring data preprocessing before being loaded into the RapidMiner software. The analysis is done using classification methods with decision trees to classify the presence/absence of disease. The authors conclude that there are various opportunities for healthcare if Big Data usage is implemented, such as:

- Improved preventive care by applying personalized treatments for each patient;
- Facilitated disease classification through cross-referencing information with other patients;
- Reduced healthcare costs as both diagnosis and medical treatment would be more accurate.

In [14], the authors highlight numerous opportunities that Big Data can unlock in this field, especially considering that a large portion of health information currently remains in static files. By leveraging real-time data instead, not only can responses be delivered more quickly, but the resulting insights become far more actionable, a critical advantage in urgent medical scenarios. Another possibility that could change the lives of not only patients but also healthcare professionals, who are often infected at work, is the use of real-time patient information to feed predictive models with this data. For example, it would be possible to identify clusters through temperature and location information, allowing the medical team to identify other patients who were in that area and isolate it. Applying cutting-edge methodologies in the use of Big Data would already be complicated in sectors where technology is highly developed, and in the healthcare sector, the problem worsens since it is the most lagging in technology use. In this sector, the challenges go beyond the 3 V's of Big Data (Volume, Velocity, and Variety), as the valence of the data is crucial, i.e., the ability to interlink information so that the data are more related. Finally, the authors mention some methods that can be used, namely: applying data analysis techniques such as segmentation and predictive modeling to extract useful insights from the data, and using technologies

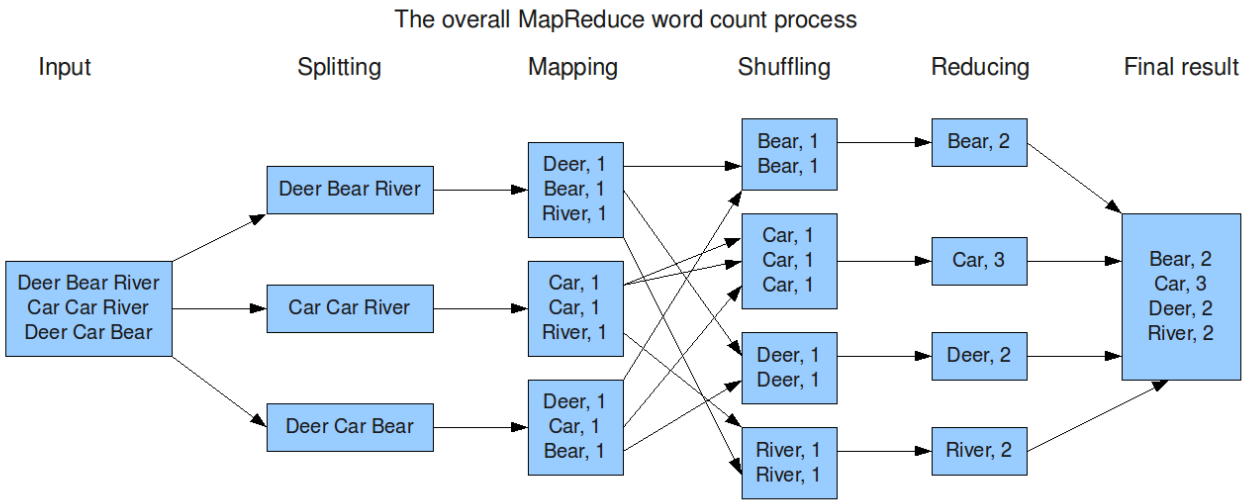

Figure 2. Hadoop MapReduce [8]

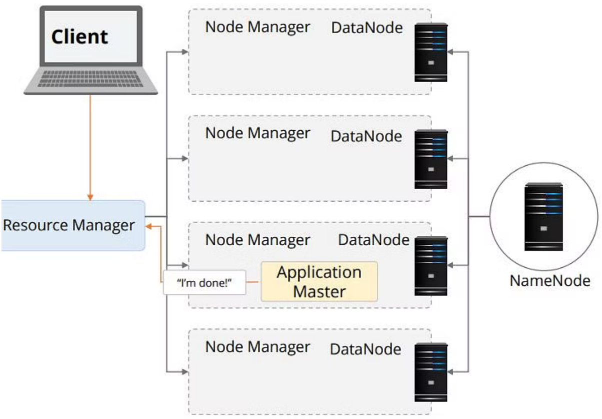

Figure 3. Hadoop YARN [10]

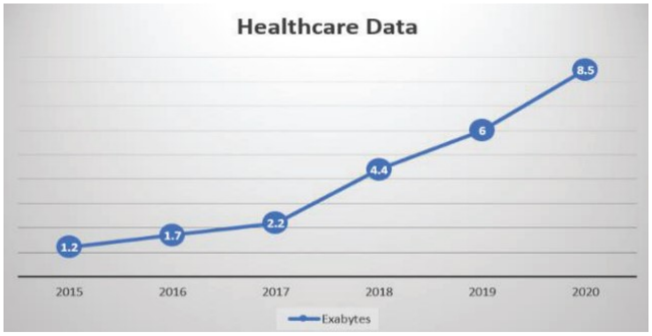

Figure 4. Growth of Healthcare Data [11]

like Hadoop, Cassandra, MongoDB, and cloud computing to facilitate efficient processing and analysis of large datasets.

According to the authors of the article [17], Big Data is still not widely used in the healthcare sector, mainly because data integrity cannot be minimally compromised. As highlighted in the Garbage-in-Garbage-out theory, if the data are inaccurate, they will produce incorrect outputs and consequently erroneous medical research. However, Big Data analysis can not only save costs and time but also detect diseases at an early stage, making the cure more effective and personalizing healthcare according to individual characteristics. Nevertheless, the challenges in this area are immense, with some of the most important being: data size, data with different formats, data security, and patient consent for data collection and sharing. The study concludes by stating that a balance needs to be found among all parties involved (government, patients, private institutions) to avoid problems such as pharmaceutical companies increasing prices based on analysis results showing an outbreak in a certain area, or the government hiding certain results for its benefit.

According to [25], the digitization of medical data has led to the exponential growth of heterogeneous data, meaning information that can appear in a variety of formats (structured, semi-structured, and unstructured) and originate from numerous sources such as IoT sensors, smartphones, clinical reports, and imaging systems. The sheer volume and diversity of this data make analytical tools indispensable for uncovering hidden patterns and supporting evidence-based decisions. Big Data Analytics (BDA) leverages techniques like data visualization, Machine Learning, and Artificial Intelligence to extract valuable insights. However, managing and storing these varied data types in a structured format for analysis is a significant challenge. This entire workflow can be divided into four stages: Data Acquisition, Data Storage, Data Analysis, and Information Management.

In the Data Acquisition stage, data collection, transmission, and preprocessing are crucial for integrating information

Table 1. Analyzed scientific articles and their respective application areas and methodologies used

Ref.	Year	Application Area	Suggested Methodologies
[13]	2015	Healthcare, Disease Prediction	Naive Bayes; Apache Mahout; Decision Trees
[14]	2016	Clinical sectors, Healthcare professionals	Predictive Models; Hadoop; MongoDB; Cloud Computing
[15]	2016	Clinic, Intensive Care Units	Hadoop; Multivariate Logistic Regression
[16]	2017	Hospital Environment	Hadoop, Cassandra, MongoDB, Big Data Analytics
[17]	2018	Health	Review
[18]	2019	Early Disease Diagnosis	Big Data Analytics; IoT; Hadoop; Machine Learning
[19]	2019	Hospital Environment	Big Data Analytics; IoT; Naive Bayes; Support Vector Machine
[20]	2020	Hospital Security	CyberSecurity; Blockchain; Denial of Service (DDoS) Attacks; SQL Vulnerabilities
[21]	2021	Hospitals, Health Centers	RapidMiner; Decision Trees
[22]	2021	Remote Consultations, Developing Countries	Cloud Computing, MongoDB, Mobile App
[23]	2022	Health	Big Data Analytics; IoT; Hadoop; Machine Learning
[24]	2022	COVID-19, Pandemics, Epidemics	Machine Learning; Supervised Learning; Hadoop; Databricks
[25]	2023	Personalized Medicine, Community Health	Big Data Analytics; Hadoop; Cloud Computing
[26]	2023	Medicine, Emergency Services	Big Data Analytics; Artificial Intelligence
[27]	2023	Clinics, Health Centers	Health Recommender System; Artificial Intelligence; CNN; Hadoop; Cassandra

from different sources. Once cleaned and standardized, the data moves into the Storage stage, where distributed architectures—such as Hadoop—are gaining popularity for their ability to handle massive datasets. The Analysis stage is where data truly gains value, yet traditional platforms often struggle to extract relevant insights from the sheer volume and complexity. Finally, the Information Management stage increasingly relies on Cloud Computing solutions, which reduce costs, improve patient care services, and facilitate real-time responses. Still, storing large quantities of unstructured information in the cloud carries its own risks—ranging from privacy exposure to integration complexity—necessitating robust preprocessing and security measures. When these challenges are effectively addressed, the possibilities for early disease detection, clinical decision support, and personalized medicine become practically limitless.

The authors of [15] highlight that data in this sector are growing in the order of petabytes, leading to many challenges such as data transfer, storage, and analysis, which cannot be resolved by relational databases as they cannot process this amount of information and can only handle structured data, while 80% of the information is unstructured. The main component for structuring Big Data is Hadoop, but Big Data encompasses not only efficient data storage and processing but also various Machine Learning techniques and cloud computing services that facilitate infrastructure maintenance. To implement this set of innovative technologies, the healthcare sector needs an infrastructure prepared to do so, which is not the case. For this reason, this sector faces several problems for the efficient use of Big Data, such as: no standards to integrate data from different sources, insufficient real-time processing, and patient privacy protection. Given all these challenges, researchers in this area develop several case studies. In one of these studies [28], a platform is developed to collect and analyze data from Intensive Care Units (ICUs). The experiment was conducted in a small ICU with 20 beds, generating approximately 1058 billion records annually. These data were used to predict mortality using multivariate logistic regression, achieving a success rate of 82%. The authors conclude that the growth potential of this sector is enormous, as people are currently willing to spend more money on healthcare, and clinics can take advantage of this to provide better healthcare to clients. However, an efficient architecture that uses Big Data and overcomes all the resulting challenges must be implemented.

According to [26], AI and Big Data are considered the major technological advances of this century, and using these two tools together can be used to extract insights quickly. In this study, the authors discuss how these techniques can be used together, as AI can analyze large volumes of data but needs the data to do so, taking advantage of Big Data Analytics techniques to obtain information more quickly. For example, through NLP (Natural Language Processing), AI can establish connections between data, detect and correct information problems, and automate tasks. It is demonstrated in this article that these methodologies can be applied in the

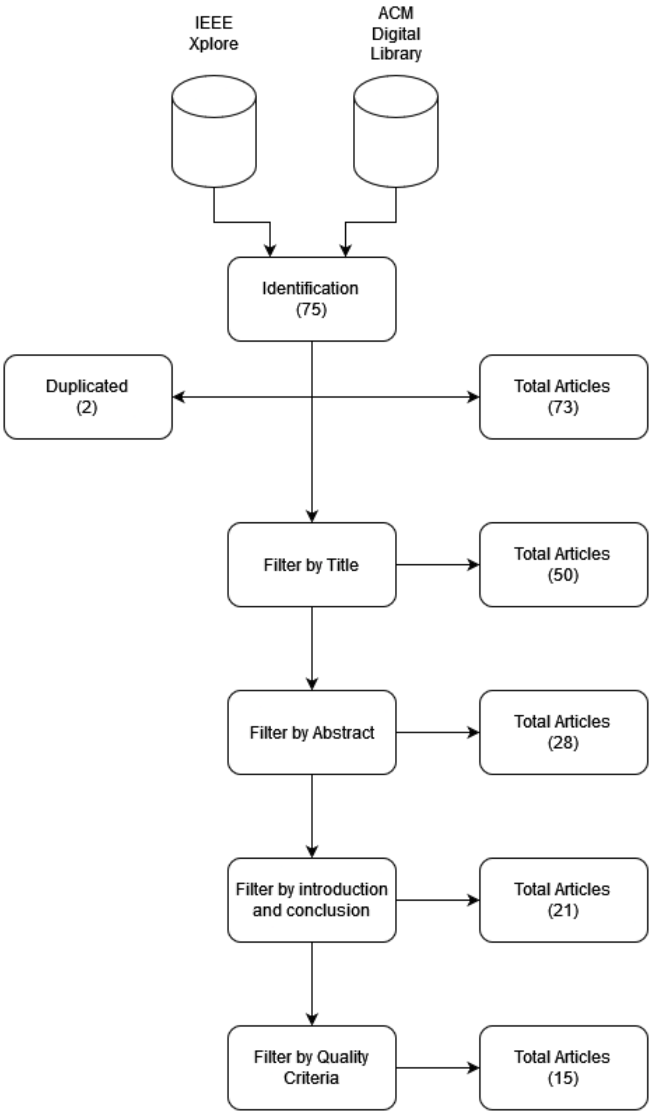

Figure 5. Flowchart of article selection.

real world, and when comparing medication doses suggested by doctors and AI, it is concluded that AI can impact reducing mortality and consequently assist a doctor or replicate their role in the future. Recent technological advances have the potential to transform healthcare as we know it today, turning it into more personalized and precise treatments for each patient, which can prevent emergency room overcrowding by using online treatments for less urgent patients.

In this article, the authors [27] propose a Health Recommender System (HRS) that uses Big Data and Artificial Intelligence to analyze large amounts of information, including social life information and medical records, to recommend treatments and medication alternatives. This system is created with a focus on ensuring the quality, reliability, authenticity, and privacy of the information. To achieve this, it is necessary to consider not only the 3 V's of Big Data but also Veracity, as we are talking about data that can affect people's lives and well-being. The suggested recommendation system uses a LeNET-type CNN to analyze large datasets and develop a useful health recommendation system, composed of several stages: initially, a patient's health record is created, and features are extracted through CNN training. From here, the system collects useful information from the database and performs sentiment analysis to understand patient opinions, thus generating reliable recommendations while preserving user privacy. Of course, for this algorithm to work, a large amount of quality information must be analyzed, and to obtain this, Big Data tools must be applied. Apache Hadoop was used for its large-scale data processing capability, and Cassandra was used as the database to facilitate the efficient management of massive data. LeNET is based on medical images such as X-rays and MRIs to extract relevant features important for recommending appropriate medical treatments or procedures. The authors conclude that by using the HRS, it is possible to analyze data such as medical conditions, allergies, and lifestyle to provide more personalized medical guidance. However, it is essential to monitor long-term behavior changes.

In the article [23], the authors emphasize the importance of IoT devices in generating Big Data, as these sensor-based gadgets, although contributing to improving people's lives in various areas such as healthcare, are also responsible for the increasing amount of information generated every second. The emergence of 5G networks resulted in high-speed connectivity, exponentially increasing the number of IoT devices used in all areas. To analyze this amount of information, estimated to be 175 Zettabytes of data by 2025, techniques such as Big Data analytics are needed to uncover hidden patterns and relationships between the data. Big Data Analytics extracts information used by Machine Learning algorithms to predict data using techniques such as clustering, classification, and statistics. The authors divide Big Data Analytics into four layers: Data storage, Data processing, Data query, and Data visualization, where different frameworks can be used in this workflow. The challenges faced by Big Data Analytics include the characteristics of the data (Volume, Velocity, Variety), as well as data storage, preprocessing, management, and integration. The authors discuss some frameworks, such as Hadoop and Spark, where the former stands out for its broad capabilities, and the latter is an alternative to Hadoop's MapReduce, offering better performance and support for different programming languages. Finally, it is concluded that although IoT is the largest source of Big Data, this information without analysis is completely useless, hence the importance of Big Data Analytics.

The authors of the article [20] focus on the importance of ensuring data integrity, as data is the most valuable asset of a company, making the management of this vast amount of information a complex task. Cybersecurity attacks on hospitals are now considered the most severe attacks because, in addition to this sector lagging in this area, the number of attacks increases yearly. It is highlighted that data integrity is more critical than availability because data can be tampered with, leading to patients receiving incorrect medications, causing fatalities. Healthcare organizations' storage systems are highly vulnerable to malware attacks, with an estimated medical record on the dark web being the second most valuable asset, ranging from $1 to $1000. The authors conducted a systematic review of the most commonly used methods to ensure data integrity in healthcare, emphasizing the use of Blockchain technology, where each transaction is recorded in blocks and chained sequentially. Multiple parties maintain identical copies of a record, making it extremely difficult to alter the data without being detected. Another technique used is a message authenticator, which is extremely useful in IoT devices in healthcare, ensuring that only authorized devices can send and receive data, with messages being encrypted, meaning they cannot be deciphered without the correct key. Despite this, several problems can affect data integrity; Blockchain is inadequate for handling large volumes of healthcare data, and since it implements a distributed network where each node has a copy of the transaction, it complicates data privacy management. Other categorized attacks include DDoS attacks, SQL vulnerabilities, and human errors, often caught in phishing attacks. The authors conclude that a more secure data integrity strategy must be implemented in the healthcare sector to avoid compromising human lives.

In the article [24], the authors discuss the importance of using effective analytical tools to analyze large datasets related to the COVID-19 pandemic, although the use of these tools can extend to other areas. Initially, an experiment with 40 participants (20 healthy, 20 sick) is conducted to diagnose the presence of depression using supervised learning techniques such as classification and linear regression. The results are positive, although they conclude that analyzing more complex data types, such as speech samples, requires more advanced machine learning techniques. Some of the machine learning tools mentioned in the study include Tensorflow, PyTorch, and Caffe, which are deep learning frameworks used to develop and train Artificial Intelligence models. These tools can be combined with technologies for processing large datasets in a distributed manner, such as MapReduce, Apache Spark, and Databricks. This methodology could have been applied in detecting COVID-19 infections, as the resulting graph from digitally tracking contacts between people can be extremely complex, making the use of common algorithms impractical. This study also addresses the growing importance of metaheuristic algorithms, used to select important features in datasets with hundreds or thousands of variables. These algorithms were widely used during the pandemic, particularly in classifying the presence of the virus in patients, efficiently allocating patients in hospitals, and distributing medical resources.

The authors of the article [22] primarily discuss the importance of cloud computing in healthcare systems. In addition to improving the quality and efficiency of services, it also facilitates the recording of information through remote consultations, mobile applications, or IoT devices. This approach would bring even more advantages to developing countries where medical services are limited. For example, according to World Health Organization (WHO), Bangladesh has only 5.26 doctors per 10,000 people. The use of cloud computing and digital technologies in healthcare (E-health) can overcome constraints such as the high cost of maintaining information systems, allowing the optimized use of limited medical resources. The authors created a cloud computing-based system, where through requirement analysis, they built a system for users to schedule appointments, make video calls, and receive medical prescriptions and real-time notifications. Building this system requires several tools, including HTML, CSS, JavaScript for the front-end, and Node.js and Express.js for back-end development, allowing the connection with MongoDB, responsible for storing and synchronizing patient data in real-time. The authors conclude, based on a usability study, that the population is ready and willing to use these systems in the future, given their advantages in providing more efficient healthcare.

The objective of the authors of [13] is to create a system for early disease identification using the Naive Bayes classification algorithm executed on Apache Mahout. Once the disease is identified based on users' symptoms, necessary treatments are suggested, which can reduce healthcare costs. It is estimated that in the USA, 1,000 people die daily due to medical errors, meaning healthcare is often inaccessible and inaccurate, revealing the lack of an electronic system to help patients obtain some preliminary information based on their symptoms. The proposed method is composed of several stages: initially, data are aggregated and preprocessed. The next stage involves applying mining algorithms to the selected data, namely the Naive Bayes algorithm to identify potential risk groups. Then, using decision trees, preventive measures and risk factors are identified. Next, the data are divided into training and test sets. Unlike traditional database models where data are stored in clusters and processed by queries, real-time processing is used in this case, analyzing data as it is transmitted to the server, allowing for faster decision-making. The authors conclude that the developed model can help predict a person's health status based on their symptoms and provide some care tips, improving healthcare efficiency.

In the article [18], the authors address the inherent challenges and possible solutions that can be applied due to the increased information from IoT devices. These devices, namely smartwatches and smartbands, can generate a large amount of relevant data and help diagnose diseases at an early stage. However, the challenge lies in the data's heterogeneity, meaning that to extract insights from the data, they must first be preprocessed, resulting in other problems such as information privacy. The authors mention numerous challenges that Big Data Analytics (BDA) faces, such as privacy and security, real-time processing, data structure and preprocessing, as well as data mining. Despite using machine learning algorithms, the challenges arise from the high speed at which data are generated and their heterogeneity. Currently, there are several tools used in the ecosystem for Big Data analysis, with some of the most common being: Apache Hadoop, Apache Spark, Cassandra, MapReduce, and HBase. Some proposed solutions in the literature are also discussed, highlighting the use of technologies such as Hadoop to process data in real-time or using data compression and aggregation techniques to simplify Big Data analysis. This method involves using machine learning algorithms to reduce the size of IoT data. Finally, the authors conclude that BDA is crucial for decision-making. However, more studies and new methods are needed to overcome all inherent challenges and make this system secure and efficient.

The authors of the article [19] use Big Data Analytics and machine learning in a case study, resulting in 95% accuracy in disease detection at a cost 90% lower than the local hospital in Bangladesh. The proposed system uses wearables like smartwatches, smart glasses, smart shoes, etc., to collect patient data. These data are temporarily stored for initial analysis. Then, the data are analyzed using machine learning algorithms to extract the most relevant features and predict diseases. However, the correct selection of the algorithm is necessary as different algorithms yield different results in predicting certain diseases. Finally, the results are presented to patients, who can consult specialists, find the nearest hospital, get medications, or schedule an appointment. This way, the doctors' work is also facilitated, ensuring secure access to patient records through fingerprint or facial recognition. The authors conclude that obtaining real-time data is a significant challenge, being the main limitation of the proposed experiment. However, in the future, more complex methods such as deep learning can be used to obtain new datasets and improve accuracy compared to Naive Bayes and Support Vector Machine methods.

In [16], it is mentioned that digitizing healthcare systems will enable the acquisition of more data sources, allowing the discovery of new patterns and signals that would otherwise remain hidden. It is also highlighted that one of the main objectives of applying Big Data in healthcare systems is the possibility of incorporating social and behavioral areas. With this, it would be possible to identify patterns in the population and improve disease prevention and provide more effective diagnostic mechanisms. This would require creating data analysis departments, necessitating the acquisition of spaces to store the information. To mitigate these problems, the authors developed a methodology that follows rules such as data heterogeneity, the use of distributed computing, and open-source software. This methodology operates in three layers: the first layer is storage, consisting of Data Lakes using HDFS; the second layer is the data layer, using services like Cassandra for data distribution, PostGIS for geospatial data, and PostgreSQL and MongoDB; the third layer is the cognitive layer, allowing visualization. With this methodology, the authors believe that this system could be implemented in almost any environment, as it is flexible, meaning it is not necessary to implement all services, and it is possible to add new services to the existing ones.

Considering all the information gathered during this systematic review, a diagram was created to demonstrate the different stages required for the efficient implementation of Big Data usage in healthcare. To implement an efficient solution in this area, several stages are necessary, each containing numerous challenges to be overcome. The diagram comprises six main components: Information; Data Aquisition; Cloud Computing; Data Processing; Storage; and Big Data Analytics, providing insights that neither doctors nor patients would previously know.

Based on the diagram shown in Figure 6, the challenges arising from each stage will be identified. Implementing a solution of this type requires collecting as much information as possible from various sources such as hospitals and wearables. In this phase, challenges include ensuring data quality and integrity. The massive volume of generated data causes numerous problems in the Data Aquisition phase, namely preprocessing data in huge and complex datasets where information is constantly generated at a high rate and in different formats. Additionally, it can compromise the privacy of the collected patient information. In this phase, encryption techniques should be applied, adding complexity to data processing.

In the Cloud Computing phase, the main challenge is ensuring data security and integrity. Although blockchain is considered secure for ensuring data authenticity, all employees must be trained in security awareness to avoid phishing attacks. The next phase (Data Processing) requires adjusting and optimizing clusters to process large volumes of data, ensuring the efficiency of distributed processing in a Hadoop environment. In the storage phase, the primary challenge is ensuring data scalability and availability to handle the increasing amount of stored data.

Finally, in Big Data Analytics, it is necessary to select and adjust the machine learning algorithm that is most accurate for the specific problem, considering data complexity and specific analytical objectives.

At the end of the implementation, it will be essential to pay attention to each of the different phases to ensure that the implemented solution meets all technical and ethical challenges. Once these issues are guaranteed, the immense advantages offered by implementing Big Data in healthcare can be enjoyed. The main possibilities identified from the review were:

- Improved Patient Care Quality: Doctors can develop personalized treatments tailored to each patient's individual needs.
- Early Disease Detection: Big Data analysis can identify patterns in complex datasets, allowing for early problem detection.
- Cost and Time Reduction: Insights from Big Data Analytics will make both diagnosis and treatment more accurate and effective.

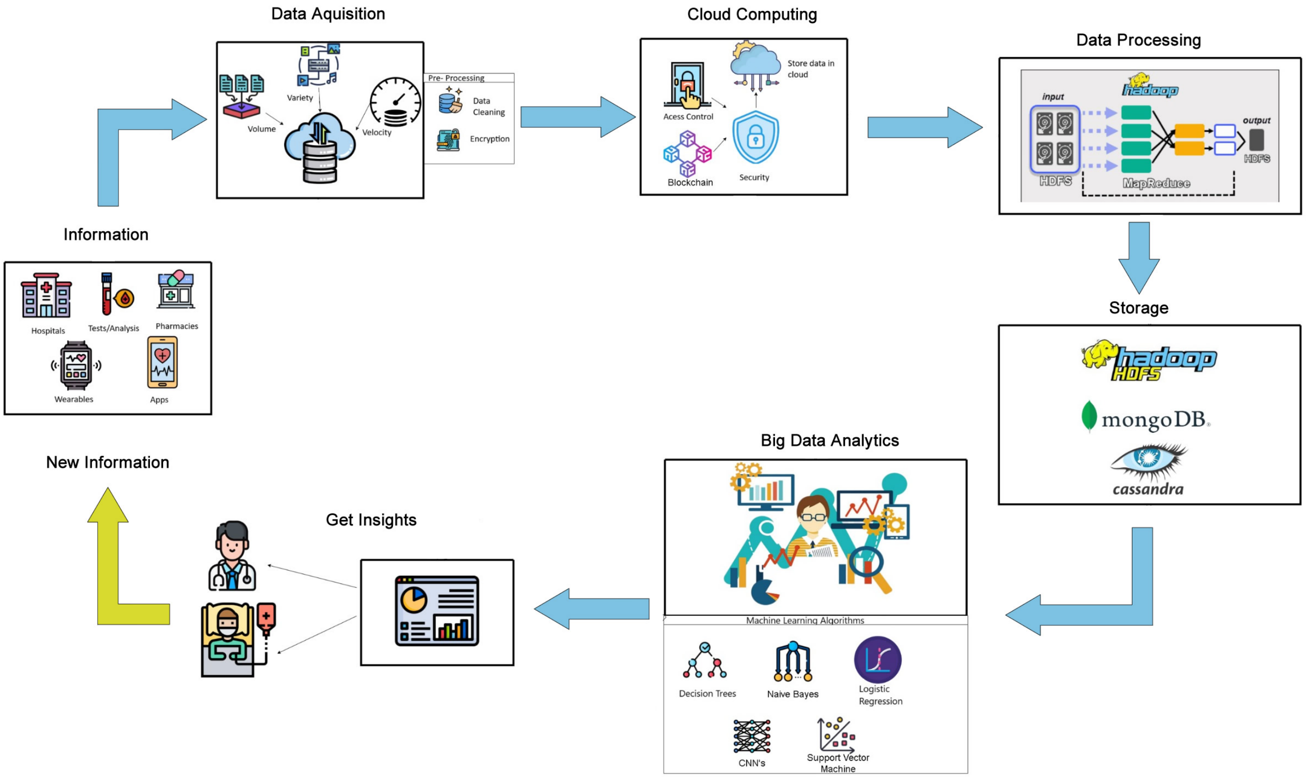

Figure 6. Overview of a typical Big Data solution, illustrating the progression from data acquisition and cloud-based storage to data processing, analytics, and the generation of actionable insights.

- Advanced Medical Research: Access to datasets with health data will allow researchers to perform advanced analyses and discover new treatments and therapies.

5. Conclusion

In this article, we conducted a systematic review of the possibilities and challenges for improving patient care. Initially, through research in scientific repositories, only 15 articles were considered relevant, as many articles in this area contain similar information, and for this reason, articles that also addressed practical methodologies were selected. Achieving a successful implementation requires combining various technologies and methods, which entails overcoming numerous challenges.

The main conclusions drawn are as follows:

(RQ1) In what aspects can Big Data transform the healthcare sector? Big Data provides data-based evidence to support not only clinical but also administrative decision-making, resulting in more efficient healthcare systems and, consequently, a better quality of life for the population.

(RQ2) What are the challenges that arise in the use of Big Data? To implement an efficient system, as represented in Figure 6, numerous challenges must be overcome, mainly in processing and managing the large amount of generated information to ensure data integrity and security.

(RQ3) What methodologies are currently used to apply Big Data in healthcare? Real-time data analysis is used, allowing immediate intervention based on current information, which can feed machine learning algorithms primarily built using open-source libraries TensorFlow and PyTorch. For real-time data processing, Hadoop and Apache Spark platforms are mainly highlighted, typically using MongoDB or Cassandra for storing and querying large volumes of unstructured health data.

Implementing an end-to-end solution that spans all necessary phases up to the Big Data Analytics stage is inherently complex, as it must address multiple factors including infrastructure limitations, data handling processes, and user acceptance. While a significant portion of health data still resides in static files, and some professionals—such as doctors, nurses, and administrative staff—may initially resist adopting new workflows, a gradual rollout backed by clear benefits can facilitate the transition. By carefully introducing modern tools and demonstrating tangible improvements, healthcare organizations can overcome these hurdles and ultimately leverage advanced analytics to improve patient outcomes.

Author contributions

Luís M. Barata: translation, validation, methodology, review, editing, and supervision; João Louro and João Afonso: original research, formal analysis, preparation of the original draft.

References

[1] TechNewsHQ. *What Is The Digital Revolution?* Available at ⟨https://www.techbusinesshq.com/what-is-the-digital-revolution/⟩ (2024-03-25).

[2] GOOGLE. *O que é Big Data?* Available at ⟨https://cloud.google.com/learn/what-is-big-data?hl=pt-br⟩ (2024-05-26).

[3] DEKATE, T. *The Origin of Big Data Analytics*. Available at ⟨https://www.analyticsvidhya.com/blog/2022/09/the-origin-of-big-data-analytics/⟩ (2024-05-24).

[4] ZHASA, M. *What Is Hadoop? Components of Hadoop and How Does It Work [Updated]*. 2023. Available at ⟨https://www.simplilearn.com/tutorials/hadoop-tutorial/what-is-hadoop⟩ (2024-05-26).

[5] YASAR, K.; ROSENCRANCE, L.; VAUGHAN, J. *What is Hadoop Distributed File System (HDFS)? | Definition from TechTarget*. 2024. Available at ⟨https://www.techtarget.com/searchdatamanagement/definition/Hadoop-Distributed-File-System-HDFS⟩ (2024-05-26).

[6] PANDA, S. *HDFS (Hadoop Distributed File System) - Data Science, AI and ML*. 2021. Available at ⟨https://discuss.boardinfinity.com/t/hdfs-hadoop-distributed-file-system/6571⟩ (2021-06-02).

[7] DATABRICKS. *What is MapReduce?* Available at ⟨https://www.databricks.com/glossary/mapreduce⟩ (2021-08-12).

[8] LăPUșAN, T. *Hadoop MapReduce deep diving and tuning*. Available at ⟨http://www.todaysoftmag.com/article/1358/hadoop-mapreduce-deep-diving-and-tuning⟩ (2024-05-26).

[9] WU, C. *Demystifying YARN: Understanding Its Architecture, Components, and How It Works*. Available at ⟨https://medium.com/@chenglong.w1/demystifying-yarn-understanding-its-architecture-\components-and-how-it-works-738dd95ad453⟩ (2021-08-12).

[10] M, S. *Yarn Tutorial*. 2022. Available at ⟨https://www.simplilearn.com/tutorials/hadoop-tutorial/yarn⟩ (2024-05-26).

[11] VIJ, A.; SAINI, S.; BATHLA, R. Big data in healthcare: Technologies, need, advantages, and disadvantages. *ICRITO 2020 - IEEE 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)*, Institute of Electrical and Electronics Engineers Inc., p. 1301–1305, 6 2020.

[12] PAGE, M. J. et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. v. 372, p. n71. ISSN 1756-1833. Publisher: British Medical Journal Publishing Group Section: Research Methods &amp; Reporting. Disponível em: ⟨https://www.bmj.com/content/372/bmj.n71⟩.

[13] YU, W. D. et al. A modeling approach to big data based recommendation engine in modern health care environment. *Proceedings - International Computer Software and Applications Conference*, IEEE Computer Society, v. 1, p. 75–86, 9 2015. ISSN 07303157.

[14] REDDY, A. R.; KUMAR, P. S. Predictive big data analytics in healthcare. *Proceedings - 2016 2nd International Conference on Computational Intelligence and Communication Technology, CICT 2016*, Institute of Electrical and Electronics Engineers Inc., p. 623–626, 8 2016.

[15] THARA, D. K. et al. Impact of big data in healthcare: A survey. *Proceedings of the 2016 2nd International Conference on Contemporary Computing and Informatics, IC3I 2016*, Institute of Electrical and Electronics Engineers Inc., p. 729–735, 2016.

[16] GONZALEZ-ALONSO, P.; VILAR, R.; LUPIAñEZ-VILLANUEVA, F. Meeting technology and methodology into health big data analytics scenarios. In: *2017 IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS)*. [s.n.]. p. 284–285. ISSN: 2372-9198. Disponível em: ⟨https://ieeexplore.ieee.org/document/8104203/figures#figures⟩.

[17] TSE, D. et al. The challenges of big data governance in healthcare. *Proceedings - 17th IEEE International Conference on Trust, Security and Privacy in Computing and Communications and 12th IEEE International Conference on Big Data Science and Engineering, Trustcom/BigDataSE 2018*, Institute of Electrical and Electronics Engineers Inc., p. 1632–1636, 9 2018.

[18] NTEHELANG, G. et al. Iot-based big data analytics issues in healthcare. *ACM International Conference Proceeding Series*, Association for Computing Machinery, p. 16–21, 11 2019. Disponível em: ⟨https://dl.acm.org/doi/10.1145/3369555.3369573⟩.

[19] ISLAM, M. D. S. et al. A case study of healthcare platform using big data analytics and machine learning. *ACM International Conference Proceeding Series*, Association for Computing Machinery, p. 139–146, 6 2019. Disponível em: ⟨https://dl.acm.org/doi/10.1145/3341069.3342980⟩.

[20] PANDEY, A. K. et al. Key issues in healthcare data integrity: Analysis and recommendations. *IEEE Access*, Institute of Electrical and Electronics Engineers Inc., v. 8, p. 40612–40628, 2020. ISSN 21693536.

[21] MADYATMADJA, E. D. et al. Analysis of big data in healthcare using decision tree algorithm. *Proceedings of 2021 1st International Conference on Computer Science and Artificial Intelligence, ICCSAI 2021*, Institute of Electrical and Electronics Engineers Inc., p. 313–317, 2021.

[22] ISHAK, M.; RAHMAN, R.; MAHMUD, T. Integrating cloud computing in e-healthcare: System design, implementation and significance in context of developing countries. *2021 5th International Conference on*

Electrical Engineering and Information and Communication Technology, ICEEICT 2021, Institute of Electrical and Electronics Engineers Inc., 2021.

[23] SHOUAIB, M.; METWALLY, K.; BADRAN, K. Survey on iot-based big data analytics. *13th International Conference on Electrical Engineering, ICEENG 2022*, Institute of Electrical and Electronics Engineers Inc., p. 81–85, 2022.

[24] FEI, Z. et al. An overview of healthcare data analytics with applications to the covid-19 pandemic. *IEEE Transactions on Big Data*, Institute of Electrical and Electronics Engineers Inc., v. 8, p. 1463–1480, 12 2022. ISSN 23327790.

[25] HUSSAIN, F. et al. Leveraging big data analytics for enhanced clinical decision-making in healthcare. *IEEE Access*, Institute of Electrical and Electronics Engineers Inc., v. 11, p. 127817–127836, 2023. ISSN 21693536.

[26] YADAV, A. et al. Role of ai, big data in smart healthcare system. *2023 6th International Conference on Information Systems and Computer Networks, ISCON 2023*, Institute of Electrical and Electronics Engineers Inc., 2023.

[27] KHAN, S. et al. Incorporating deep learning methodologies into the creation of healthcare systems. *2023 International Conference on Artificial Intelligence and Smart Communication, AISC 2023*, Institute of Electrical and Electronics Engineers Inc., p. 994–998, 2023.

[28] CONSTANTINOU, I. et al. *Big Data in Healthcare: Intensive Care Units as a Case Study*. 2014. Available at ⟨https://ercim-news.ercim.eu/en97/ri/big-data-in-healthcare-intensive-care-units-as-a-case-study⟩ (2024-53-25).