Multimedia Systems (2022) 28:1339–1371
https://doi.org/10.1007/s00530-020-00736-8

SPECIAL ISSUE PAPER

Leveraging big data analytics in healthcare enhancement: trends, challenges and opportunities

Arshia Rehman[1] · **Saeeda Naz**[1] · **Imran Razzak**[2]

Received: 13 August 2020 / Accepted: 9 December 2020 / Published online: 21 January 2021

Abstract
Clinical decisions are more promising and evidence-based, hence, big data analytics to assist clinical decision-making has been expressed for a variety of clinical fields. Due to the sheer size and availability of healthcare data, big data analytics has revolutionized this industry and promises us a world of opportunities. It promises us the power of early detection, prediction, prevention, and helps us to improve the quality of life. Researchers and clinicians are working to inhibit big data from having a positive impact on health in the future. Different tools and techniques are being used to analyze, process, accumulate, assimilate, and manage large amount of healthcare data either in structured or unstructured form. In this review, we address the need of big data analytics in healthcare: why and how can it help to improve life?. We present the emerging landscape of big data and analytical techniques in the five sub-disciplines of healthcare, i.e., medical image analysis and imaging informatics, bioinformatics, clinical informatics, public health informatics and medical signal analytics. We present different architectures, advantages and repositories of each discipline that draws an integrated depiction of how distinct healthcare activities are accomplished in the pipeline to facilitate individual patients from multiple perspectives. Finally, the paper ends with the notable applications and challenges in adoption of big data analytics in healthcare.

Keywords Big data analytics · Medical image processing and Imaging informatics · Bioinformatics and Genomics · Clinical informatics · Public health informatics · Medical signal analytics

1 Introduction

Due to the sheer size and availability of multidimensional data, the rate of technological innovation has bought huge potential to make an extra ordinary impact on our daily life in different disciplines especially in healthcare sector. The rapidly growing and exploited data will refer to introduce a new gigantic term known as big data. Uncovering information from such complicated nature of data is often a complex process. The development and analysis of tools and methods for analysis of such large quantities of data provide us with an opportunity to make the transition into new era far easier. Having data-driven, real-time insights accessible to the organization through analytics can be a critical enabler for executing the organization strategies. Big data analytic's greatest asset is its possibilities and its need to find new ways to provide the services that we are looking for.

Unlike other fields, big data analytics is so promising in healthcare sector and received much more attention in the last few years. Clinician decisions are becoming evidence-based, meaning that they are relying more on large swathes of research and clinical data as opposed to solely their schooling and professional opinion. Big data in terms of healthcare is defined as the name given to larger and complex electronic healthcare datasets that are problematic or almost impossible to manage by employing common traditional methods, tools or software [1–3]. Big data in healthcare are generated by healthcare record (such as patients record, disease surveillance, hospital, medicine, health management, doctor, clinical decision support or feedback of patient [4–7]) and clinical data (like imaging, personal,

✉ Imran Razzak
imran.razzak@ieee.org

Arshia Rehman
arshiar29@gmail.com

Saeeda Naz
saeedanaz292@gmail.com

[1] Computer Science Department, Govt. Girls Postgraduate College No.1, Abbottabad, KPK, Pakistan

[2] Deakin University, Geelong, Australia

financial record, genetic and pharmaceutical data, Electronic Medical Records (EMR), etc. [8]). The generation and management of these enormous healthcare record are considered to be very complex; thus, big data analytics is introduced [9, 10]. With the rise of technological innovation and personalised medicine, big data analytics has the potential to make a huge impact on our life, i.e., how it helps to predict, prevent, manage, treat and cure disease. Furthermore, it helps government agencies, policy maker, and hospital to manage resources, improving medical research, planning preventative methods, and managing epidemic.

With the advancement in information and communication technology system, hard copy medical data tend to move towards Electronic Health Records (EHR) and Electronic Medical Records (EMR) systems. These systems generated exponential growth of data [11, 12]. Healthcare data are not only collected from clinical record, tele-monitoring or medical tests but also there are a larger number of healthcare apps. These apps have tremendous amount of subscriptions. According to the *Ericsson Mobility Report* of 2019, there were a total of 7.9 billion mobile subscriptions, with 49 million new subscriptions added during the quarter as the growth of people on this planet subscribe new and valuable data about health and well-being everyday [13]. These apps contain voluminous data due to the world of social media. There are more than 4 billions people [14] who use internet for the purpose of mailing, downloading, surfing, blogging, entertainment, etc. These amount of data also tend to move towards the concept of big data. Figure 1 depicts the ecosystem of healthcare assisted by big data and cloud computing approaches.

Moving towards the five characteristics of big data in healthcare sector, *Volume* refers to the medical record of personal data, clinical data, radiology images, genetics and

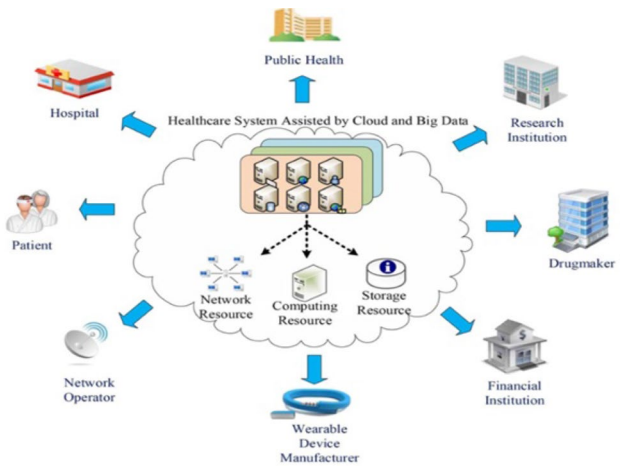

Fig. 1 Healthcare ecosystem assisted by big data and cloud computing [15]

population information, resource intensive applications like 3D imaging genomics and biological sequences. Likewise, rapid increase in diseases and medications produces exponential growth of data that are to be stored, manipulate and managed. For the effective capturing, management and manipulation of data modern techniques like advances in data management, cloud computing, visualization, etc. play a vibrant role for healthcare systems. Volume is rapidly increasing in bio-medical informatics like Proteomics DB [16] which contains data volume of 5.17 TB covering 92% of human genes information explained in Swiss-Prot database. Vast amount of volume is produced from medical images like Visible Human Project which comprehends female datasets of 39 GB [17]. It is estimated that volume of big data in healthcare increased to 35 zeta-bytes by 2020 [18, 19].

Variety in healthcare divulges that there is a gigantic amount of healthcare record either in structured, unstructured, or semi-structured format. There is a variety of unstructured healthcare record generated daily like patient information, doctor notes, prescriptions, clinical or official medical records, images of MRI, CT, radio films, etc. Furthermore, structured and semi-structured variety regarding EMS and EHS comprises actuarial data, electronic apps and automated databases information like physician name, hospital name, treatment reimbursement codes, patient name, address, etc., information of electronic billings and accounting, and some of the clinical and laboratory instrument reading observations. For the conversion of unstructured data into structured datasets, data analytics provides different facilities; one of them is natural language processing in health fidelity.

Another important characteristic is *velocity* that can be at rest or motion pace. At rest velocity, healthcare record encompasses doctor or nurse notes, scripts, documentary files, renders record, X-ray films, etc. Moreover, medium-velocity healthcare data include blood pressure readings, measurement of daily diabetic glucose by insulin pumps, ECG/EKG, etc. However, sometimes high velocity is required, as it become a staple of life or death. This type of data embroils on real-time data like monitoring of inside heart, anesthesia and trauma for blood pressure, room operations, detecting infections or diseases like cancer, etc. at early stage.

Value describes how much data are beneficial for healthcare ecosystem. For example, raw data like paper prescriptions, official record or patient information are less valuable than diagnostics record, medicines and laboratory instruments reading record. *Veracity* tells the reliability or understandability of healthcare record that explains the capturing of diagnosis, procedures, treatments, etc. and to verifying the information of patient, hospital, reimbursement code, etc. Different domains of healthcare and medical care were proposed in the literature. This review paper discusses five

sub-disciplines (i.e., medical image processing and imaging informatics, bioinformatics, clinical informatics, public health informatics, medical signal analytics) that directly or indirectly involve in healthcare and bio-medical.

In Sect. 2, we present the theoretical information of big data and data analytics. Different architectures of big data analytics deployed in the domain of healthcare are explained in Sect. 3. We also present the advantages of big data to healthcare in Sect. 4 that give the insights how healthcare can be improved by big data analytics. Section 5 presents the review methodology to give the insight about the criteria of paper selection. Based on the review methodology, the big data in five sub-disciplines of healthcare (i.e., medical image processing and imaging informatics, bioinformatics, clinical informatics, public health informatics, medical signal analytics) are comprehensively explained in Sect. 6. We also summarize our main findings in Sect. 7. Then, Sect. 8 presents the notable applications of healthcare analytics based on the main findings. Section 9 discusses the challenges and open research issues. Finally, the Sect. 10 draws conclusion of this paper.

2 Background of big data and data analytics

The concept of big data was introduced in 1990s by Cox and Ellsworth [20], when they considered visualization as a big data problem. The significant academic references of big data in computer science was first discovered by Weiss and Indurkhya [21]. In 2000, Diebold [22] introduced big data in statistics/econometrics when they referred to exploited quality information. The concept was enriched by Douglas Laney at Gartner in an unpublished 2001 research [23]. In short, the term Big data is attributed to Weiss and Indurkhya, Diebold, and Laney. Big data is the name given to the larger and enormous datasets that are usually complex so that traditional information processing techniques are not enough to deal with them. Mostly, the difficulties or challenges regarding big data are how to capture, store, share and analyze data, how to visualize, update or query information privacy. From the view of Radar [24], big data deal with the huge amount of data that are not fit into the conventional databases; thus, an alternative way is chosen to extract and process the data from it. According to ZDNet[1], big data involve techniques and procedures for the creation, formation, manipulation, and organization of larger datasets and facilities offering for its storage. Techopedia[2] suggests that unstructured large complex data are processed by massive parallelism on readily available hardware because relational database engines

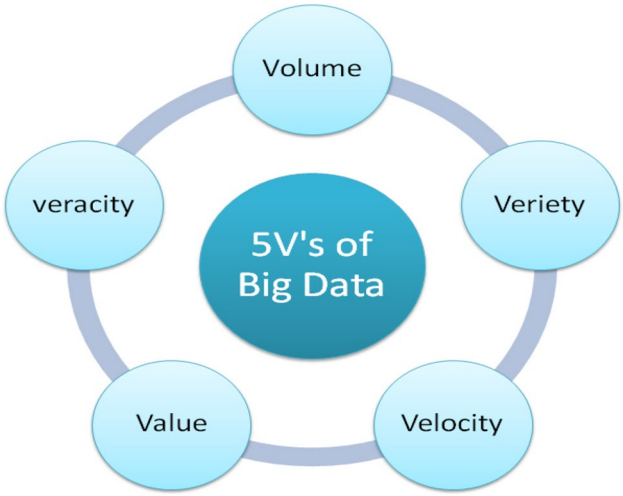

Fig. 2 5Vs characteristics of big data

are unable to process those data. In short, literature divulges that big data are larger data sets, enormous growth of data, massive data, unstructured or complex data [25–29].

Basically, main characteristics of big data are complexity and massive size [30–32]. However, big data are deliberated by three characteristics known as 3Vs—volume, variety and velocity [33–35]. Two additional characteristics are extended to make 5Vs properties of big data as depicted in Fig. 2. These additional characteristics are—value and veracity [26, 36, 37]. Volume leads to the size or quantity of stored and generated data. When the volume of data is large, they become big data [15, 38]. Variety is the type or nature of data when grouped from several sources. Data are varied in terms of format like CSV, text or Excel format in which data are stored in a database. Likewise, various forms of data also vary such as video, audio, SMS or PDF data [15]. This variety is also one of the decisive characteristics of big data. Velocity specifies the speed of data at which it is generated or processed. Value describes how much data are beneficial or valuable. The big data and the value are strongly co-related as storage of raw data is useless and inoperable. Huge data are valuable due to the costs and benefits while collecting and evaluating data [15]. The term veracity is the quality of data understandability. In other words, reliability, quality and accuracy of big data depend on the veracity property because it prevents low-quality data.

In the early stages of big data, the framework was defined using three Vs: Volume, Velocity, and Variety [27, 39]. Later, this framework was expanded to include two more dimensions: Value and Veracity [40, 41]. These 5Vs (Volume, Velocity, Variety, Value, and Veracity) are frequently referred to as the 5Vs of Big Data. The 5Vs framework has been a useful approach to addressing the issues and challenges of big data. Thus big data is defined

[1] http://www.zdnet.com/blog/virtualization/what-is-big-data/1708

[2] https://www.techopedia.com/definition/27745/big-data

as a holistic approach to manage, process, and analyze 5Vs (i.e., volume, variety, velocity, veracity and value) to create actionable insights for sustained value delivery, measuring performance and establishing competitive advantages.

Data analytics is the amalgamation of two words where data refer to raw facts, figures and information, and analytics means use of several tools to analyze data although data are small or big. Analytics is an umbrella term for all data analysis applications [33]. The big data analytics is the process of analyzing large voluminous data using different strategies. As aforementioned, big data are integrated from multiple sources; thus, big data analytics is used to explore how to extract valuable and hidden patterns and connections from this integrated data. In other words, big data analytics is simply analysis of data with the intention of extracting information and supporting conclusion-making from the inclusive procedure of scrutinizing, modeling, cleansing, and transforming of Big data.

Data analytics can be analyzed by three general methods: descriptive, predictive and prescriptive analytics. Descriptive analytics is used to summarize the big data. The simplest way to define descriptive analytics is that, it answers the question "What has happened?". Descriptive analytics was found to have more applications in analyzing what implications the different health care decisions had on the service delivery systems and clinical outcomes.The focus of descriptive analytics in healthcare organizations is to collect the patient's operational data with respect to the health organization and to arrive at patterns of patient's care leading to evidence-based clinical practice, identifying unnoticed trends in patients, imbalances between the cost, capacity and patients' needs. Predictive analytics is use to predict the future analysis by deploying a diversity of machine learning, statistical, modeling and data mining techniques to study latest recent and historical data. Prescriptive analytics is basically the predictive analytics that is used to take action and make the business decision. It is used by the health organizations when a selection is to be made from the available, feasible alternative solutions.

Most extensively used approaches for predictive and descriptive analytics on big data are based on either supervised, unsupervised, or semisupervised learning. An exponential time increase in data has made it difficult to extract valuable information from this data. Despite the strong performance of traditional methods, their predictive power is limited as traditional analysis only deals with primary analysis whereas data analytics deals with secondary analysis. Data mining involves the digging or mining of data from many dimensions or perspectives through data analysis tools to find previously unknown patterns and associations from data that may be used as valid information [42–45]. Moreover, it makes use of this extracted information to build predictive models. It has been deployed intensively and extensively by many organizations, especially in the healthcare sector.

Data mining is not a magical wand but in fact a big powerful tool that does not discover solutions without guidance. Data mining is convenient for the succeeding purposes:

- Exploratory data analysis to examine the data corpus to summarize their main characteristics.
- Descriptive modeling to segregating the data into clusters based on their properties.
- Predictive modeling to forecasting information from existing data.
- Discovering pattern to find patterns that occur frequently.
- Content retrieval to discover hidden patterns.

Several techniques are deployed for reduction, optimization, regression analysis, etc. of big data. On account of the voluminous amount of big data, its dimensionality is reduced by linear mapping approaches like Principal Component Analysis (PCA) [46], and Singular Value Decomposition (SVD) [47]. Some non-linear mapping methods for dimensonality reduction are Kernel Principal Component Analysis (KPCA) [48], Sammon's mapping [49, 50], Laplacian eigenmaps [51].

Mathematical optimization is another analytics tool that involves multi-objective and multi-modal optimization approaches like pareto optimization [52, 53], evolutionary algorithms [54, 55]. Extracting meaningful information and cluster development and analysis is achieved by various clustering algorithms like Clustering LARge Applications (CLARA) [56], Balanced Iterative Reducing using Cluster Hierarchies (BIRCH) [57], etc.

3 Architectures for big data analytics

Our anticipated general framework of big data analytics for healthcare is an abstraction of several conceptual steps that describe the generic functionalities of the domain. The first step in the framework is data collection, in which health and the clinical data are collected from internal or external sources. Variety of data includes Electronic Healthcare Records (EHRs), clinical images, health monitoring devices logs, etc. After the collection of data, next step is Data processing in which healthcare data are stored, extract and load in the data ware houses, middle-ware or in traditional formats like CSV, tables, etc. Data transformation is the next step in which data are transformed, aggregated and loaded in database file systems like Hadoop cloud or in a Hadoop distributed file systems (HDFS). Analytical phase is used to examine the big data using big data tools and platforms like Hadoop, Mapreduce, Hive, Hbase, Jaql, Avro and several

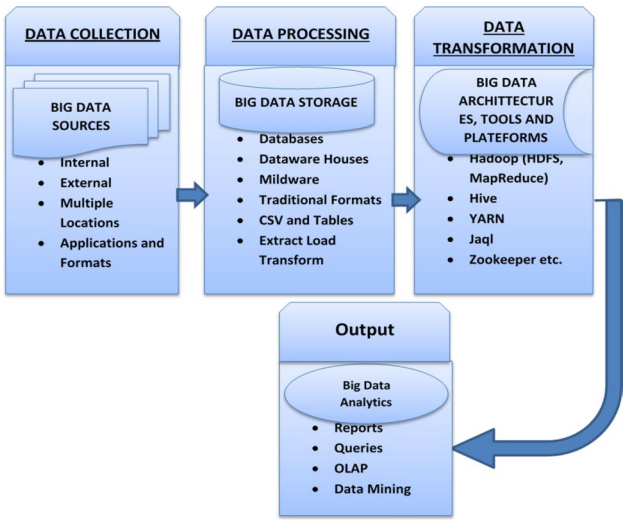

Fig. 3 Conceptual journey of data to information in big data analytics environment

others. Finally, the output is generated in the form of reports and queries using data mining and OLAP tools. The self-explanatory general and conceptual architecture is elaborated in Fig. 3 along with Fig. 4.

Based on the domain abstraction and identification, there are several definitions of big data architectures proposed and developed by researchers for big data analytics. Some of the important architectures are Hadoop, MapReduce [58], Streaming graph [59], Fault tolerant graph, etc. We present some of the renowned architectures along with its core component comprehensively in detail. One of the major frameworks on Apache platform is Hadoop developed by Doug Cutting and Apache Lucene. It is a collection of open-source software utilities used for distributed computation, processing, and storage of large data sets or big data. Succeeding Figs. 5 and 6 depict the core components and basic framework of Hadoop.

3.1 Hadoop distributed file system (HDFS)

HDFS [60] is the master–slave architecture intended to run on the commodity hardware. It provides great throughput access to application data. It allows the underlying storage for the Hadoop cluster and enhances healthcare data analytics system by segregating huge expanse of data into smaller one and disseminated it across various servers/nodes. The architecture of HDFS is divided into Name-node and Data-node where Name-node is master and Data-node is slave. Documents are stored in the data node having size of 64M that cannot be changed. Following Fig. 7 illustrates the architecture of HDFS.

According to Fig. 7, client is a HDFS user. Name-node is responsible to manage the name space in the file system. It stores and maintains the files and folders into a file system tree. Data-node is the place where the real data are saved and handled.

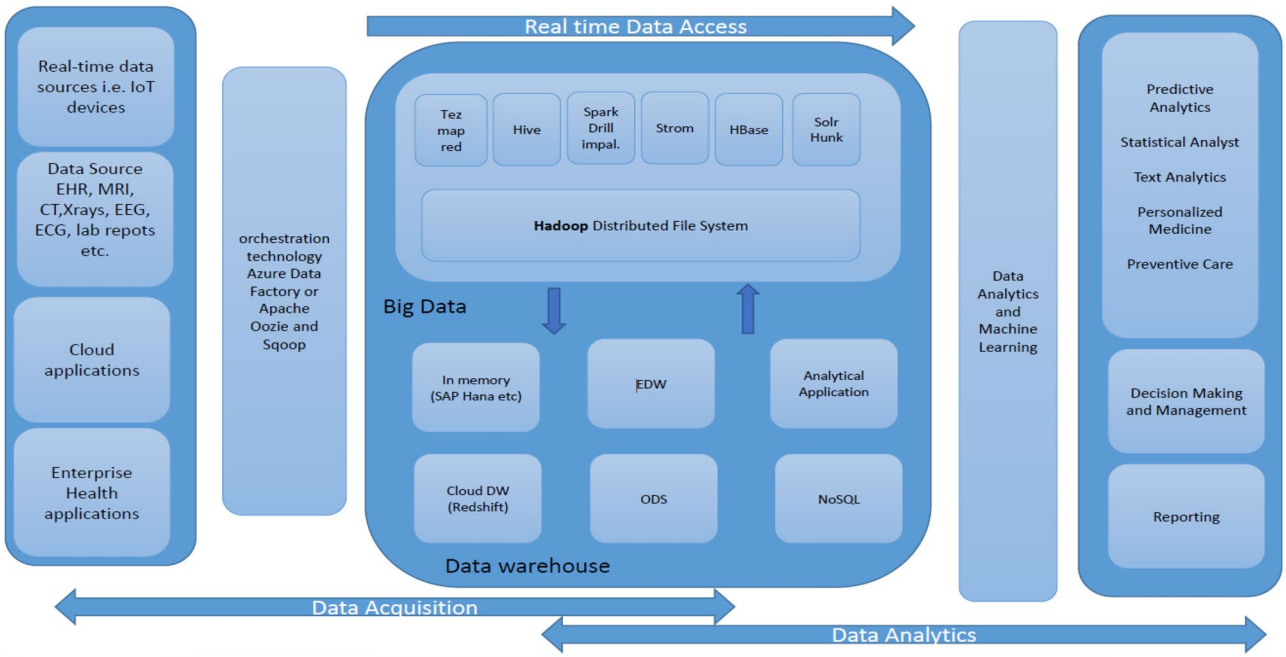

Fig. 4 Architecture of big data analytics platform

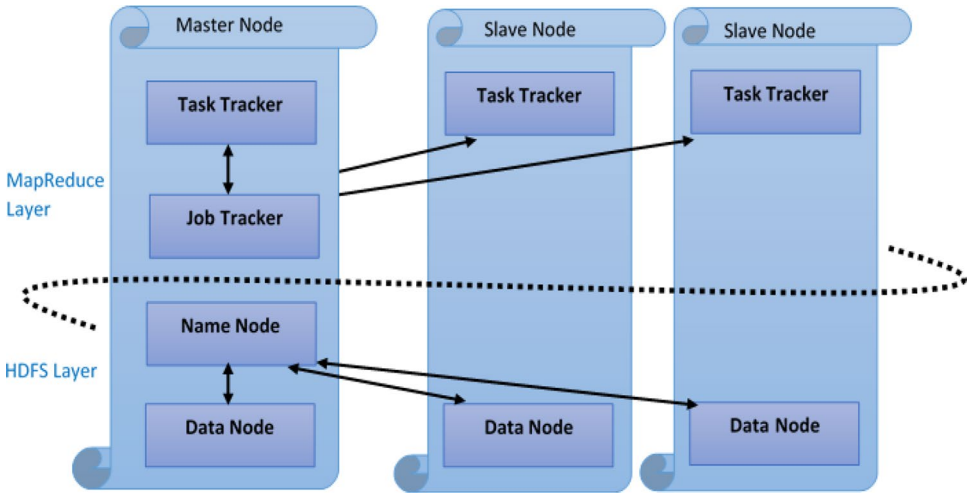

Fig. 5 Core components of Hadoop

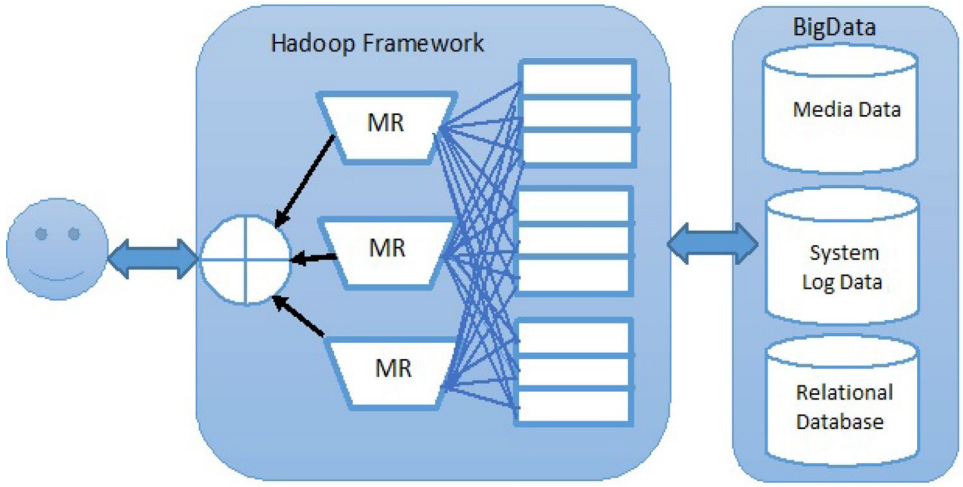

Fig. 6 Framework of Hadoop

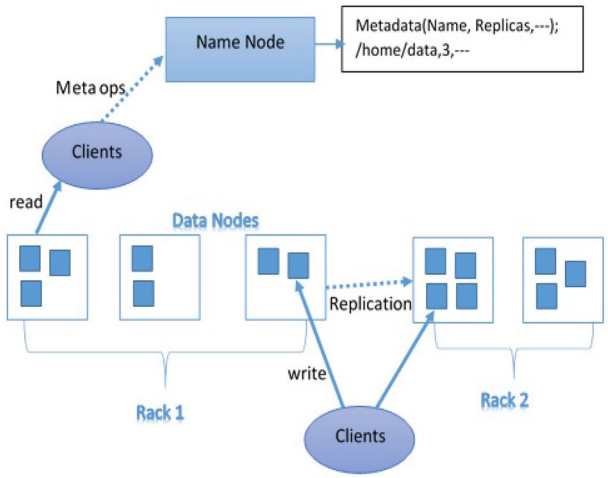

Fig. 7 Architecture of HDFS

3.2 MapReduce

Mapreduce is another cornerstone of Apache Hadoop that is developed in 2004 when Google published a thesis [58]. MapReduce is a standard functional programming model that processes and analyzes the big data . It breaks task into sub-tasks, gathering its output and analyzes efficiently large datasets in parallel mode. Data analysis and processing employed two steps, namely: Map phase and Reduce phase.

The architecture of MapReduce operation is split into three main components: Client, Job-Tracker, and Task-Tracker. Client submit its job to the Job-Tracker in the form of JAR file. Job-Tracker maintains all the jobs that are executed on the MapReduce and, thus, acts as master service. Task-Tracker executes the jobs that are assigned

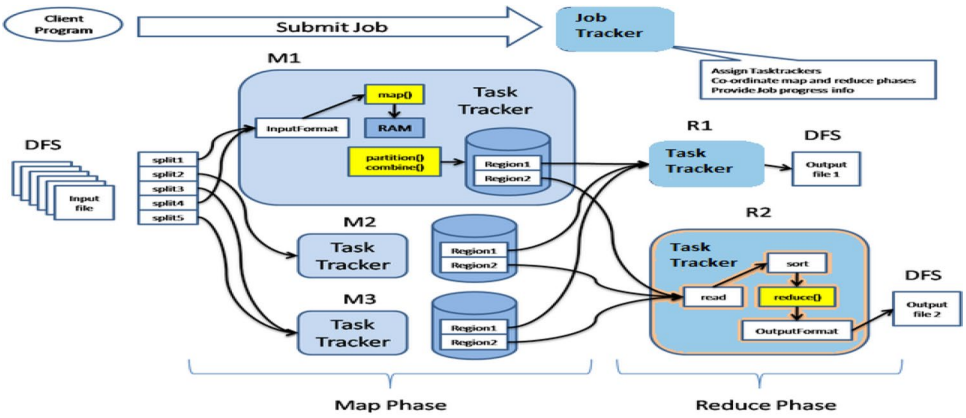

Fig. 8 MapReduce architecture

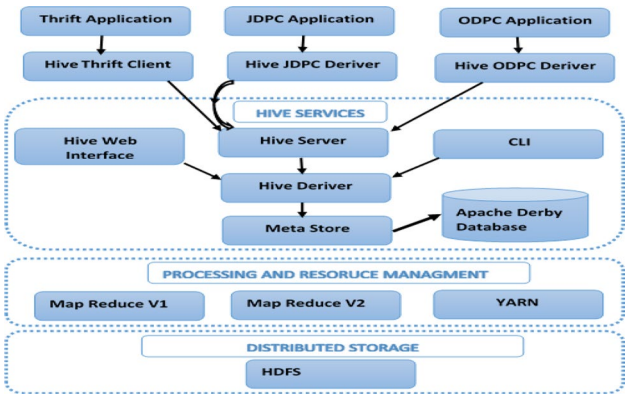

Fig. 9 Hive architecture

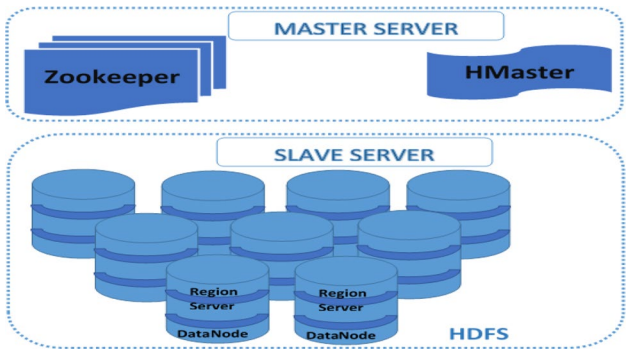

Fig. 10 Hbase architecture

by Job-Tracker and, thus, acts as slave service. Figure 8 demonstrates the generic architecture of MapReduce operation.

3.3 Apache hive

Apache Hive [61] is a Structured Query Language (SQL) based on Extract Transform Load (ETL) and dataware house on Hadoop plateform. It is a runtime Hadoop provision framework that works on Hive Query Language (HQL) that converts SQL queries into MapReduce jobs. The main operations performed by Hive are data encapsulation, analyzing, ad hoc querying and summarizing large datasets. Apache Hive have four major components: hive clients, services, processing framework and distributed storage. Hive client like thrift clients, JDBC clients, ODBC clients, etc. can be written in any supportive language like C++, Java, Python, etc. Services are used to perform queries. Services of Hive may include command line interface (CLI), web interface (WI), hive server, driver, meta-store, etc. Queries are processed, executed, and managed using internal Hadoop MapReduce framework. Finally, the distributed data are deposited in HDFS. The core components are illustrated in Fig. 9.

3.4 Apache HBase

Apache HBase works on non-SQL and non-relational approach. It is a database management approach using column-oriented structure lies on the top of HDFS. It used the key/value data that perform read/write operations on large HDFS database. Apache Hbase is categorized into three main components: HMaster Server, HBase Region Server, and Zookeeper. HMaster server is the main component that manages and monitors HBase region servers, and performs database operations using DDL to create, update, and delete tables. Hbase tables are divided into several regions that manage, handle, and execute operations through Hbase region servers. Hbase is a distributed system that is coordinated by Zookeeper. The components of Apache HBase are depicted in Fig. 10.

3.5 Presto

Presto [62] is a distributed structured query language engine that is used to analyze large amount of data ranging in size from gigabytes to petabytes. The architecture of Presto is

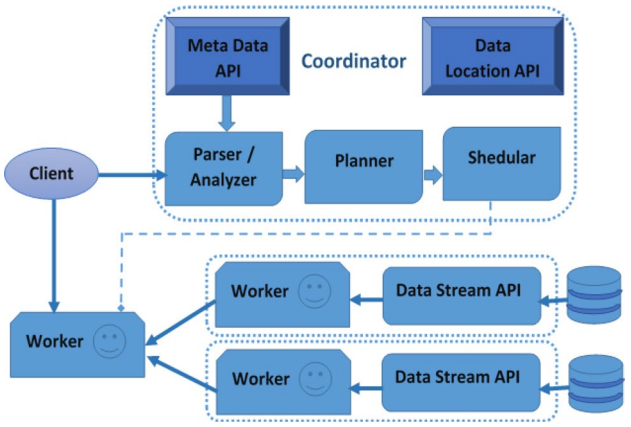

Fig. 11 Presto architecture

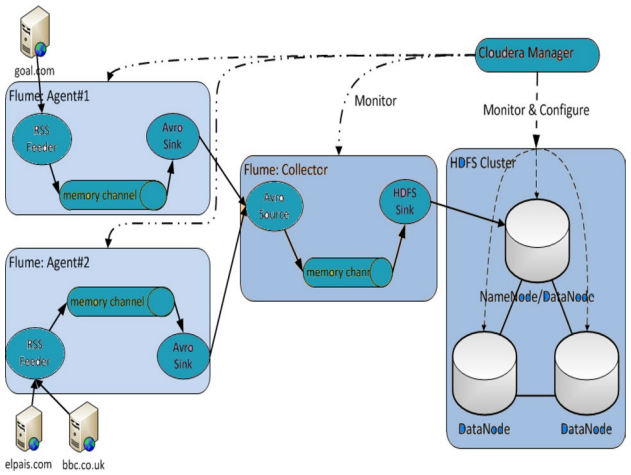

Fig. 12 Avro architecture

composed of coordinators and workers. User queries are submitted to the co-ordinator that is accountable for planning, executing, scheduling, and parsing the queries of Workers. The architecture is explained from the succeeding Fig. 11.

3.6 Mahout

Mahout [63] is an apache scheme that is used to produce unrestricted applications of disseminated and accessible machine learning algorithms that support healthcare data analytics on Hadoop systems. It is designed to support big data analytics that provide free application on Hadoop platform like applications of distributed and accessible machine learning algorithms.

3.7 Avro

Avro [64] assists serialization and data encoding that advances structure of data by identifying data types, meaning, and scheme. It has the functionalities of serialization and versioning control features. Avro configuration is illustrated from the Fig. 12.

4 Advantages of big data to healthcare

How big data analytics can improve healthcare? Simple answer to this question is: Analyzing big data can aid healthcare stakeholders to deliver efficient procedures and insights into the patients and their health. Numerous benefits can be obtained with big data analytics. Main source of healthcare data are: Electronic Health Records (EHR), Laboratory Information Management system (LIMS), Pharmacy, Monitoring and diagnostic instruments (MDI), Finance (Insurance claim and billing) and hospital resources. With the advancement of data acquisition devices and analytics techniques, data source are getting enriched with newer forms of data, i.e., hospitals start to collect genetic information in EHR as well. Within this vast variety of patient data lie valuable insights for both patients and organizations, which when applied judiciously can bring in wonderful results. Potential benefits include advanced patient care:

Quality of care EHR helps in assembling demographic and medical data such as clinical data, lab test, diagnoses, and medical conditions. Discovering associations and patterns within this data helps healthcare practitioners to provide quality care, save lives and lower costs.

Disease prevention Spending more on health does not guarantee health system efficiency. The investment in prevention can help to reduce the cost as well as improve health quality and efficiency. Health systems face considerable challenges in endorsing and protecting health at a time when the burden on finances and resources is substantial in many countries. The early detection and prevention of disease plays a very important role in reducing deaths as well as healthcare costs. Thus, the core question are: How can we diminish the level of ill health in the population? And how can we prevent the disease to occur based on early symptoms of patient?

Efficiency Managing healthcare data using traditional analytical tools is nearly impossible due to the diversity and volume of data. Healthcare stakeholders use big data as a part of their business intelligence strategy to examine historical patient admission rates and to analyze staff efficiency.

Disease cureness Healthcare practices have largely been reactive where the patient has to wait until the onset of disease after which treatment is prescribed which hopefully leads to a cure. However, no two persons in the world would have the same genetic sequence. Furthermore, environmental factors associated with the onset of the disease are not known, which is the motive why particular medication

seems to work for few people but not for others. Since there are millions of things to be considered in a single genome, it is almost impossible to study them comprehensively. On the other hand, big data in healthcare have been revolutionizing the expanse of genomics medicine. Big data analytics can extract hidden patterns, unknown correlations, and insights by exploring large datasets. Scientists are banking on big data to discover the cure for cancer.

Cost Healthcare cost can be cut down by analyzing big data i.e., predictive analytics can help to detect disease at early stage. Moreover, big data also reduce medication errors by advancing economic, administrative performance, and re-admissions. For example, patient groups affected by a disease and are treated with different drug regimens can be compared to determine which treatment plans work best for the same or similar disease which result in saving resources and money.

Finding diseases cure A particular medication seems to work for a few people but not for others, and there are numerous things to be discovered in a single genome. It is not feasible to observe all of them in element. However, big statistics can help in uncovering unknown correlations, hidden styles, and insights by analyzing large sets of statistics. By applying machine learning in big data, practitioners find the big facts to have a look at human genomes and find the correct remedy or drugs to deal with cancer [42, 65].

5 Review methodology

The review methodology is the systematic process of finding the relevant literature from different sources. The main objectives of review methodology are:

- To deploy the definitions and concepts of Big data in healthcare.
- To explore the five sub-disciplines (i.e., medical image processing and imaging informatics, bioinformatics, clinical informatics, public health informatics, medical signal analytics) that directly or indirectly involve in healthcare and bio-medical.
- To illustrate the repositories and complex datasets of five sub-disciplines.
- To determine the big data analytical architectures and techniques in healthcare.
- To discuss the potential advantages and applications of big data in healthcare.
- To present the open challenges and research issues of big data in healthcare and the strategies tackling the challenges facing in the domain.

The main steps of review methodology are information sources, selection criteria, and search and selection procedure. *Information Sources*: The first step in the systematic process of research methodology is to collect the relevant articles. To search the relevant articles, we used Google Scholar. We scanned the references to present a thorough review. *Selection Criteria*: In second step, we selected the literature on the basis of following inclusion–exclusion criteria:

- Studies were based on articles, conferences, and reviews
- Studies written in English language
- Studies related to the big data analytics in healthcare
- Studies published from 2000 to 2019

Search and selection procedure: In the third step, we searched the studies from the information sources containing the keywords of "big data", "big data analytics", "healthcare", "biomedical" and "healthcare analytics" to provide the background information, advantages, and architectures of big data analytics. As mentioned earlier, our goal is to expand the research in healthcare using five sub-disciplines, we used the additional keywords: "medical" , "medical image processing" "imaging informatics" ,"bioinformatics", "clinical informatics", "public health informatics, "medical signal analytics". On the basis of initial search criteria, 47,130 papers were found; thus, we scrutinized the title, keywords and abstract and excluded 28,280 papers. We also performed the screening on the basis of full-text reading and excluded 18,020 papers that are irrelevant to the big data or healthcare domain. We ended with 830 papers that are included in this review paper.

The abstract symbols are used to present schematic process of review methodology in Fig. 13.

6 Key application in healthcare

Health professionals, just like business entrepreneurs, are capable of collecting massive amounts of data and look for best strategies to use these numbers to reduce costs of treatment, predict outbreaks of epidemics, avoid preventable diseases and improve the quality of life in general.

Different domains of healthcare and medical care had been proposed in the literature. The general overview, analysis, and examples of big data in healthcare analytics were presented in the studies of Raghupathi [2] and Ward et al. [10]. The meaning of big data in healthcare was presented in the literature reviews of Baro et al. [3] and Wamba et al. [66]. In 2017, Zhang and Li [67] presented the literature review of specialized healthcare and HIV self-management. Jacofsky [68] discussed the pitfalls of analytics related to the physicians from metadata sets in healthcare. Another case study of healthcare analytics was presented in 2018 by Wang et al. [69] that presented IT-enabled procedures, advantages,

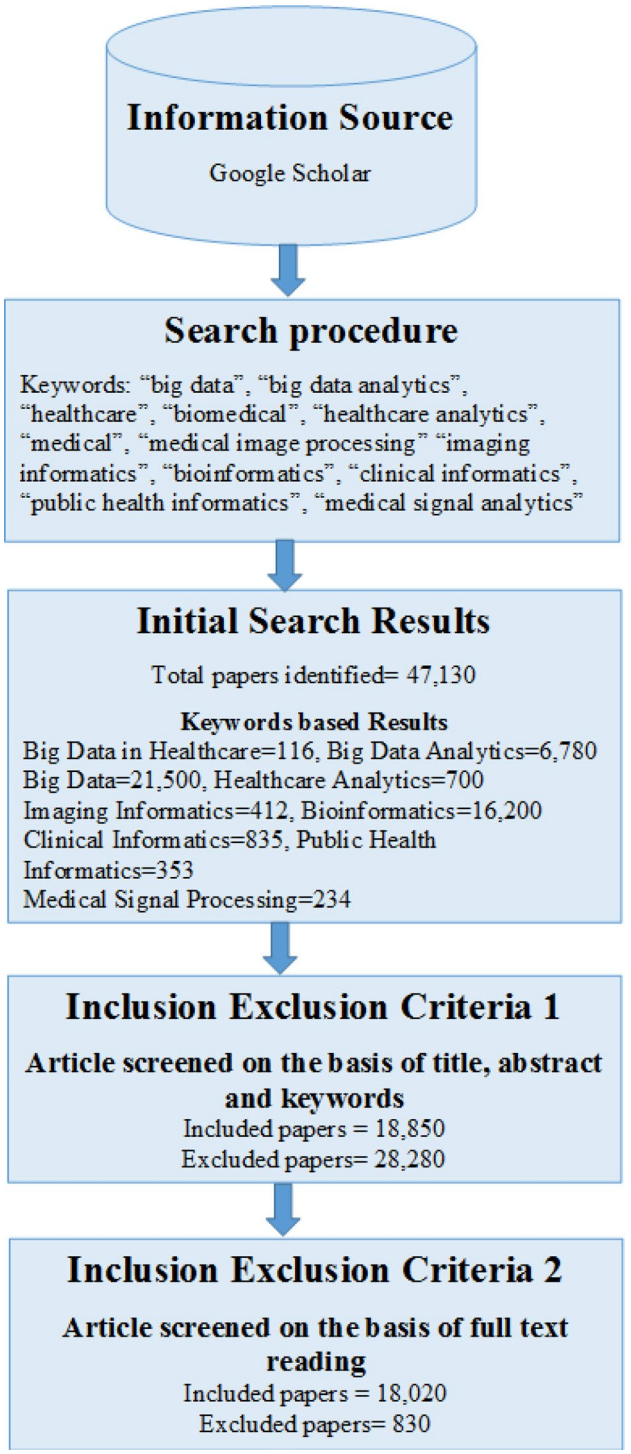

Fig. 13 Schematic process of review methodology

and capabilities of big data analytics. Galetsi and Katsaliaki [70] reviewed the articles of big data analytical techniques for healthcare from 2000 to 2016.

In this review, we discuss five sub-disciplines (i.e., medical image processing and imaging informatics, bioinformatics, clinical informatics, public health informatics, medical signal analytics) that directly or indirectly involve in healthcare and bio-medical. As mentioned earlier, we cover the literature from 2000 to 2019 that provides the comprehensive evaluation of big data techniques in healthcare domains. The literature review of five sub-disciplines of healthcare is explained comprehensively in the following subsections.

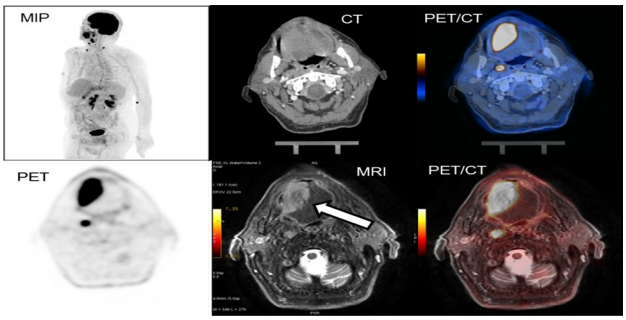

Fig. 14 Popular image modalities in healthcare Like CT, MRI, PET images

6.1 Medical image processing and imaging informatics

Medical image processing and imaging informatics are the main applications that play a vital role in healthcare and bio-medical. One of the acceptable uses of medical imaging is to detect diseases like tumors detection of brain and lungs [71, 72], artery stenosis detection, organ delineation detection [73], aneurysm detection and the diagnosis of spinal deformity and so on. Image processing and machine learning techniques were deployed in these applications for the accurate and effective use of computer-aided medical diagnostics and decision-making. In complex healthcare and bio-medical, information is generated, managed, analyzed, exchanged, and represented imaging information using imaging informatics [75–76].

After the brief introduction, we will elaborate the related work of medical imaging and informatics, techniques and applications deployed in big data healthcare.

Medical imaging is used in image acquisition. Magnetic Resonance Imaging (MRI), Computed Tomography (CT), photo-acoustic and ultrasound images are used for single-dimensional medical data like visualizing the structure of blood vessels [73, 75, 77]. However, for multidimensional medical data like 3d ultrasound, functional MRI (fMRI), Positron-emission tomography (PET), etc. are used as shown in Fig. 14.[3] There are publicly available medical images

[3] Available at: http://www.en.nuk.usz.ch/expert-knowledge/PublishingImages/pages/pet-center/PETMR.png

Table 1 Medical image modalities

Databases	Images	Patients	Data size	Modalities	Applications
Image CLEF Database [a]	306,549	–	316 GB	CT,MRI, PET, Ultra-sound	Modality Classification , Visual Image Annotation, Scientific Multimedia Data Management
Digital Mammography database [b]	9,428	2620	211 GB	DX	Computer Algorithm research development for screening
Cancer Imaging Archive Database [c]	244,527	1010	241 GB	CT, DX, CR	image validation of drug response, detection and classification of Lesion, Diagnostic Image Decision, etc.
Public Lung Image Database [d]	28,227	119	28 GB	CT	Screening Images for identification of lungs cancer
MS Lesion Segmentation [e]	145	41	36 GB	MRI	3D MS Lesion Segmentation Techniques development and comparison
ADNI Database [f]	67,871	2851	16GB	MRI, PET	Alzheimer's disease progression

[a] http://www.imageclef.org/2013/medical

[b] http://marathon.csee.usf.edu/Mammogr aphy/Database.html

[c] https://public.cancerimagingarchive.net/ ncia/dataBasketDisplay.jsf

[d] https://eddie.via.cornell.edu/crpf.html

[e] http://www.ia.unc.edu/MSseg/download .php

[f] http://adni.loni.ucla.edu/data-samples/acscess-data/

repositories that contain medical images of patients in different sizes and modalities depicted in the Table 1.

Shackelford [78] used fMRI images and single-nucleotide polymorphism (SNP) for the classification of schizophrenia and healthy subjects. They retrieved 87% classification using hybrid machine learning method. Chen et al. [79] introduced a computer-aided decision support system for the treatment of patients with traumatic brain injury (TBI). They predicted the intracranial pressure (ICP) level from CT scans images. They combined CT scans images for features extraction, medical records and patient's demographics. They achieved 70.3% accuracy, 65.2% sensitivity and 73.7% specificity correspondingly.

Yao et al. [80] introduced a system for retrieval of medical images based on Hadoop. They applied the local binary pattern algorithm and Brushlet transform for feature extraction of medical images. They implemented MapReduce for storing features in HDFS. They reported highest precision rate of 95.04% and recall of 92.21% on brain CT images. They concluded that retrieval efficiency of medical images were improved but retrieval time decreased.

Jai-Andaloussi et al. [81] employed the MapReduce for computation and HDFS for storage in content-based image retrieval systems. They used mammography image database and applied Bi-dimensional Empirical Mode Decomposition with Generalized Gaussian Density functions (BEMD-GGD) method and Bi-dimensional Empirical Mode Decomposition with Huang-Hilbert Transform (BEMD-HHT) method. They used Kernal Linear Discriminant (KLD) and euclidean distance. They produced promising results to prove the hypothesis that MapReduce technique can be effectively employed for content-based medical image retrieval.

Dilsizian and Siegel [82] worked on cardiac imaging and medical data by integrating several techniques like data mining, AI, and parallel computing. Their system used AI and big data for the diagnostic imaging of 55 participating sites from the group of formation of optimal cardiovascular utilization strategies. The system result decreased from 10 to 5% in such case.

Istephan et al. [83] conducted a feasibility study in the epilepsy domain. They used the distributing computation of hadoop clusters. Their framework deals with the structured and unstructured medical data.

6.2 Bioinformatics

Bioinformatics is a discipline of sciences which deals with mathematical, computerized and IT-based methods, techniques, algorithms and software tool for capturing, storing, analyzing, compiling, simulating and modeling information of life science and biological data. Role of big data in bioinformatics is to provide efficient data manipulation tools for investigation to analyze biological information of patient. Hadoop and MapReduce are currently used extensively for bioinformatics analytics.

Basically, bioinformatics is the combination of biology and computer science [84]. The biological analysis system analyzes variations at the molecular level. The bioinformatics consists of a variety of data types like genomics (genes sequencing), RNA, DNA, proteomics (protein sequencing), gene onthology, protein–protein interaction, pathway data, association network of the disease gene and a network of human disease as shown in Fig. 15. With the current trends in personalized care, there is an increasing

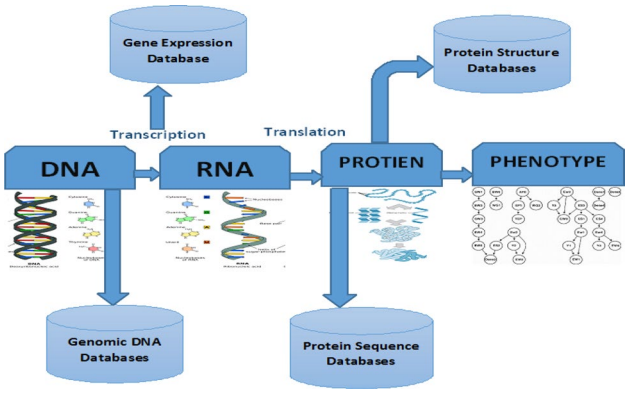

Fig. 15 Bioinformatics types

demand to analyze massive size of personalized patient data in a manageable time frame.

The size of bioinformatics'data is increasing exponentially day by day. For example, a single human's sequence of the genome is almost up to 200 GB [85]. A database produced by European Bio-informatics Institution (EBI) has getting double volume after each year [86]. *Genomics or Genome sequencing data* are currently being annotated as big data of bioinformatics problem because human genomics consists of 30,000–35,000 genes [87, 88]. Genomics data are usually the data related to gene sequencing, DNA sequencing, genotyping, gene expression, etc. [89, 90] Gene is made of DNA comprising 3 billion pairs of four building blocks or bases known as Adenine, Thymine, Cytosine and Guanine. The single genome has the size of about 3 GB. Genome analysis employing micro-arrays has been profitable in examining traits across a population and widely contributed in treatments of several complicated diseases like bipolar disease, hypertension, rheumatoid arthritis, diabetes, muscular degeneration, coronary heart disease, Crohn's disease, etc. [91]. This genomics information tends to move towards big data analytics.

In bioinformatics, protein sequencing and protein–protein interaction are sophisticated problems in functional genomics. This is due to huge number of enormous features in feature vector that is not only cost effective and complex analysis, but also reduces accuracy. Thus, feature selection of big data problem was overcome by the method proposed by Bagyamathi et al. [103]. They combined improved harmony search algorithm to improve the accuracy and feature selection. Likewise, another feature selection methodology was introduced by Barbu et al. [104]. They reduced the dimensionality of an instance using annealing technique for big data learning. Similarly, adaptiveness or behavior of big data is predicted by incremental learning approach. For this purpose, Zeng et al. [105] implemented incremental feature selection method called FRSA-IFS-HIS. They applied fuzzy rough set theory on hybrid information systems and reported better performance in big data feature selection.

Once the features were extracted and selected, next step is classification or clustering. Classification is the supervised learning procedure of finding a model that describes and discriminates data classes or concepts. The model is used to predict the class label of test instances from already trained instances. Among numerous models described in the literature, linear and non-linear density-based classifiers, neural networks, decision trees, Support Vector Machines (SVMs), Naive Bayes, and K-nearest Neighbour (KNN) are the most often used methods in numerous applications [107–109]. In big data analytic, advanced models had been reported in the literature like neural networks approaches, divide-and-conquer SVM [110], Multi-hyper-plane Machine (MM) classification model [111], etc. for big data parallel and distributed learning.

Giveki et al. [112] diagnosed automatic detection of diabetics using weighted SVM on mutual information and modified cuckoosearch. They conducted experiment on diabetics datasets by selecting features from PCA. Haller et al. [113] classified parkinson patients by employing SVM. They performed pre-processing using DTI fractional anisotropy data and select most discriminated voxels as features and then classified using SVM. Son et al. [114] predicted the heart failure patients by deploying SVM. Likewise, Bhatia et. al. [115] classified heart disease by SVM. They selected optimal feature subset using integer-coded genetic algorithm.

The big data classification and regression is effectively performed using advanced decision tree. In bioinformatics, Ye et. al. [116] implemented Gradient Boosted Decision Trees (GBDT) techniques to distribute and parallelize big data. Calaway et al. [117] estimated efficiency of decision tree on big data by employing rxDTree. Hall et al. [118] modified decision tree learning by generating rules for large training dataset.

Clustering is the unsupervised learning that analyzes data objects without labeled responses. To handle big data, CLARA [56], CLARANS [119] DBSCAN [120], DENCLUE [121], and CURE [122], k-mode and k-prototype methods [123], PDBCSCAN [124], and IGDCA [125] methods were used in the literature. Literature divulges several bioinformatics repositories [126] explained in the Table 2.

Along that there were several techniques and tools employed in bioinformatics for specific task. One of the bioinformatics type is microarray data analysis. Tools used for this type were caCORRECT [127] and omniBiomarker [128]. For gene–gene network analysis, FastGCN [129], UCLA Gene Expression, Tool (UGET) [130], WGCNA [131] tools were used for specific tasks like finding disease associated with genes, parallelism with GPU, etc. Several tools had been proposed for Protein–Protein interaction (PPI) that is a complex and time consuming process. NeMo

Table 2 Bio-informatics databases

Database	Database Type	Size	Description
European Molecular Biology Laboratory (EMBL) [92]	DNA Sequences	185,000 organisms	EMBL is the part of an international alliance with DDBJ (Japan) and GenBank (USA). It is used to analyze collection of nucleotide sequences and annotation from sources that are publically available.
Genetic Sequence Data Bank (GenBank) [93, 94]	DNA Sequence	15,000 DNA and RNA sequences entries	This database contains nucleotide sequences that provide information based on functional and physical contexts of the sequences.
DDBJ [95]	DNA Sequences	1,880,115 entries and 1,134,086,245 bases	This dataset is known as All-round Retrieval for Sequence and Annotation that enable its users to search keywords from Nucleotide Sequence Database Collaboration
The GDB Human Genome [96]	Genomics Database		Public Database of human genes, clones, STSs, polymorphisms and maps
SWISS-PROTT [97, 98]	Protein Sequences	557,012 sequence entries, comprising 199,714,119 amino acids	It contains information of protein variety, function and associated disorders
UniProtKB / TrEMBL [99]	Protein Sequences		Computer-annotated protein sequence database. It contains sequence translation of coding sequences present in the EMBL/GenBank/DDBJ
PROSITE [100]	Protein Sequences	1329 patterns and 552 profile entries	This database contains meaningful biologically signatures that described patterns or profiles
PDP [101]	Protein Structure	32,500 structures	This repository is informative with online reports, summaries, tools and information related to structural genomics initiatives
BiowareHouse [102]	Comprehensive Database		This detailed repository is the integration of the set of databases including ENZYME, KEGG, and BioCyc, and in addition the UniProt, GenBank, NCBI Taxonomy, and CMR databases, and the Gene Ontology

[132], MCODE [133], and ClusterONE [134], PathBLAST [135] had been developed for PPI analysis. For pathway analysis, GO-Elite [136], PathVisio [137], directPA [138], pathway processor [139], Pathway-PDT [140], and pathview [141] tools had been employed.

In protein–protein interaction and protein sequence, sequencing data are mapped with the specific genomes for the analysis of various tasks like genotype and expression variation. One of the major task of genomes is DNA sequencing that is produced from millions of sequencing machines data. There were several techniques for matching DNA sequence with reference gene. A parallel computing model for matching genomes is CloudBurst [142]. It uses 24 core clusters for evaluation that is 24 times faster in speed than single-core system. It has the capability of short read mapping of 7 million reads that improved the scalability of reading huge sequencing data. On the basis of CloudBurst, Contrail [143] was developed to accumulate hefty genomes and for the identification of single nucleotide polymorphisms (SNP), Crossbow [144] was prepared.

A proteomic search engine based on Hadoop distributed framework is Hydra [145] software package. It is a distributed computing environment that processes large peptide and spectra databases to support searching of immense volumes of spectrometry data. It has the fast processing of performing 27 billion peptide scorings on a 43-node Hadoop cluster in approximately 40 min. Another query engine for bioinformatics and genomics researchers is SeqWare [146] built on Apache HBase [147]. Th SeqWare has an interactive interface with genome browsers and tools. It includes loaded U87MG and 1102GBM tumor databases used for the comparison with other prototypes.

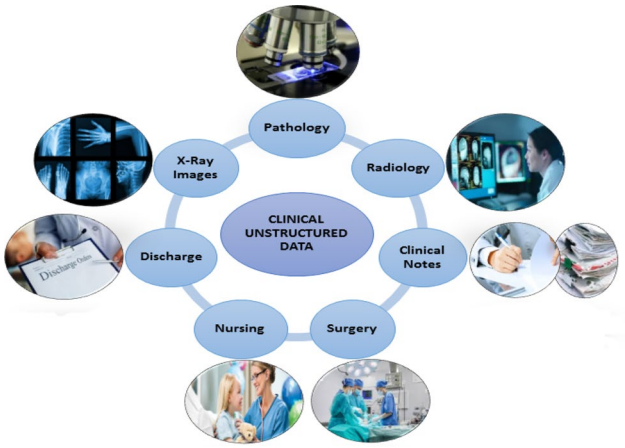

Fig. 16 Unstructured clinical informatics

There are certain tools used for the error identification of sequencing data. SAMQA [148] is the error identification tool that provides a scaleable quality for standards for large scale genomic data. ART [149] can identify three types of errors from sequencing data like base insertion, deletion, and substitution. CloudRS [150] is a parallel algorithm for error correction. It is based on RS algorithm [151]. For the analysis of data sequencing and genomic analysis, several frameworks and toolkits were developed. CloVR [152, 153] is a distributed virtual machine package for sequencing analysis that supports both local and cloud systems. Another virtual machine tool is CloudBioLinux [154] that provides 135 bioinformatics packages for analysis. Genome Analysis Toolkit (GATK) [155, 156] analyzes large sequence and genomics. It is based on MapReduce-based programming framework that had been used in 1000 Genomes Projects. BlueSNP [157] analyzed 1000 phenotypes and found association based on R package and Hadoop platform.

6.3 Clinical informatics

The clinical laboratory is a major source of data related to patients' diseases and health issue. There are approximately 80% unstructured data like clinical documents, radiology, pathology, patient discharge summaries, diagnostic testing reports, X-ray and radiological images, transcribed notes, etc. as shown in Fig. 16. Clinical informatics is the study of Information Technology (IT) and healthcare for organizing the patient's clinical data and laboratory test, reports, etc. into structured and computerized form to increase data retrieval and extraction efficiently that will assist in evaluations and reports effectively. It divulge the development of electronic health informatics systems for improvement of care and management of patients and sharing of data in seconds using computer and internet. Increasingly laboratory data are being integrated with other data of patient to improve the diagnostic process efficiency, and increase its meaningful use to improve patient outcomes. IT-based systems replace the manual data entry in records, reports, documents, and also save time and cost associated with records, hospital data and reports on daily bases, like billing and schedules of patients [158]. However, clinical informatics is currently not practiced in small clinics, hospitals, laboratories in rural and county side areas due to implementation of clinical informatics technology [159]. For boosting the implementation, the Electronic Care Records (ECR) system as a clinical informatics in the whole government hospitals in USA, HITEC [160] made some interesting incentives for the medical organizations, hospital and clinics. The doctors and physicians should use EHR systems for data of patients which they can share with any others and can provide to patients online and or can access anywhere.

In big data analytics, the first step is to store and manage data in some structured form. Clinical data are stored to observe the information of patients, hospitals and other relevant structured and unstructured record. It can be then used to settle on clinical decision, assessing patients and make treatment plans. Data warehouses and relational databases are the traditional and structured methods to store and retrieve data. However, to use clinical data, they are first transformed and classified when they are integrated from multiple sources [161, 162]. A detailed systematic review paper was published in [163] till 2011. We here present the further related work. Dutta et al. [164] stored EEG data using Hadoop and HBase in data warehouses. Jin et al. [165] analyzed and stored distributed EHR data using big data tools like Hadoop HDFS and HBase. Similarly, Nguyen et al. [166] stored signal clinical data using HBase. Jayapandian et al. [167] and Sahoo et al. [168] developed a system named 'Cloudwave' for storing and querying EEG clinical data that are voluminous. Mazurek [169] stored unstructured data in Not Only Structured Query Language (NoSQL) repositories to provide fast processing speed and data mining capabilities. For this purpose, relational and multidimensional technologies were combined with NoSQL.

Clinical data were often retrieved and shared interactively for data integration and knowledge sharing, so the cloud computing was usually considered for this purpose. Bahga and Madisetti [170] proposed a system based on cloud approach for inter-operable EHRs. Chen et al. [171] translated the informatics aspects of present and future using cloud computing. For multi-site clinical traits, the interactions of researchers were enhanced by the conceptual software architecture developed by Sharp [172] using cloud approach.

Clinical data were analyzed to predict the disease, risk, diagnosis, and progression. Literature revealed a lot of data analysis strategies for the prediction of clinical record. One of the predictive modeling platforms was "PARAMO"

designed by Ng et al. [173] for analyzing EHR and the generation and reuse of clinical data using a Hadoop cluster. They analyzed the EHR from 5000 to 300,000 patients and reported promising time effective results. Chawla and Davis [174] formulated the framework for patient-centered to explain the big data approaches for personalized medicine. Similarly, the big data for perioperative medicine were illustrated by Abbott [175]. Zolfaghar et al. [176] implemented big data techniques for the predictive model. They conducted an experiment on patient data of "National Inpatient Dataset and the MultiCare Health System" for the congestive heart failure. They reported the maximum accuracy upto 77% and recall upto 61%, respectively. Rangarajan et al. [177] proposed data lake architecture that used HDFS for data storage. Similar health conditions of patients were clustered using K-means. From each cluster, the successful recommendation was found by deploying SVM. Wang and Hajli [178] examined 109 case description of 63 healthcare organizations. They modeled the big data analytics for business transformation using RBT theory and capability building view in the model. Each case occurrences along with pair-wise connections, constructs and path-to-value chains were used to find business value (see Table 3).

6.4 Public health informatics

Informatics is an "Applied Information Science". It synthesizes the practices and theories of information technology, computer science, management sciences, and behavioral sciences into concepts, tools, and methods for implementing information systems into health for public. Informatics transform raw data into information effectively according to requirement of users. Healthcare informatics research is a scientific attempt that improves both health service organizations' performance and patient care outcomes as shown in the following Fig. 17.

Public healthcare is determined through Epidemiology. Epidemiology is the study of analyzing how frequently diseases arise in different groups of people and why. Epidemiological information is used to formulate and evaluate

Table 3 Clinical informatics databases

Database	Database type	Description
Texas Inpatient Public Use Data File (PUDF) [a]	Structured EHR	This dataset contains record of patients, hospitals,admission type/source, claims, admit day and discharge details. In 2017 dataset contains 699 hospitals, 776,554 base date records, 12,486,488 charges date records in First quarter. In Second quarter there were 694 hospitals, 761,921 base date records and 11,985,920 charges date records
Multi-parameter Intelligent Monitoring in Intensive Care II (MIMIC-II) Clinical Database [179]	Structured EHR	This dataset encompasses detailed clinical data, including physiological wave forms and records subsets from minute-by-minute. It contains 32,536 subjects with 40,426 ICU admissions and 25,328 intensive care unit stays
Patient Discharge Data By Admission Type [b]	Unstructured	Dataset contains the information of inpatient discharges by type of admission for each California hospital for years 2009–2015 containing 9322 entries
Framingham Heart Study Database [c]	Structured EHR	It is a genetic dataset for cardiovascular diseases like Heart. It include 5209 men and women having age between 30 and 62 years. 1948, participants had been assessed every 2 years
Basic Stand Alone (BSA) Medicare Claims Public Use Files (PUFs) [d]	Unstructured	CSV format that contain non-identifiable claim-specific information and are within the public domain
Nationwide Inpatient Sample [e]	Structured EHR	This dataset contains discharge information including diagnosis, procedures, status, demographics, cost and length of stay. It comprises 1051 hospitals of 45 states.
i2b2 Informatics for Integrating Biology & the Bedside [f]	Unstructured Clinical Data	Clinical notes used for clinical NLP challenges like deidentification, Smoking, Obesity, Medication, Relations and co-reference challenges

[a] http://www.dshs.texas.gov/thcic/hospitals/Inpatientpudf.shtm

[b] https://data.chhs.ca.gov/dataset/patient-discharge-data-by-admission-type/resource/460bd2e8-3b0e-4a41-b2a6-1044f7c82178

[c]3 https://epi.grants.cancer.gov/pharm/pharmacoepi.html

[d] https://www.cms.gov/Research-Statistics-Data-and-Systems/Downloadable-Public-Use-Files/BSAPUFS/index.html

[e] https://www.hcup-us.ahrq.gov/news/exhibit_booth/nis_brochure.jsp

[f] https://www.i2b2.org/

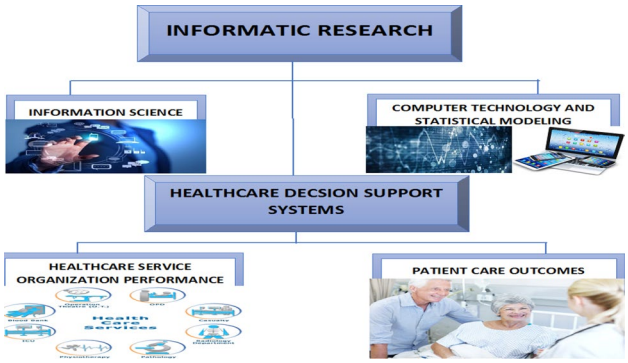

Fig. 17 Healthcare informatics researches

techniques to prevent illness. This information also serves as a guideline to the management of patients in whom disease has already evolved. Traditionally, epidemiology has been based on the data collected by public health agencies through health personnel in hospitals, doctors' offices, and out in the field.

The healthcare mechanism is an usual first line of reaction to clinical activities, whether of large or less severity. Informatics are used to figure out sentinel occasions, leading to analysis that can keep away from doubtlessly devastating effects. An example of response is *war on cancer* announced in 1973 when the programmers of National Institutes of Health feed the data from registries to the information system entitled with Surveillance Epidemiology and End Results (SEER) system. This system provides the information to the public health planner and epidemiologists to analyze the distribution of cancer throughout the population [180]. After many years of monitoring and evaluation, Age-adjusted mortality rates as a consequence of cancer were dropped step by step since early 1990, with important development in areas including lung. Most cancers reflecting fulfillment in public health efforts aimed at controlling precipitants to the disease [181].

Another example of that capacity can be seen inside the response to the 2001 bio-terrorism assaults. During September 2001, anthrax spores had been traced to postal facilities in Trenton, New Jersey and Brentwood, Washington. Epidemiologists face daunting venture: the new Jersey facility was a facility of 281,387 square ft, staffed by 250 employees according to shift and processing over 2 million items of mail in line with day [182]. Informatics help to become aware of the those who could have been exposed to anthrax, monitored the screening system, and recorded who obtained antibiotics and distribution of recognized cases and known deaths. Further analytical strategies and significant healthcare researches were explained in [183, 184].

In latest years, innovative data sources have introduced that are used to collect data in a second from individuals directly using electronic devices. Social media change the life of society and make global World. The exponential amount of data is produced daily. Big data are produced from Public Health (PH) information and can be generally characterized as big data. Public healthcare data are collected, analyzed, assured, and accessed so that big data analytics techniques are deployed to extract hidden informative patterns. Public or social media information is further used to predict, monitor, and diagnose of diseases, i.e., effective use of PH data determines the extent to which social health concerns can be determined. Literature divulges several survey papers based on data mining [185, 186], deep learning [187, 188] and other [189]. We here present some of the public healthcare work using social media. The data-sets for public healthcare data corpus are explained in the Table 4.

Young et al. [198] gathered 553,186,016 tweets from the Twitter. They extracted more than 9800 keywords and geographic annotations that contains HIV risk words. They revealed that social media monitor global HIV occurrence and concluded that positive correlation of greater than 0.01 was retrieved between HIV-related tweets and HIV cases. Hay et al. [199] facilitated public health surveillance using online social media combined with epidemiological information. They developed atlas for real-time disease monitoring.

Nambisan et al. [200] detected depression from messages and tweets of social media thus big data analytic tools were used to extract the hidden valuable patterns for detecting mental disorders. They concluded that behavioral and emotional patterns in messages showed the symptoms of depression. Tsugawa et al. [201] implemented multiple regression models to detect the depressive tendencies. They extracted frequency of words form messages and Twitter from the popular micro-blogging services to detect depression and achieved a correlation of approximately 0.5.

Park et al. [202] analyzed depression of 60 participants from their activities on tweeter from sentiment words of depressed users. Another contribution by the same author was to detect the symptoms of depressive users through Facebook [203]. Choudhury et al. [204, 205] developed a large dataset from Twitter posts using crowd sourcing methodology. They implemented the probabilistic model to indicate the depression level form social media. Postpartum depression affects up to 20% of mothersand has negative consequences for both mother and child [206]. Similarly, same authors in [207] detected and predicted the onset of post-partum depression of 165 mothers through Facebook shared data.

Sadelik et al. [208] predicted infectious diseases through the social network. They used 1000 Twitter messages related to healthcare. They applied statistical models on geo-tagged postings made on Twitter for prediction of diseases that cause an infection like flu, etc. Digital media are widely used to improve healthcare monitoring and its effectiveness.

Table 4 Public health databases

Database	Database type	Size	Description
Ohio Hospital Inpatient/Outpatient Database [190]	Public Patients Records	35 million patient records per year	This repository contains hospital record such as number of admissions, discharges, stay length, transfers, number of patients with specific codes
Behavioral Risk Factor Surveillance System (BRFSS) [191]	Survey System	50 states data of more than 400,000 adult interviews each year	This system contains record of mental illness, smoking, alcohol, lifestyle (diet, exercise) and diseases (diabetes, cancer), etc.
Surveillance Epidemiology and End Results (SEER) Program [192, 193]	Cancer Dataset	7.7M cases and more than 350,000 cases are added each year	This program contains survival data from population-based cancer registries covering approximately 28% US population.
PatientsLikeMe Online Patient Network Database [194, 195]	Online Patient Network Database	More than 200,000 patients and is tracking 1500 diseases	These data corpus contain information of disease-specific functional scores, symptoms, etc. through which people having same symptoms connect with each others.
Human Mortality Database [196, 197]	Public Mortality	39 countries or areas	This database contains information about population and mortality in detail along with Birth, death, population size by country.

Ginsberg et al. [209] used trends models and search queries on Google to detect influenza and flu-like diseases. One of the most earlier comprehensive review papers of public healthcare informatics using social media was presented by Hagg et al. [210]

6.5 Medical signal analytics

Nowadays, technology is advancing rapidly that provides effectiveness in every walk of life, especially in healthcare. Currently, healthcare systems use a variety of continuous monitoring devices that generate signals. Physiological signal monitoring devices and telemetry devices are pervasive [211] because these devices improve healthcare management and patient healthcare [212, 213]. These devices use discretized or physiological waveform data and generate alert mechanisms in case of an overt event. There are certain issues in medical signals that tend to move towards big data. The most notable obstacle is volume and velocity of continuous and high-resolution multitude monitors connected to each patient. The generated alarm systems are unreliable and cause alarm exhaustion for both caregivers and patients [214, 215]. The primary failure of these systems is due to relying on single source of information.

The first step in streaming data analytics in healthcare is the acquisition of signals. It is usually rare to store the streaming signals from continuous acquisition devices. However, to access the live streaming data from devices is one of the foremost tasks for big data analytics applications. There are many challenges posed to healthcare systems during streaming data collection like network bandwidth, scalability, and cost [216]. Thus, Research communities are developing continuous monitoring technologies [217] to capture live monitor signals. Next step is to store the signals data from monitoring devices using Big Data analytics tools like HDFS, MapReduce, MongoDB [218, 219], etc. Medical data including signals are complex due to interconnected and interdependent data among several sources. Thus, data are integrated and aggregation techniques are deployed for effective performance [220, 221]. The workflow of generalized streaming healthcare is depicted in Fig.18. The most notable data repositories containing signals information in healthcare are shortlisted in Table 5.

After introducing medical signal analytics, we present some of the related work of big data analytics in medical signaling. Han et al. [224] developed a patient care management system using a scaleable infrastructure. This system combined static and continuous data from monitored ICU devices. It analyzed and mined medical data in real time.

Bressan et al. [225] implemented an architecture for neonatal ICU. It used data of EEG monitors, infusion pumps,

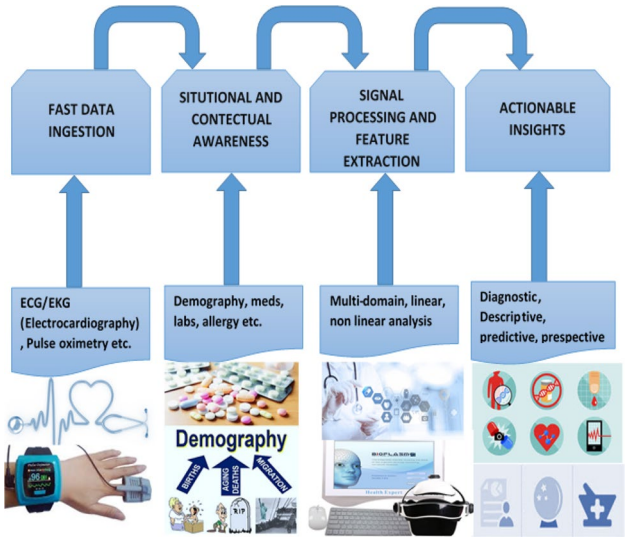

Fig. 18 Generalized work-flow of streaming healthcare

and cerebral oxygenation monitors. Their proposed system provides effective decision system for clinics.

Lee and Mark [226] conducted experiment on MIMIC II database for therapeutic intervention to hypotensive episodes. Their system predicted intensive care based using blood pressure and cardiac time series data.

Sun et al. [227] also used MIMIC II database to extract the physiological waveform data along with clinical data. They selected cohorts and find the similarity of patients from them that is beneficial for healthcare. The similarity measure was used for the treatment of similar diseases and deducted effective decisions from them. Another study on MIMIC II database was to detect the cardiovascular instability in patients at an early stage. For this purpose Cao et al. [228] developed a system that combined multiple waveform data from MIMIC II corpus.

Roux et al. [229] discussed the neuro-critical care of the patient's disorders using different physiological monitoring systems. They provide a platform for the researchers

Table 5 Medical signal databases (Shortlisted from Physionet)

Database	Database type	Size	Applications
The Multi-parameter Intelligent Monitoring in Intensive Care II (MIMIC-II) Clinical Database [179]	Structured EHR	32,536 subjects with 40,426 ICU admissions and 25,328 ICU stays	It contains comprehensive clinical data, including physiological waveforms and minute-by-minute records subsets.
MIMIC-III Database [222]	hospital Database	38,597 patients, 49,785 hospital admissions	This data corpus is informative with vital signs, laboratory measurements,medications, imaging reports, details of observations , fluid balance, diagnostic codes, procedure codes, stay length of hospital, survival data, etc.
The ECG-ID Database [a]	Signals Database	90 persons, 310 ECG recordings	EEG signal recordings each have 10 annotated beats, digitized at 500 Hz with 12-bit resolution and recorded for 20 seconds.
CinC Challenge 2000 data sets [223]	EEG signal based database	583 megabytes, 70 records	This dataset contain EEG signals of 70 records, used 35 records for learning set and 35 for testing
MIT-BIH Polysomnographic Database [223]	Physiologic Signals Database	18 records, each have 4 files	This database is the collection of recordings of multiple physiologic signals during sleep
EEG Motor Movement/Imagery Dataset [223]	EEG Signals Database	109 volunteers,1500 recordings	Two minutes EEG recordings, 64-channel EEG were recorded using the BCI2000 system
American Heart Association (AHA)[b]	EEG Signals Database	80 recordings	ECG recordings of 80 two-channel records digitized at 250 Hz per channel with 12-bit resolution with range of 10 mV.

[a] https://www.physionet.org/pn3/ecgiddb/

[b] https://www.physionet.org/physiobank/database/ahadb/

with guidelines by examining the potentials and implications of neuro-monitoring. Rajan et al. [230] used a multichannel signal acquisition method for the development of physiological signal monitoring system using NI myRIO connected with the wireless network. They also used the Internet of Things(IOT) techniques for better performance in healthcare.

Zhang et al. [231] recognized the Lung cancer using sensor-based wrist pulse signal processing with the technique of cubic support vector machine (CSVM). They implemented iterative slide window (ISW) algorithm for signal segmentation and extract 26 features. Using these strategies, they achieved 78.13% accuracy. Nanda et al. [232] distinguished between essential tremor and Parkinson's tremor using non-invasive recording techniques. They employed Neural Network for the classification of tremor sEMG signals and achieved 91.66% accuracy.

7 Key findings

This survey presents the emerging landscape of big data and analytical techniques in the five sub-disciplines of healthcare. We present various domains of healthcare in which big data technology has played a significant role in modern-day healthcare revolution, as it has totally changed the perception of people about healthcare activities. Big data analytical techniques deployed in five sub-disciplines such as, medical image processing and imaging informatics, bioinformatics, clinical informatics, public health informatics, medical signal analytics are explained comprehensively that draw an integrated depiction of how distinct healthcare activities are accomplished in a pipeline to facilitate individual patients from multiple perspectives. The existing reviews did not provide the detailed explanation in multiple sub-disciplines of healthcare. There is no comprehensive evaluation of studies in the existing reviews.

The existing studies discussed the different sources of healthcare for big data such as pharmaceutical firms, healthcare providers, diagnostic companies, laboratories, not-for-profit organizations insurance companies and web-health portals [10, 235–237]. The big data techniques used for the analysis of healthcare data are machine learning, data mining, cluster analysis, pattern recognition, neural networks, deep leaning and spatial analysis. Most of the studies processed the patient data using Hadoop and its tools, but they are batch processing tools [240–242]. There are some studies that used newer tools like Spark, Storm, GraphLab, etc. for the processing of real time and streaming data [242]. Most of the studies discussed the applications of big data analytics in different fields of healthcare like personalized medicine, clinical decision support, clinical operations optimization and cost effectiveness of healthcare. It can be shown that healthcare analytics improves the quality and early identification of patients. There are researches related to diabetes, gynecology, oncology, cardiovascular diseases and so on that enable to save time and cost [69, 245–247] (see Fig. 19).

With the rapid increase of publications in biomedical and healthcare industry, we have conducted the detailed review regarding healthcare analytics in five sub-disciplines. We summarized the usability studies of each discipline in Table 6, including image visualization, image

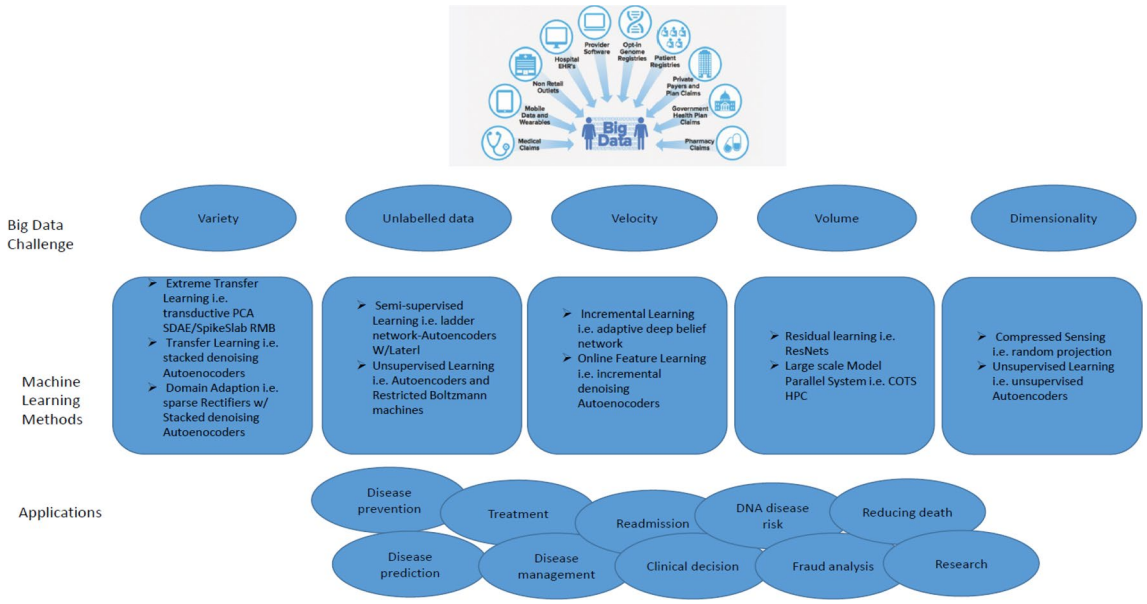

Fig. 19 Deep learning architecture for big data analytics

Table 6 Comparative analysis of the literature

Healthcare discipline	Big data analytical technique	Studies
Medical image processing and Imaging informatics	Image visualization	[77, 248–253]
	Image Classification	[71, 73, 78, 254–256]
	Image retrieval	[80, 81, 257–259]
	Data and Workflow sharing	[260–262, 262]
	Data analysis	[82, 259, 263]
Bioinformatics	Feature selection	[103–105, 264–266]
	Classification	[112, 113, 115–118, 264, 267, 268]
	Clustering	[56, 119–121, 124, 125, 269]
	Microarray Data Analysis	[127–131]
	Protein–Protein interaction	[132–135]
	Pathway analysis	[136–141]
	Protein sequencing	[142–144]
	Protein query and Search engine	[145–147]
	Error identification of sequencing data	[148–157]
Clinical informatics	Storage of EHR	[160–162, 164–169, 177]
	Retrieval of EHR	[172, 270]
	Interactive data retrieval for data sharing	[54, 170, 170, 171, 271–273]
	Treatment recommendation	[79, 82, 83, 274]
	Business transformation	[69, 178, 275–278]
	Disease predication, Diagnosis and progression	[173–176, 279]
	Data Security	[280–283]
Public health informatics	Infectious disease surveillance	[198, 199, 284]
	Population health management	[285–289]
	Mental health management	[200, 290–292]
	Chronic disease management	[293–295]
Medical signal analytics	Signal acquisition	[216, 217, 226, 227, 230]
	Signal storing from monitoring devices	[218, 219, 224, 225, 228, 229]
	Signal integration and aggregation	[220, 221, 231, 232]

classification, image retrieval, data and workflow sharing, data analysis, feature selection, bioinformatics classification and clustering, micro-array data analysis, protein–protein interaction, pathway analysis, protein sequencing, query and search engine, error identification of sequencing data, storage and retrieval of EHR, treatment recommendation, business transformation, disease prediction, diagnosis and progression, data security, infectious disease surveillance, population health management, mental health management, chronic disease management, signal acquisition, signal storing from monitoring devices, signal integration and aggregation respectively. It is concluded from this survey, that bioinformatics is one of the primary disciplines in which big data analytics is currently evolved and playing a scientific role, due to the complex and massive bioinformatics data. There are a lot of tools, techniques and platforms for bioinformatics used to analyze biological, genomics, proteins and gene sequencing data. However, there is less potential of big data applications in other disciplines such as medical imaging informatics, clinical informatics, public health informatics and medical signal analytics.

8 Big data analystics applications

Healthcare sector produces huge amounts of patient data on a daily basis. Traditionally, most of these data were used to be in the form of hard copies but, due to the advancement in data acquisition devices, healthcare organizations are gathering data electronically. Healthcare data analytics has the potential to bring in dramatic changes in healthcare industry to smooth the process and improving the quality of care. Data analytic researchers, healthcare providers, government agencies and the pharmaceutical companies identify range of different ways that big data techniques can help us to significantly improve patient outcome through policy making and evidenced-based

decisions. Below are the major areas in healthcare sector where big data analytics has a huge impact:

Strategic planning *'Management is based on early measure: you cant manage if you cant measure'*. Healthcare is a time critical service. Hospitals are struggling with patient flow. Machine learning and data analytics plays important role in the prediction of patient flow and ensuring smooth patient flow as well as reducing waiting period. Early predicting of hospital visit helps the management to decide and take the necessary step to reduce patient waiting time thereby giving timely treatment. There are different applications like Patient Flow Manager and Q-nomy's etc. that provide a comprehensive graphical view patient low information, drawing of inpatient, elective, emergency, outpatient and other hospital systems. For example, care mangers can analyze check-up results among patients in different demographic groups that help to identify what factors discourage patient from taking up treatment. The classical example is staff management: how many clinicians and nurses staff should be give at specific time.

For our first example of big data in healthcare, we will look at one classic problem that any shift manager faces: how many people do I put on staff at any given time period? If you put on too many workers, you run the risk of having unnecessary labor costs add up; too few workers, you can have poor customer service outcomes—which can be fatal for patients in that industry. In other example, we can predict admission trend based on admission history of last few years, i.e., using 10 years worth of hospital admissions records, which data scientists crunched using "time series analysis" techniques followed by machine learning relevant to predicted future admissions trends.

Fraud detection *'Suspect, detect and protect'*. Fraud, waste, and abuse have caused significant cost and it range from honest mistakes that result in erroneous billings, inefficiencies that may result in wasteful diagnostic tests, overpayments due to false claims. Personal data are extremely sensitive due to its profitable value in black-markets; thus, healthcare industry is 200% more likely to experience data breaches than any other. With that in mind, effective detection of frauds is very important for reducing the cost and improving the quality of healthcare system. Fraud detection in healthcare is an important yet difficult problem. Big data have inherent security issues and healthcare organization are more vulnerable than they already are. Many organizations are using analytics to reduce security threats by analyzing the changes in network traffic, or suspicious behavior that reflects a cyber-attack. WhiteHatAI Centaur system, NICE ACTIMIZE, NHCAA, SAS, Optum, etc. are being used for medical claims processing that identify and detect healthcare fraud, waste, and abuse before it happens. Likewise, data analytic can help to prevent fraud and inaccurate claims in a systemic, repeatable way by streamlining the process of insurance claims. For example, the Centers for Medicare and Medicaid Services saved over $210.7 million in frauds in just 1 year.

Resource management *'How you use a facility, many factors pushing and pulling'*. Big data are making huge advances in reducing hospital waiting lists. Despite expensive efforts by the government and healthcare organizations, waiting times barely changed, with the median even increasing slightly, i.e., Australia has been trying hard to reduce the waiting list times on its hospital for more than two decades. Efficient and timely resource utilization helps to over come the patient flow and reduces the financial burden on organization. Data analytics continues to make inroads the manage hospital resources efficiently with respect to patient flow and risk. Examples are readmission, ambulance, bed utilization, etc.

The common example is 30-day patient readmission or return visit to an emergency department. 30-day readmission identifies the patients who have high possibility to return to hospital with 30 days of discharge. The development of risk prediction model helps to identify patients who would benefit from the disease management program in an effort to not only reduce the patient readmissions but also healthcare cost.

Personalized medicine *'Disease and its treatment is unique as we are'*. The promise of personalized medicine is the shift away from 'one size fits all' medicine. Through the datafication and genomic fingerprints, much more information of each patient can be analyzed without requiring multiple rounds of testing. Best treatment can be made on an individual basis at a faster rate using personalized data.

Genomics *'The more is the data you have the better you can treat'*. Human body consists of 30,000–35,000 genes [88, 296]. From the DNA structure of the human, it is estimated that there are 23 chromosomes with the distribution of 3.2 billion base pairs [297, 298]. These data increase dramatically to about 200 gigabytes. Thus, big data analytics is required for genomics and sequencing practices that are used for the treatment of complex diseases like Crohn's disease and age-related muscular degeneration [91]. The impact of genomic data analytics has the great potential to improve healthcare outcomes, quality, and safety, as well as cost savings.

Disease prediction and prevention *'Precaution and care can help live longer'* . Many healthcare organizations, research labs, hospitals are leveraging big data analytics are by changing the models of treatment delivery. Thus, big data analytics has tremendous applications in the healthcare domain for reducing cost overhead, detecting and curing diseases, predicting epidemics and enhancing the worth of human life by averting deaths. Number of projects from "Google", "DeepMind", "IBM", "Royal Free London NHS Foundation Trust" and "Imperial College Healthcare NHS

Trust" and others have proved the importance of deep learning and machine learning for detection, identification, diagnostics and predictive analytics. DeepMind collaborated with Moorfields Eye Hospital to the analyze anonymized eye scans, searching for early signs of diseases leading to blindness. There are also projects signed with the Royal Free London NHS Foundation Trust and Imperial College Healthcare NHS Trust to develop new clinical mobile apps linked to EHR.

Big data have transformed healthcare by putting data to work, revealing clinical and operational insights. The most applicable applications of IBM are IBM Content and Predictive Analytics. "IBM Content and Predictive Analytics" for healthcare is the first industry-specific analytics solution to enable organizations to analyze the past, see the present and predict the future by simultaneously. For example, we can predict admission trend based on admission history of last few years, i.e., using 10 years worth of hospital admissions records, which data scientists crunched using "time series analysis" techniques followed by machine learning relevant to predicted future admissions trends. One of the major applications of big data analytics in the healthcare domain is medical image processing, as in healthcare enormous amount of medical images are produced like X-ray, CT and PET–CT images, MRI, ultrasound, fluoroscopy and photoacoustic imaging. These medical images produced big data that are used for various purposes like detection, diagnoses, assessment, decision-making of therapy, etc. [299].

Heart is the basic organ of the body. If the heart stops its working, human body does not exist. There are several disorders of heart; one among them is the heart attack. Big data analytics facilitates to predict the heart attack at the early stage using early heart attack detection system based on medical biosensor [300, 301] that detects heart attack at the early stage. There are some online systems [302] and healthcare information system [303] that provides guidance about heart diseases using IOT and Hadoop techniques.

The brain is the vital organ of the body that controls all the activities of the body just like CPU of the computer. Thus, data mining and data analytics tools are deployed to detect the brain disorders like Parkinson's brain disease prediction [304] ,[?],[72, 305, 306]. Diabetics is one of the common diseases in this world. Big data analytics tools like *'Hive'* and *'R'* are used for the analysis of diabetics using descriptive dataset [307, 308]. Efficient predictive models are established to reveal the data related to the investigation of diabetics.

There are online applications that are remotely facilitating the healthcare domain. AmWell, Practo, Portea, Isabel, etc. are the most popular apps that are used for various purposes like appointment of doctors at hospitals, clinics, etc., patient diagnosis, ordering medicines, consultation with the doctor remotely for treatment, etc [309]. Summarizing the applications of big data analytics in healthcare [2, 310], it is concluded that big data are beneficial to identify and diagnose the patient accurately and precisely. It is used for the prediction and management of health risks and obesity to efficiently detect the level of frauds. It reduced the cost, variations, and elimination of duplicate care and improper claim submission.

9 Challenges and open research issues

The healthcare sector suffers from multiple challenges, ranging from new disease outbreaks to preserving an optimal operational efficiency. To overcome these challenges, data mining and data analytics in the development of applications of healthcare have tremendous potential; however, success hinges on the availability of quality data but there is no magic recipe to successfully apply data analytics methods on any problem. Thus, the successful development of data analytics-based applications depends on how data are stored, prepared and mined. However, chemical analytics poses a series of challenges when dealing with a enormous amount of complex data. These challenges involve data complexity, access to data, regulatory compliance, information security and efficient analytics methods, inter-operability, manageability, security, development, re-usability,open data, missing data and data heterogeneity.

9.1 Multiple source information management

In healthcare data analytics, the main goal is to analyze the real-world medical data to perform prediction or classification task. One of the biggest hurdles in development of such application depends upon on the data structure, i.e., how medical data are spread across many sources, how data are stored, prepared and mined. One of the worst examples of lack of data sharing is: a woman who was suffering from mental illness and substance abuse visited variety of local hospitals more than 900 times in a period of less than 3 years in Oakland, California, USA. It results in heavy cost, extensive use of hospital resources and more important, harder for woman to get good care.

Healthcare data correlations are leveraging in longitudinal records i.e., complex, heterogeneous, distributed and dynamic data i.e., in the US alone, healthcare data extended to 150 exabytes in 2011 and is expected to reach the zettabyte scale soon. Despite the rapid increase in EHR adoption, there are several challenges around making this information useful, readable and relevant to the physicians and patients who need it most. One of the key challenges in the healthcare industry is how to manage, store and exchange all of these data. Inter-operability is considered to be one of the solutions to this problem. There exists a poor inter-operability in

EHRs that creates big data analytics challenging in healthcare. Integration of different data sources would require developing a new infrastructure where all data providers can collaborate each other to share. Another challenge is data privacy that limits the sharing of data by blocking out significant patient identification information such as MRN and SSN. Healthcare needs to catch up with other industries that have already moved from standard regression-based methods to more future-oriented like predictive analytics, machine learning, and graph analytics. Big data technologies like Data ingestion, data modeling, and data visualization are integrated with existing tools to provide a supported enterprise solution.

Big data management is one of the hard tasks as there is a big cluster of data that are monitored and managed. Most patients visit multiple clinics to try to find a reason for their disease and medical solution for their illness. To overcome this issue, several management tools are integrated that is overwhelming and cost effective strategy. Proficiently handling large capacities of medical imaging data and extracting possibly useful information is another hard task. Hospitals have yet to achieve a level of inter-operability, and without it, it is almost impossible to improve patient care. The US Health Department is aiming for inter-operability between disparate EHRs by 2024. Medical stakeholders (physicians, administrators, patients, etc.) believe that inter-operability will improve patient care, reduce medical errors and save costs. Imagine having the insight and opinions of hundreds of IVF/PGD patients to assist your decision before undergoing treatment rather than only relying on a physician's recommendations. Due to the importance of data integration, healthcare organizations are turning to the implementation of inter-operability. To achieve a high level of inter-operability, HL7, HIPPA, HITECH and other health standardization bodies have demarcated several standards and guidelines to assist organizations to know whether they meet inter-operability and security standards. The Authorized Testing and Certifying Body (ATCB) provides a sovereign, third-party opinion on EHR. Two types of certification (CCHIT and ARRA) are used to evaluate the system. The review process comprises standardized test scripts and exchange tests of standardized data. Healthcare industry needs to catch up with other fields that have already progress from standardization.

9.2 Security, privacy, and confidentiality

Every stakeholder in the health industry has a role to play in ensuring the security and privacy of patient information. It is a shared responsibility. Patient privacy and information security are fundamental components of a well-functioning healthcare system that helps to accomplish better health outcomes, healthier people, and smarter spending. For example, a patient may not disclose certain information or may ask a physician not to record his health information due to a lack of trust and the perception that this information might not be kept confidential. This attitude puts the patient at risk and deprives physicians and researchers of important information as well as putting the organization at risk in terms of clinical outcomes and operational efficiency analysis. To reap the benefits, providers and individuals must belief that patients' health information is kept private and secure. On the other hand, providers are facing several challenges in ensuring that privacy and security issues are managed at a standard that meets the patients' satisfaction, i.e., efficient data analysis without providing access to precise data in specific patient records. Security and privacy in data analytics poses several challenges, especially when it draws information from multiple sources.

The major goal in healthcare is not to protect the patient's privacy rather it is to save lives. The HIPAA (Health Insurance Portability and Accountability Act) of 1996 comes to mind when privacy is debated in the health sector. It delivers legal rights to patients concerning their personally identifiable information and establishes responsibilities for healthcare providers to defend and restrict its use or disclosure. With the escalation in the amount of healthcare data, data analytics researchers envisage huge challenges in ensuring the anonymity of patient information to avoid its use or disclosure. Limiting data access, unfortunately, reduces information content which might be very important. Moreover, real data are not static but grow larger and vary over time and none of the existing techniques result in any convenient content being released in this scenario.

9.3 Advanced analyzing techniques

Technological advancements (wearable devices, patient-centered care, etc.) are transforming the entire healthcare industry. The nature of healthcare data has progressed, and currently, EHRs have simplified the data acquisition process with the help of the latest technology, but, unfortunately, they do not have the ability to aggregate, transform, or perform analytics on it. Intelligence is restricted to retrospective reporting that is insufficient for data analysis. A plethora of algorithms, techniques, and tools are available for the examination of complex data. Traditional machine learning deploys statistical analysis based on a sample of a total dataset. The use of traditional machine learning methods for these data is not efficient and is computationally infeasible. The combination of the huge volume of healthcare data and computational power lets the analysts to focus on analytics techniques which are scaled up to accommodate the volume, velocity, and variety of complex data. During the last decade, there has been a melodramatic change in the size

and complexity of data; thus, several emerging data analysis techniques have been presented.

Healthcare needs to catch up with other industries that have already progressed from traditional methods to advance methods like predictive analytics, deep machine learning, and graph analytics. Innovative analytics techniques need to be developed to interrogate healthcare data and gain insight into hidden patterns, trends, and associations in the data. It deduces relationships without the need for a specific model and enables the machine to identify the patterns of interest in huge unstructured data. As one example, a deep learning algorithm that observed data from Wikipedia learned on its own that California and Texas are both states in the U.S. It does not have to be modeled to understand the conception of a country and state, and this is a gigantic difference between older machine learning and emerging deep learning methods.

9.4 Data quality: open data, missing data, and data heterogeneity

Gone are the days when healthcare data were small, structured and collected exclusively in electronic health records. Due to the tremendous advancements in IT, wearable technology and other body sensors, data have become quite large (moving to big data), unstructured (80% of electronic healthcare data are unstructured), non-standard as well as in a multimedia format. This variety in data makes it challenging and interesting for analysis. Currently, the quality of healthcare data is a cause of concern for four reasons, incompleteness (missing data), inconsistency (data mismatch between within same or various EHR sources), inaccuracy (non-standard, incorrect or imprecise data), heterogeneity, and data fragmentation. Data quality involves a group of different techniques, these being data standardization, verification, validation, monitoring, profiling, and matching. The problem of poor data in the health industry has reached epidemic proportions and introduces several pernicious effects, particularly in relation to disease prevention. The problem with dirty data is mostly related to missing values, duplication, outliers and stale records.

Although real-time data monitors (especially in ICUs) are partially used in most hospitals, real-time data analytics is not in practice. Hospitals are moving to real-time data collection and in the near future, real-time data analytics will revolutionize the healthcare industry, enabling such things as the early identification of infections, the continuous monitoring of the progress of treatment, the selection of the right drugs, etc. which could help to reduce morbidity and mortality. To achieve real-time data processing, we need data standardization and device inter-operability.

The other common issue is data standardization. Structuring of only 20 percent of data has shown its importance but on the other hand, clinical notes are still in practice and created in billions due to the reason that the physician can best explain the clinical encounter. Empower physicians as well as maintaining the data quality is quite challenging. So far, these data are excluded from data analytics as it is available in the natural language and not discrete. Transforming this unstructured data into a discreet form requires efficient intelligent technology and it is has been a very difficult problem for medical IT until now. The only way this unstructured and nonstandard data can be used is using NLP to translate the data using ICD or SNOMED CT into discrete data.

Heterogeneous data are any information with high variability of information types. They are low-quality and ambiguous data due to high information access, data redundancy, missing values, and untruthfulness. It is difficult of integrate heterogeneous information to satisfy the business data needs. For instance, heterogeneous information are frequently produced from Internet of Things (IoT). The challenges of Big Data algorithms concentrate on algorithm design in tackling the difficulties raised by big data volumes, distributed data distributions, and complex and dynamic data characteristics. The challenges include the following stages. First, heterogeneous, incomplete, uncertain, sparse, and multi-source data are pre-processed by data fusion techniques. Second, dynamic and complex data are mined after pre-processing. Third, the global knowledge obtained by local learning and model fusion is tested and relevant information is fed back to the preprocessing stage. Then, the model and parameters are adjusted according to the feedback. In the whole process, information sharing is not only a promise of smooth development of each stage, but also a purpose of big data processing [311].

10 Conclusion

The exponential growth of big data analytics has rapidly increased that plays a vital role in the progression of healthcare practices and research. It includes providing tools to collect, analyze, manage and store a large volume of structured, unstructured and large complex data. Big data have brought a dramatic change in healthcare which reduce the cost of treatment and accelerate the identification of disease, cancer, etc. and improve the life's quality. They have been recently applied in aiding in the process of healthcare personnel, care delivery, early disease detection, disease exploration, patient care, and community services.

In this paper, we have discussed the big data analytics methods, tools, techniques and architectures in the healthcare domain. We have focused on five major sub-disciplines of healthcare, i.e., medical image processing and imaging informatics, bioinformatics, clinical informatics, Public Health informatics and Medical Signal analytics along with

techniques, tools, and repositories deployed in each discipline. These disciplines play a vital role in healthcare and bio-medical due to the enormous amount of data.

Healthcare providers had no direct incentive in sharing the patient information with each other, that made it harder to efficiently utilize the power of analytics in healthcare industry. We can possibly change the way to healthcare providers use modern advances and sophisticated technologies to pick up understanding from their clinical, data warehouses, information storehouses for extracting informative patterns and decision-making. Later on, we will see the quick, across the board execution and utilization of big data analytics over the social insurance association and the medicinal services industry. Keeping that in mind, the few difficulties must be tended to. Its potential is extraordinary; however, issues, for example, multiple source information management, ensuring protection, shielding security, setting up models and administration, advance analyzing techniques and data quality are the notable challenges in the domain. Regardless, the future trends of big data in the social insurance framework have the capability of enhancing and quickening communications among clinicians, executive, logistic manger, and analyst by diminishing costs, reducing risks and improving personalized care.

Implementation of big data analytic is the responsibility for all stakeholders in healthcare industry. It is the responsiblity of stakeholders to make and review the polices of big data to improve the patient outcomes. Government agencies, healthcare professionals, hardware companies, pharmaceutical industries, people, data scientist, researchers, and vendors must be involved in developing the big data framework that will provide the future direction of big data analytics in healthcare industry.

References

1. Frost, S.: Drowning in big data? reducing information technology complexities and costs for healthcare organizations (2015)
2. Raghupathi, W., Raghupathi, V.: Big data analytics in healthcare: promise and potential. Health Inf. Sci. Syst. **2**(1), 3 (2014)
3. Baro, E., Degoul, S., Beuscart, R., Chazard, E.: Toward a literature-driven definition of big data in healthcare. BioMed Res. Int. (2015)
4. Burghard, C.: Big data and analytics key to accountable care success. In: IDC Health Insights pp. 1–9 (2012)
5. Dembosky, A.: Data prescription for better healthcare. Financial Times **11**(12), 2012 (2012)
6. Feldman, B., Martin, E.M., Skotnes, T.: Big data in healthcare hype and hope. Dr. Bonnie **360**, 122–125 (2012)
7. Fernandes, L.M., O'Connor, M., Weaver, V.: Big data, bigger outcomes. J. AHIMA **83**(10), 38–43 (2012)
8. Vayena, E., Salathé, M., Madoff, L.C., Brownstein, J.S.: Ethical challenges of big data in public health. PLoS Comput. Biol. **11**(2), e1003904 (2015)
9. Wyber, R., Vaillancourt, S., Perry, W., Mannava, P., Folaranmi, T., Celi, L.A.: Big data in global health: improving health in low- and middle-income countries. Bull. World Health Org. **93**(3), 203–208 (2015)
10. Ward, M.J., Marsolo, K.A., Froehle, C.M.: Applications of business analytics in healthcare. Bus. Horizons **57**(5), 571–582 (2014)
11. Sessler, D.I.: Big data-and its contributions to peri-operative medicine. Anaesthesia **69**(2), 100–105 (2014)
12. Razzak, M.I., Imran, M., Xu, G.: Big data analytics for preventive medicine. Neural Comput. Appl., 1–35 (2019)
13. Ericsson Mobility Report February Interim 2018. https://www.ericsson.com/491b06/assets/local/mobility-report/documents/2019/ericsson-mobility-report-q4-2019-update.pdf (2018)
14. Available:: Internet world stats. https://www.internetworldstats.com/stats.htm (2018)
15. Manogaran, G., Lopez, D.: A survey of big data architectures and machine learning algorithms in healthcare. Int. J. Biomed. Eng. Technol. **25**(2–4), 182–211 (2017)
16. Wilhelm, M., Schlegl, J., Hahne, H., Gholami, A.M., Lieberenz, M., Savitski, M.M., Ziegler, E., Butzmann, L., Gessulat, S., Marx, H., et al.: Mass-spectrometry-based draft of the human proteome. Nature **509**(7502), 582 (2014)
17. Ackerman, M.J.: The visible human project: a resource for education. Acad. Med. **74**(6), 667–670 (1999)
18. Gui, H., Zheng, R., Ma, C., Fan, H., Xu, L.: An architecture for healthcare big data management and analysis. In: International Conference on Health Information Science, pp. 154–160. Springer (2016)
19. Razzak, M.I., Naz, S., Zaib, A.: Deep learning for medical image processing: Overview, challenges and the future. In: Classification in BioApps, pp. 323–350. Springer (2018)
20. Cox, M., Ellsworth, D.: Application-controlled demand paging for out-of-core visualization. In: Proceedings of the 8th Conference on Visualization'97, pp. 235–ff. IEEE Computer Society Press (1997)
21. Kochański, A.: Data preparation. Comput. Methods Mater. Sci. **10**(1), 25–29 (2010)
22. Diebold, F.X.: Big data dynamic factor models for macroeconomic measurement and forecasting. In: Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress of the Econometric Society, "(edited by M. Dewatripont, LP Hansen and S. Turnovsky), pp. 115–122 (2003)
23. Laney, D.: 3d data management: Controlling data volume, velocity and variety. META Group Res. Note **6**(70), 1 (2001)
24. O'Reilly, T., Steele, J., Loukides, M., Hill, C.: Solving the wanamaker problem for healthcare (2012)
25. Shin, D.: Demystifying big data: anatomy of big data developmental process. Telecommun. Policy **40**(9), 837–854 (2016)
26. Emani, C.K., Cullot, N., Nicolle, C.: Understandable big data: a survey. Comput. Sci. Rev. **17**, 70–81 (2015)
27. Chen, M., Mao, S., Liu, Y.: Big data: A survey. Mob. Netw. Appl. **19**(2), 171–209 (2014)
28. Groves, P., Kayyali, B., Knott, D., Van Kuiken, S.: The 'big data'revolution in healthcare. McKinsey Q. **2**, 3 (2013)
29. Eynon, R.: The rise of big data: what does it mean for education, technology, and media research? (2013)
30. Porche, D.J.: Men's health big data (2014)
31. Berger, M.L., Doban, V.: Big data, advanced analytics and the future of comparative effectiveness research. J. Comp. Effect. Res. **3**(2), 167–176 (2014)
32. BERNARD, E.: Supporting diagnosis and treatment in medical care based on big data processing. In: Cross-Border Challenges in Informatics with a Focus on Disease Surveillance and Utilising Big Data: Proceedings of the EFMI Special Topic Conference,

27-29 April 2014, Budapest, Hungary, vol. 197, p. 65. IOS Press (2014)
33. Watson, H.J.: Tutorial: Big data analytics: concepts, technologies, and applications. CAIS **34**, 65 (2014)
34. McAfee, A., Brynjolfsson, E., Davenport, T.H., Patil, D., Barton, D.: Big data: the management revolution. Harv. Bus. Rev. **90**(10), 60–68 (2012)
35. Russom, P., et al.: Big data analytics. TDWI Best Pract. Rep. Fourth Quarter **19**(4), 1–34 (2011)
36. Saporito: The 5 v's of big data: value and veracity join three more crucial attributes that carriers should consider when developing a big data vision. https://www.thefreelibrary.com/The+5+V (2021)
37. Sathi, A.: Big Data Analytics: Disruptive Technologies for Changing the Game. MC Press, Chennai (2012)
38. Manogaran, G., Lopez, D.: Health data analytics using scalable logistic regression with stochastic gradient descent. Int. J. Adv. Intel. Paradig. **10**(1–2), 118–132 (2018)
39. Tsai, C.W., Lai, C.F., Chao, H.C., Vasilakos, A.V.: Big data analytics: a survey. J. Big Data **2**(1), 1–32 (2015)
40. James, R.: Out of the box: Big data needs the information profession-the importance of validation. Bus. Inf. Rev. **31**(2), 118–121 (2014)
41. Gandomi, A., Haider, M.: Beyond the hype: Big data concepts, methods, and analytics. Int. J. Inf. Manag. **35**(2), 137–144 (2015)
42. Razzak, M.I., Saris, R.A., Blumenstein, M., Xu, G.: Robust 2d joint sparse principal component analysis with f-norm minimization for sparse modelling: 2d-rjspca. In: 2018 International Joint Conference on Neural Networks (IJCNN), pp. 1–7. IEEE (2018)
43. Naz, S., Umar, A.I., Ahmad, R., Siddiqi, I., Ahmed, S.B., Razzak, M.I., Shafait, F.: Urdu nastaliq recognition using convolutional–recursive deep learning. Neurocomputing **243**, 80–87 (2017)
44. Razzak, I., Saris, R.A., Blumenstein, M., Xu, G.: Integrating joint feature selection into subspace learning: A formulation of 2dpca for outliers robust feature selection. Neural Netw. (2019)
45. Naz, S., Umar, A.I., Ahmad, R., Ahmed, S.B., Shirazi, S.H., Siddiqi, I., Razzak, M.I.: Offline cursive urdu-nastaliq script recognition using multidimensional recurrent neural networks. Neurocomputing **177**, 228–241 (2016)
46. Holland, S.M.: Principal components analysis (pca), pp. 30602–2501. Department of Geology, University of Georgia, Athens, GA pp (2008)
47. SVD, S.V.D.: Singular value decomposition. 593–594 (2014)
48. Schölkopf, B., Smola, A., Müller, K.R.: Kernel principal component analysis. In: International Conference on Artificial Neural Networks, pp. 583–588. Springer (1997)
49. Sammon, J.W.: A nonlinear mapping for data structure analysis. IEEE Trans. Comput. **100**(5), 401–409 (1969)
50. De Ridder, D., Duin, R.P.: Sammon's mapping using neural networks: a comparison. Pattern Recognit. Lett. **18**(11–13), 1307–1316 (1997)
51. Belkin, M., Niyogi, P.: Laplacian eigenmaps for dimensionality reduction and data representation. Neural Comput. **15**(6), 1373–1396 (2003)
52. Pareto, V.: Cours d'économie politique, vol. 1. Librairie Droz, Geneva (1964)
53. Horn, J., Nafpliotis, N., Goldberg, D.E.: A niched pareto genetic algorithm for multiobjective optimization. In: Evolutionary Computation, 1994. IEEE World Congress on Computational Intelligence., Proceedings of the First IEEE Conference on, pp. 82–87. Ieee (1994)
54. Deb, K.: Multi-Objective Optimization Using Evolutionary Algorithms, vol. 16. Wiley, Hoboken (2001)
55. Bäck, T.: Evolutionary computation: toward a new philosophy of machine intelligence (1997)
56. Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data: An Introduction to Cluster Analysis, vol. 344. Wiley, Hoboken (2009)
57. Zhang, T., Ramakrishnan, R., Livny, M.: BIRCH: an efficient data clustering method for very large databases. ACM Sigmod Rec. **25**(2), 103–114 (1996)
58. Dean, J., Ghemawat, S.: Mapreduce: simplified data processing on large clusters. Commun. ACM **51**(1), 107–113 (2008)
59. Stanton, I., Kliot, G.: Streaming graph partitioning for large distributed graphs. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1222–1230. ACM (2012)
60. Shvachko, K., Kuang, H., Radia, S., Chansler, R.: The hadoop distributed file system. In: 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST), pp. 1–10. IEEE (2010)
61. Capriolo, E., Wampler, D., Rutherglen, J.: Programming Hive: Data Warehouse and Query Language for Hadoop. O'Reilly Media Inc, Newton (2012)
62. Wulff, F.: Presto. https://prestodb.io/ (2013)
63. Hortonworks: Apache mahout. http://hortonworks.com/hadoop/mahout/ (2015)
64. Confluent: Avro. http://docs.confluent.io/1.0/avro.html (2015)
65. Razzak, I., Blumenstein, M., Xu, G.: Multiclass support matrix machines by maximizing the inter-class margin for single trial eeg classification. IEEE Trans. Neural Syst. Rehabil. Eng. (2019)
66. Wamba, S.F., Akter, S., Edwards, A., Chopin, G., Gnanzou, D.: How 'big data'can make big impact: findings from a systematic review and a longitudinal case study. Int. J. Prod. Econ. **165**, 234–246 (2015)
67. Zhang, Y., Li, X.: Uses of information and communication technologies in hiv self-management: A systematic review of global literature. Int. J. Inf. Mang. **37**(2), 75–83 (2017)
68. Jacofsky, D.J.: The myths of 'big data' in health care. Bone Jt J. **99**(12), 1571–1576 (2017)
69. Wang, Y., Kung, L., Wang, W.Y.C., Cegielski, C.G.: An integrated big data analytics-enabled transformation model: application to health care. Inf. Mang. **55**(1), 64–79 (2018)
70. Galetsi, P., Katsaliaki, K.: A review of the literature on big data analytics in healthcare. J. Oper. Res. Soc., 1–19 (2019)
71. Rehman, A., Naz, S., Razzak, M.I., Akram, F., Imran, M.: A deep learning-based framework for automatic brain tumors classification using transfer learning. Circ. Syst. Signal Process. **39**(2), 757–775 (2020)
72. Naz, A.R.S., Naseem, U., Razzak, I., Hameed, I.A.: Deep autoencoder-decoder framework for semantic segmentation of brain tumor. Austral. J. Intell. Inf. Process. Syst., 53
73. Rehman, A., Khan, F.G.: A deep learning based review on abdominal images. Multimed. Tools Appl. 1–32 (2020)
74. Shirazi, S.H., Umar, A.I., Naz, S., Razzak, M.I.: Efficient leukocyte segmentation and recognition in peripheral blood image. Technol. Health Care **24**(3), 335–347 (2016)
75. Razzak, I., Imran, M., Xu, G.: Efficient brain tumor segmentation with multiscale two-pathway-group conventional neural networks. IEEE J. Biomed. Health Inf. (2018)
76. Naz, S., Umar, A.I., Ahmad, R., Ahmed, S.B., Shirazi, S.H., Razzak, M.I.: Urdu nasta'liq text recognition system based on multi-dimensional recurrent neural network and statistical features. Neural Comput. Appl. **28**(2), 219–231 (2017)
77. Gessner, R.C., Frederick, C.B., Foster, F.S., Dayton, P.A.: Acoustic angiography: a new imaging modality for assessing microvasculature architecture. J. Biomed. Imaging **2013**, 14 (2013)
78. Shackelford, K.: System & method for delineation and quantification of fluid accumulation in efast trauma ultrasound images. US Patent App. 14/167,448 (2014)
79. Chen, W., Cockrell, C., Ward, K.R., Najarian, K.: Intracranial pressure level prediction in traumatic brain injury by extracting

features from multiple sources and using machine learning methods. In: Bioinformatics and Biomedicine (BIBM), 2010 IEEE International Conference on, pp. 510–515. IEEE (2010)
80. Yao, Q.A., Zheng, H., Xu, Z.Y., Wu, Q., Li, Z.W., Yun, L.: Massive medical images retrieval system based on hadoop. J. Multimed. **9**(2), 216–222 (2014)
81. Jai-Andaloussi, S., Elabdouli, A., Chaffai, A., Madrane, N., Sekkaki, A.: Medical content based image retrieval by using the hadoop framework. In: Telecommunications (ICT), 2013 20th International Conference on, pp. 1–5. IEEE (2013)
82. Dilsizian, S.E., Siegel, E.L.: Artificial intelligence in medicine and cardiac imaging: harnessing big data and advanced computing to provide personalized medical diagnosis and treatment. Curr. Cardiol. Rep. **16**(1), 441 (2014)
83. Istephan, S., Siadat, M.R.: Unstructured medical image query using big data—an epilepsy case study. J. Biomed. Inf. **59**, 218–226 (2016)
84. O'Driscoll, A., Daugelaite, J., Sleator, R.D.: 'big data', hadoop and cloud computing in genomics. J. Biomed. Inf. **46**(5), 774–781 (2013)
85. Robison, R.J.: https://medium.com/precision-medicine/how-big-is-the-human-genome-e90caa3409b0 (2014)
86. Kashya, H., Ahmed, H.A., Hoque, N., Roy, S., Bhattacharyya, D.K.: Big data analytics in bioinformatics: a machine learning perspective. J. Latex Class Files **13**(9), 837–854 (2014)
87. Lander Eric, S., Linton Lauren, M., Bruce, B., Chad, N., Zody Michael, C., Jennifer, B., Keri, D., Ken, D., Michael, D., William, F., et al.: Initial sequencing and analysis of the human genome. (2001)
88. Drmanac, R., Sparks, A.B., Callow, M.J., Halpern, A.L., Burns, N.L., Kermani, B.G., Carnevali, P., Nazarenko, I., Nilsen, G.B., Yeung, G., et al.: Human genome sequencing using unchained base reads on self-assembling dna nanoarrays. Science **327**(5961), 78–81 (2010)
89. Chen, H., Chiang, R.H., Storey, V.C.: Business intelligence and analytics: from big data to big impact. MIS Q. **36**(4), 1165–1188 (2012)
90. Priyanka, K., Kulennavar, N.: A survey on big data analytics in health care. Int. J. Comput. Sci. Inf. Technol. **5**(4), 5865–5868 (2014)
91. Koboldt, D.C., Steinberg, K.M., Larson, D.E., Wilson, R.K., Mardis, E.R.: The next-generation sequencing revolution and its impact on genomics. Cell **155**(1), 27–38 (2013)
92. Kanz, C., Aldebert, P., Althorpe, N., Baker, W., Baldwin, A., Bates, K., Browne, P., van den Broek, A., Castro, M., Cochrane, G., et al.: The embl nucleotide sequence database. Nucl. Acids Res. **33**(suppl_1), D29–D33 (2005)
93. Bilofsky, H.S., Christian, B.: The genbank® genetic sequence data bank. Nucl. Acids Res. **16**(5), 1861–1863 (1988)
94. Yao, Y.G., Salas, A., Logan, I., Bandelt, H.J.: mtdna data mining in genbank needs surveying. Am. J. Hum. Genet. **85**(6), 929–933 (2009)
95. Sugawara, H., Ogasawara, O., Okubo, K., Gojobori, T., Tateno, Y.: DDBJ with new system and face. Nucl. Acids Res. **36**(suppl_1), D22–D24 (2007)
96. Letovsky, S.I., Cottingham, R.W., Porter, C.J., Li, P.W.: Gdb: the human genome database. Nucl. Acids Res. **26**(1), 94–99 (1998)
97. Boeckmann, B., Blatter, M.C., Famiglietti, L., Hinz, U., Lane, L., Roechert, B., Bairoch, A.: Protein variety and functional diversity: Swiss-prot annotation in its biological context. Compt. rendus Biol. **328**(10–11), 882–899 (2005)
98. Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'donovan, C., Phan, I., et al.: The swiss-prot protein knowledgebase and its supplement trembl in 2003. Nucl. Acids Res. **31**(1), 365–370 (2003)
99. GDB: http://www.bioinfo.pte.hu/more/TrEMBL.htm. Accessed 15 Mar 2018
100. Hulo, N., Bairoch, A., Bulliard, V., Cerutti, L., De Castro, E., Langendijk-Genevaux, P.S., Pagni, M., Sigrist, C.J.: The prosite database. Nucl. Acids Res. **34**(suppl_1), D227–D230 (2006)
101. Kouranov, A., Xie, L., de la Cruz, J., Chen, L., Westbrook, J., Bourne, P.E., Berman, H.M.: The rcsb pdb information portal for structural genomics. Nucl. Acids Res. **34**(suppl_1), D302–D305 (2006)
102. Lee, T.J., Pouliot, Y., Wagner, V., Gupta, P., Stringer-Calvert, D.W., Tenenbaum, J.D., Karp, P.D.: Biowarehouse: a bioinformatics database warehouse toolkit. BMC Bioinform. **7**(1), 170 (2006)
103. Bagyamathi, M., Inbarani, H.H.: A novel hybridized rough set and improved harmony search based feature selection for protein sequence classification. In: Big Data in Complex Systems, pp. 173–204. Springer (2015)
104. Barbu, A., She, Y., Ding, L., Gramajo, G.: Feature selection with annealing for big data learning. arXiv preprint (2013)
105. Zeng, A., Li, T., Liu, D., Zhang, J., Chen, H.: A fuzzy rough set approach for incremental feature selection on hybrid information systems. Fuzzy Sets Syst. **258**, 39–60 (2015)
106. Mitchell, T.M., et al.: Machine learning. 1997. Burr Ridge, IL: McGraw Hill **45**(37), 870–877 (1997)
107. Duda, R.O., Hart, P.E., Stork, D.G., et al.: Pattern Classification, vol. 2. Wiley, New York (1973)
108. Anzai, Y.: Pattern Recognition and Machine Learning. Elsevier, Amsterdam (2012)
109. Mohri, M., Rostamizadeh, A., Talwalkar, A.: Foundations of Machine Learning. MIT Press, Cambridge (2012)
110. Hsieh, C.J., Si, S., Dhillon, I.: A divide-and-conquer solver for kernel support vector machines. In: International Conference on Machine Learning, pp. 566–574 (2014)
111. Djuric, N.: Big Data Algorithms for Visualization and Supervised Learning. Temple University, Philadelphia (2013)
112. Giveki, D., Salimi, H., Bahmanyar, G., Khademian, Y.: Automatic detection of diabetes diagnosis using feature weighted support vector machines based on mutual information and modified cuckoo search. arXiv preprint arXiv:1201.2173 (2012)
113. Haller, S., Badoud, S., Nguyen, D., Garibotto, V., Lovblad, K., Burkhard, P.: Individual detection of patients with parkinson disease using support vector machine analysis of diffusion tensor imaging data: initial results. Am. J. Neuroradiol. **33**(11), 2123–2128 (2012)
114. Son, Y.J., Kim, H.G., Kim, E.H., Choi, S., Lee, S.K.: Application of support vector machine for prediction of medication adherence in heart failure patients. Healthc. Inf. Res. **16**(4), 253–259 (2010)
115. Bhatia, S., Prakash, P., Pillai, G.: Svm based decision support system for heart disease classification with integer-coded genetic algorithm to select critical features. In: Proceedings of the World Congress on Engineering and Computer Science, pp. 34–38 (2008)
116. Ye, J., Chow, J.H., Chen, J., Zheng, Z.: Stochastic gradient boosted distributed decision trees. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management, pp. 2061–2064. ACM (2009)
117. Calaway, R., Edlefsen, L., Gong, L., Fast, S.: Big data decision trees with r. Revolution (2016)
118. Hall, L.O., Chawla, N., Bowyer, K.W.: Decision tree learning on very large data sets. In: Systems, Man, and Cybernetics, 1998. 1998 IEEE International Conference on, vol. 3, pp. 2579–2584. IEEE (1998)

119. Ng, R.T., Han, J.: Clarans: a method for clustering objects for spatial data mining. IEEE Trans. Knowl. Data Eng. **14**(5), 1003–1016 (2002)
120. Ester, M., Kriegel, H.P., Sander, J., Xu, X., et al.: A density-based algorithm for discovering clusters in large spatial databases with noise. Kdd **96**(34), 226–231 (1996)
121. Hinneburg, A., Keim, D.A., et al.: An efficient approach to clustering in large multimedia databases with noise. KDD **98**, 58–65 (1998)
122. Guha, S., Rastogi, R., Shim, K.: CURE: an efficient clustering algorithm for large databases. ACM Sigmod Rec. **27**(2), 73–84 (1998)
123. Huang, Z.: Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Min. Knowl. Discov. **2**(3), 283–304 (1998)
124. Xu, X., Jäger, J., Kriegel, H.P.: A fast parallel clustering algorithm for large spatial databases. In: High Performance Data Mining, pp. 263–290. Springer (1999)
125. Chen, N., Chen, A.Z., Zhou, L.X.: An incremental grid density-based clustering algorithm. J. Softw. **13**(1), 1–7 (2002)
126. Kumar, V., Sharma, R.M., Thakur, R.: Big data analytics: Bioinformatics perspective. (2016)
127. Stokes, T.H., Moffitt, R.A., Phan, J.H., Wang, M.D.: chip artifact correction (cacorrect): a bioinformatics system for quality assurance of genomics and proteomics array data. Ann. Biomed. Eng. **35**(6), 1068–1080 (2007)
128. Phan, J.H., Young, A.N., Wang, M.D.: omnibiomarker: a web-based application for knowledge-driven biomarker identification. IEEE Trans. Biomed. Eng. **60**(12), 3364–3367 (2013)
129. Liang, M., Zhang, F., Jin, G., Zhu, J.: FastGCN: a GPU accelerated tool for fast gene co-expression networks. PloS one **10**(1), e0116776 (2015)
130. Day, A., Dong, J., Funari, V.A., Harry, B., Strom, S.P., Cohn, D.H., Nelson, S.F.: Disease gene characterization through large-scale co-expression analysis. PLoS one **4**(12), e8491 (2009)
131. Langfelder, P., Horvath, S.: Wgcna: an r package for weighted correlation network analysis. BMC Bioinform. **9**(1), 559 (2008)
132. Rivera, C.G., Vakil, R., Bader, J.S.: Nemo: network module identification in cytoscape. BMC Bioinform. **11**(1), S61 (2010)
133. Bader, G.D., Hogue, C.W.: An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinform. **4**(1), 2 (2003)
134. Nepusz, T., Yu, H., Paccanaro, A.: Detecting overlapping protein complexes in protein-protein interaction networks. Nat. Methods **9**(5), 471 (2012)
135. Kelley, B.P., Yuan, B., Lewitter, F., Sharan, R., Stockwell, B.R., Ideker, T.: Pathblast: a tool for alignment of protein interaction networks. Nucl. Acids Res. **32**(suppl_2), W83–W88 (2004)
136. Zambon, A.C., Gaj, S., Ho, I., Hanspers, K., Vranizan, K., Evelo, C.T., Conklin, B.R., Pico, A.R., Salomonis, N.: Go-elite: a flexible solution for pathway and ontology over-representation. Bioinformatics **28**(16), 2209–2210 (2012)
137. van Iersel, M.P., Kelder, T., Pico, A.R., Hanspers, K., Coort, S., Conklin, B.R., Evelo, C.: Presenting and exploring biological pathways with PathVisio. BMC Bioinformatics **9**(1), 1–9 (2008)
138. Yang, P., Patrick, E., Tan, S.X., Fazakerley, D.J., Burchfield, J., Gribben, C., Prior, M.J., James, D.E., Hwa Yang, Y.: Direction pathway analysis of large-scale proteomics data reveals novel features of the insulin action pathway. Bioinformatics **30**(6), 808–814 (2013)
139. Grosu, P., Townsend, J.P., Hartl, D.L., Cavalieri, D.: Pathway processor: a tool for integrating whole-genome expression results into metabolic networks. Genome Res. **12**(7), 1121–1126 (2002)
140. Park, Y.S., Schmidt, M., Martin, E.R., Pericak-Vance, M.A., Chung, R.H.: Pathway-pdt: a flexible pathway analysis tool for nuclear families. BMC Bioinform. **14**(1), 267 (2013)
141. Luo, W., Brouwer, C.: Pathview: an r/bioconductor package for pathway-based data integration and visualization. Bioinformatics **29**(14), 1830–1831 (2013)
142. Schatz, M.C.: Cloudburst: highly sensitive read mapping with mapreduce. Bioinformatics **25**(11), 1363–1369 (2009)
143. Schatz, M., Sommer, D., Kelley, D., Pop, M.: Contrail: Assembly of large genomes using cloud computing. In: CSHL Biology of Genomes Conference (2010)
144. Gurtowski, J., Schatz, M.C., Langmead, B.: Genotyping in the cloud with crossbow. Curr. Protoc. Bioinform., 3–15 (2012)
145. Lewis, S., Csordas, A., Killcoyne, S., Hermjakob, H., Hoopmann, M.R., Moritz, R.L., Deutsch, E.W., Boyle, J.: Hydra: a scalable proteomic search engine which utilizes the hadoop distributed computing framework. BMC Bioinform. **13**(1), 324 (2012)
146. O'Connor, D.B., Merriman, B., Nelson, S.F.: Seqware query engine: storing and searching sequence data in the cloud. BMC Inform **11**(12), S2 (2010)
147. George, L.: HBase: The Definitive Guide: Random Access to Your Planet-Size Data. O'Reilly Media, Inc., Newton (2011)
148. Robinson, T., Killcoyne, S., Bressler, R., Boyle, J.: SAMQA: error classification and validation of high-throughput sequenced read data. BMC Genom. **12**(1), 1–7 (2011)
149. Huang, W., Li, L., Myers, J.R., Marth, G.T.: Art: a next-generation sequencing read simulator. Bioinformatics **28**(4), 593–594 (2011)
150. Chen, C.C., Chang, Y.J., Chung, W.C., Lee, D.T., Ho, J.M.: Cloudrs: an error correction algorithm of high-throughput sequencing data based on scalable framework. In: Big Data, 2013 IEEE International Conference on, pp. 717–722. IEEE (2013)
151. Gnerre, S., MacCallum, I., Przybylski, D., Ribeiro, F.J., Burton, J.N., Walker, B.J., Sharpe, T., Hall, G., Shea, T.P., Sykes, S., et al.: High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc. Natl. Acad. Sci. **108**(4), 1513–1518 (2011)
152. Angiuoli, S.V., Matalka, M., Gussman, A., Galens, K., Vangala, M., Riley, D.R., Arze, C., White, J.R., White, O., Fricke, W.F.: Clovr: a virtual machine for automated and portable sequence analysis from the desktop using cloud computing. BMC Bioinformatics **12**(1), 356 (2011)
153. Eelmets, M.: Clovr: A virtual machine for automated and portable sequence analysis from the desktop using cloud computing. – **-**(-) (2011)
154. Krampis, K., Booth, T., Chapman, B., Tiwari, B., Bicak, M., Field, D., Nelson, K.E.: Cloud biolinux: pre-configured and on-demand bioinformatics computing for the genomics community. BMC Bioinf. **13**(1), 42 (2012)
155. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al.: The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. Genome Res. **20**(9), 1297–1303 (2010)
156. Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., et al.: From fastq data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. Curr. Protoc. Bioinform. **43**, 10–11 (2013)
157. Huang, H., Tata, S., Prill, R.J.: Bluesnp: R package for highly scalable genome-wide association studies using hadoop clusters. Bioinformatics **29**(1), 135–136 (2012)
158. Abbott, P.A., Coenen, A.: Globalization and advances in information and communication technologies: the impact on nursing and health. Nurs Outlook **56**(5), 238–246 (2008)

159. Bhattacherjee, A., Hikmet, N.: Physicians' resistance toward healthcare information technology: a theoretical model and empirical test. Eur. J. Inf. Syst. **16**(6), 725–737 (2007)
160. Blumenthal, D.: Launching hitech. N. Engl. J. Med. **362**(5), 382–385 (2010)
161. Bakshi, K.: Considerations for big data: architecture and approach. In: Aerospace Conference, 2012 IEEE, pp. 1–7. IEEE (2012)
162. Herodotou, H., Lim, H., Luo, G., Borisov, N., Dong, L., Cetin, F.B., Babu, S.: Starfish: a self-tuning system for big data analytics. Cidr **11**, 261–272 (2011)
163. Buntin, M.B., Burke, M.F., Hoaglin, M.C., Blumenthal, D.: The benefits of health information technology: a review of the recent literature shows predominantly positive results. Health Affairs **30**, 464–471 (2011)
164. Dutta, H., Kamil, A., Pooleery, M., Sethumadhavan, S., Demme, J.: Distributed storage of large-scale multidimensional electroencephalogram data using hadoop and hbase. In: Grid and Cloud Database Management, pp. 331–347. Springer (2011)
165. Jin, Y., Deyu, T., Yi, Z.: A distributed storage model for ehr based on hbase. In: Information Management, Innovation Management and Industrial Engineering (ICIII), 2011 International Conference on, vol. 2, pp. 369–372. IEEE (2011)
166. Nguyen, A.V., Wynden, R., Sun, Y.: Hbase, mapreduce, and integrated data visualization for processing clinical signal data. In: AAAI Spring Symposium: Computational Physiology, vol. 2011. California, CA: Association for the Advancement of Artificial Intelligence (2011)
167. Jayapandian, C.P., Chen, C.H., Bozorgi, A., Lhatoo, S.D., Zhang, G.Q., Sahoo, S.S.: Cloudwave: distributed processing of "big data" from electrophysiological recordings for epilepsy clinical research using hadoop. In: AMIA Annual Symposium Proceedings, vol. 2013, p. 691. American Medical Informatics Association (2013)
168. Sahoo, S.S., Jayapandian, C., Garg, G., Kaffashi, F., Chung, S., Bozorgi, A., Chen, C.H., Loparo, K., Lhatoo, S.D., Zhang, G.Q.: Heart beats in the cloud: distributed analysis of electrophysiological 'big data'using cloud computing for epilepsy clinical research. J. Am. Med. Inf. Assoc. **21**(2), 263–271 (2013)
169. Mazurek, M.: Applying nosql databases for operationalizing clinical data mining models. In: International Conference: Beyond Databases, Architectures and Structures, pp. 527–536. Springer (2014)
170. Bahga, A., Madisetti, V.K.: A cloud-based approach for interoperable electronic health records (ehrs). IEEE J. Biomed. Health Inf. **17**(5), 894–906 (2013)
171. Chen, J., Qian, F., Yan, W., Shen, B.: Translational biomedical informatics in the cloud: present and future. BioMed Res. Int., (2013)
172. Sharp, J.: An application architecture to facilitate multi-site clinical trial collaboration in the cloud. In: Proceedings of the 2nd International Workshop on Software Engineering for Cloud Computing, pp. 64–68. ACM (2011)
173. Ng, K., Ghoting, A., Steinhubl, S.R., Stewart, W.F., Malin, B., Sun, J.: Paramo: a parallel predictive modeling platform for healthcare analytic research using electronic health records. J. Biomed. Inf. **48**, 160–170 (2014)
174. Chawla, N.V., Davis, D.A.: Bringing big data to personalized healthcare: a patient-centered framework. J. Gener. Intern. Med. **28**(3), 660–665 (2013)
175. Abbott, R.: Big data and pharmacovigilance: using health information exchanges to revolutionize drug safety. Iowa L. Rev. **99**, 225 (2013)
176. Zolfaghar, K., Meadem, N., Teredesai, A., Roy, S.B., Chin, S.C., Muckian, B.: Big data solutions for predicting risk-of-readmission for congestive heart failure patients. In: Big Data, 2013 IEEE International Conference on, pp. 64–71. IEEE (2013)
177. Rangarajan, S., Liu, H., Wang, H., Wang, C.L.: Scalable architecture for personalized healthcare service recommendation using big data lake. In: Service Research and Innovation, pp. 65–79. Springer (2015)
178. Wang, Y., Hajli, N.: Exploring the path to big data analytics success in healthcare. J. Bus. Res. **70**, 287–299 (2017)
179. Saeed, M., Villarroel, M., Reisner, A.T., Clifford, G., Lehman, L.W., Moody, G., Heldt, T., Kyaw, T.H., Moody, B., Mark, R.G.: Multiparameter intelligent monitoring in intensive care ii (mimic-ii): a public-access intensive care unit database. Crit. Care Med. **39**(5), 952 (2011)
180. Hankey, B.F., Ries, L.A., Edwards, B.K.: The surveillance, epidemiology, and end results program: a national resource. Cancer Epidemiol. Prev. Biomark. **8**(12), 1117–1121 (1999)
181. Hiatt, R.A., Rimer, B.K.: A new strategy for cancer control research. Cancer Epidemiol. Prev. Biomark. **8**(11), 957–964 (1999)
182. Zubieta, J.C., Skinner, R., Dean, A.G.: Initiating informatics and gis support for a field investigation of bioterrorism: The new jersey anthrax experience. Int. J. Health. Geogr. **2**(1), 8 (2003)
183. Wan, T.T.: Healthcare informatics research: from data to evidence-based management. J. Med. Syst. **30**(1), 3–7 (2006)
184. Revere, D., Turner, A.M., Madhavan, A., Rambo, N., Bugni, P.F., Kimball, A., Fuller, S.S.: Understanding the information needs of public health practitioners: a literature review to inform design of an interactive digital knowledge management system. J. Biomed. Inf. **40**(4), 410–421 (2007)
185. Herland, M., Khoshgoftaar, T.M., Wald, R.: A review of data mining using big data in health informatics. J. Big Data **1**(1), 2 (2014)
186. Kamesh, D., Neelima, V., Priya, R.R.: A review of data mining using bigdata in health informatics. Int. J. Sci. Res, Publ. **5**(3), (2015)
187. Ravı, D., Wong, C., Deligianni, F., Berthelot, M., Andreu-Perez, J., Lo, B., Yang, G.Z.: Deep learning for health informatics. IEEE J. Biomed. Health Inf. **21**(1), 4–21 (2017)
188. Miotto, R., Wang, F., Wang, S., Jiang, X., Dudley, J.T.: Deep learning for healthcare: review, opportunities and challenges. Brief. Bioinform. (2017)
189. Aziz, H.A.: A review of the role of public health informatics in healthcare. J. Taibah Univ. Med. Sci. **12**(1), 78–81 (2017)
190. Association, T.O.H.: https://www.ericsson.com/491b06/assets/local/mobility-report/documents/2019/ericsson-mobility-report-q4-2019-update.pdf0 (2018). Accessed 29 Mar 2018
191. National Center for Chronic Disease Prevention and Health Promotion, Division of Population Health. https://www.cdc.gov/brfss/about/index.htm (2014). Accessed 29 Mar 2018
192. Hayat, M.J., Howlader, N., Reichman, M.E., Edwards, B.K.: Cancer statistics, trends, and multiple primary cancer analyses from the surveillance, epidemiology, and end results (seer) program. The oncologist **12**(1), 20–37 (2007)
193. (NIH), N.C.I.: https://www.ericsson.com/491b06/assets/local/mobility-report/documents/2019/ericsson-mobility-report-q4-2019-update.pdf2 (2018). Accessed 29 Mar 2018
194. Smith, C.A., Wicks, P.J.: Patientslikeme: Consumer health vocabulary as a folksonomy. In: AMIA annual symposium proceedings, vol. 2008, p. 682. American Medical Informatics Association (2008)
195. Heywood, J.: https://www.ericsson.com/491b06/assets/local/mobility-report/documents/2019/ericsson-mobility-report-q4-2019-update.pdf3 (2005-2018). 18 Apr 2018

196. Wilmoth, J.R., Shkolnikov, V.: Human Mortality Database. University of California, Berkeley (2010)
197. Shkolnikov, V., Barbieri, M., Wilmoth, J.: https://www.ericsson.com/491b06/assets/local/mobility-report/documents/2019/ericsson-mobility-report-q4-2019-update.pdf4. Accessed 18 Apr 2018
198. Young, S.D., Rivers, C., Lewis, B.: Methods of using real-time social media technologies for detection and remote monitoring of hiv outcomes. Prev. Med. **63**, 112–115 (2014)
199. Hay, S.I., George, D.B., Moyes, C.L., Brownstein, J.S.: Big data opportunities for global infectious disease surveillance. PLoS Med. **10**(4), e1001413 (2013)
200. Nambisan, P., Luo, Z., Kapoor, A., Patrick, T.B., Cisler, R.A.: Social media, big data, and public health informatics: Ruminating behavior of depression revealed through twitter. In: 2015 48th Hawaii International Conference on System Sciences, pp. 2906–2913. IEEE (2015)
201. Tsugawa, S., Mogi, Y., Kikuchi, Y., Kishino, F., Fujita, K., Itoh, Y., Ohsaki, H.: On estimating depressive tendencies of twitter users utilizing their tweet data. In: Virtual Reality (VR), 2013 IEEE, pp. 1–4. IEEE (2013)
202. Park, M., Cha, C., Cha, M.: Depressive moods of users portrayed in twitter. In: Proceedings of the ACM SIGKDD Workshop on Healthcare Informatics (HI-KDD), vol. 2012, pp. 1–8. ACM New York, NY (2012)
203. Park, S., Lee, S.W., Kwak, J., Cha, M., Jeong, B.: Activities on facebook reveal the depressive state of users. J. Med. Internet Res. **15**(10), (2013)
204. De Choudhury, M., Gamon, M., Counts, S., Horvitz, E.: Predicting depression via social media. ICWSM **13**, 1–10 (2013)
205. De Choudhury, M., Counts, S., Horvitz, E.: Social media as a measurement tool of depression in populations. In: Proceedings of the 5th Annual ACM Web Science Conference, pp. 47–56. ACM (2013)
206. Zoubovsky, S. P., Hoseus, S., Tumukuntala, S., Schulkin, J. O., Williams, M. T., Vorhees, C. V., et al. (2020). Chronicpsychosocial stress during pregnancy affects maternal behavior and neuroendocrine function and modulateshypothalamic CRH and nuclear steroid receptor expression. *Translational psychiatry*, *10*(1), 1–13.
207. De Choudhury, M., Counts, S., Horvitz, E.J., Hoff, A.: Characterizing and predicting postpartum depression from shared facebook data. In: Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing, pp. 626–638. ACM (2014)
208. Sadilek, A., Kautz, H., Silenzio, V.: Modeling spread of disease from social interactions. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 6, no. 1, pp. 1–8 (2012)
209. Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., Brilliant, L.: Detecting influenza epidemics using search engine query data. Nature **457**(7232), 1012 (2009)
210. Hagg, E., Dahinten, V.S., Currie, L.M.: The emerging use of social media for health-related purposes in low and middle-income countries: a scoping review. Int. J. Med. Inf. **115**, 92–105 (2018)
211. Belle, A., Thiagarajan, R., Soroushmehr, S., Navidi, F., Beard, D.A., Najarian, K.: Big data analytics in healthcare. BioMed Res. Int. (2015)
212. Bodo, M., Settle, T., Royal, J., Lombardini, E., Sawyer, E., Rothwell, S.W.: Multimodal noninvasive monitoring of soft tissue wound healing. J. Clin. Monit. Comput. **27**(6), 677–688 (2013)
213. Hu, P., Galvagno, S.M., Sen, A., Dutton, R., Jordan, S., Floccare, D., Handley, C., Shackelford, S., Pasley, J., Mackenzie, C.: Identification of dynamic prehospital changes with continuous vital signs acquisition. Air Med. J. **33**(1), 27–33 (2014)
214. Drew, B.J., Harris, P., Zègre-Hemsey, J.K., Mammone, T., Schindler, D., Salas-Boni, R., Bai, Y., Tinoco, A., Ding, Q., Hu, X.: Insights into the problem of alarm fatigue with physiologic monitor devices: a comprehensive observational study of consecutive intensive care unit patients. PLoS One **9**(10), e110274 (2014)
215. Graham, K.C., Cvach, M.: Monitor alarm fatigue: standardizing use of physiological monitors and decreasing nuisance alarms. Am. J. Crit. Care **19**(1), 28–34 (2010)
216. McCullough, J.S., Casey, M., Moscovice, I., Prasad, S.: The effect of health information technology on quality in us hospitals. Health Affairs **29**(4), 647–654 (2010)
217. Ahmad, S., Ramsay, T., Huebsch, L., Flanagan, S., McDiarmid, S., Batkin, I., McIntyre, L., Sundaresan, S.R., Maziak, D.E., Shamji, F.M., et al.: Continuous multi-parameter heart rate variability analysis heralds onset of sepsis in adults. PLoS One **4**(8), e6642 (2009)
218. Adrián, G., Francisco, G.E., Marcela, M., Baum, A., Daniel, L., de Quirós Fernán, G.B.: Mongodb: an open source alternative for hl7-cda clinical documents management. In: Proceedings of the Open Source International Conference (CISL'13) (2013)
219. Kaur, K., Rani, R.: Managing data in healthcare information systems: many models, one solution. Computer **48**(3), 52–59 (2015)
220. Santos, M., Portela, F.: Enabling ubiquitous data mining in intensive care: features selection and data pre-processing. In: International Conference on Enterprise Information Systems, vol. 2, pp. 261–266. SCITEPRESS (2011)
221. Berndt, D.J., Fisher, J.W., Hevner, A.R., Studnicki, J.: Healthcare data warehousing and quality assurance. Computer **34**(12), 56–65 (2001)
222. Johnson, A.E., Pollard, T.J., Shen, L., Li-wei, H.L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L.A., Mark, R.G.: Mimic-iii, a freely accessible critical care database. Sci. Data **3**, 160035 (2016)
223. Goldberger, A.L., Amaral, L.A., Glass, L., Hausdorff, J.M., Ivanov, P.C., Mark, R.G., Mietus, J.E., Moody, G.B., Peng, C.-K., Stanley, H.E.: PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation **101**(23), e215–e220 (2000)
224. Han, H., Ryoo, H.C., Patrick, H.: An infrastructure of stream data mining, fusion and management for monitored patients. In: Computer-Based Medical Systems, 2006. CBMS 2006. 19th IEEE International Symposium on, pp. 461–468. IEEE (2006)
225. Bressan, N., James, A., McGregor, C.: Trends and opportunities for integrated real time neonatal clinical decision support. In: Biomedical and Health Informatics (BHI), 2012 IEEE-EMBS International Conference on, pp. 687–690. IEEE (2012)
226. Lee, J., Mark, R.: A hypotensive episode predictor for intensive care based on heart rate and blood pressure time series. In: Computing in Cardiology, 2010, pp. 81–84. IEEE (2010)
227. Sun, J., Sow, D., Hu, J., Ebadollahi, S.: A system for mining temporal physiological data streams for advanced prognostic decision support. In: Data Mining (ICDM), 2010 IEEE 10th International Conference on, pp. 1061–1066. IEEE (2010)
228. Cao, H., Eshelman, L., Chbat, N., Nielsen, L., Gross, B., Saeed, M.: Predicting icu hemodynamic instability using continuous multiparameter trends. In: Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE, pp. 3803–3806. IEEE (2008)
229. Le Roux, P., Menon, D.K., Citerio, G., Vespa, P., Bader, M.K., Brophy, G.M., Diringer, M.N., Stocchetti, N., Videtta, W., Armonda, R., et al.: Consensus summary statement of the international multidisciplinary consensus conference on multimodality monitoring in neurocritical care. Neurocrit. Care **21**(2), 1–26 (2014)

230. Rajan, J.P., Rajan, S.E.: An internet of things based physiological signal monitoring and receiving system for virtual enhanced health care network. Technol. Health Care **26**(2), 1–7 (2018)
231. Zhang, Z., Zhang, Y., Yao, L., Song, H., Kos, A.: A sensor-based wrist pulse signal processing and lung cancer recognition. J. Biomed. Inf **79**, 107–116 (2018)
232. Nanda, S.K., Lin, W.Y., Lee, M.Y., Chen, R.S.: A quantitative classification of essential and parkinson's tremor using wavelet transform and artificial neural network on semg and accelerometer signals. In: Networking, Sensing and Control (ICNSC), 2015 IEEE 12th International Conference on, pp. 399–404. IEEE (2015)
233. Rouse, W.B., Serban, N.: Understanding and Managing the Complexity of Healthcare. MIT Press, Cambridge (2014)
234. Mohammed, E.A., Far, B.H., Naugler, C.: Applications of the mapreduce programming framework to clinical big data analysis: current landscape and future trends. BioData Min. **7**(1), 22 (2014)
235. Swan, M.: The quantified self: Fundamental disruption in big data science and biological discovery. Big Data **1**(2), 85–99 (2013)
236. Huang, B.E., Mulyasasmita, W., Rajagopal, G.: The path from big data to precision medicine. Expert Rev. Precis. Med. Drug Dev. **1**(2), 129–143 (2016)
237. Bradley, P.S.: Implications of big data analytics on population health management. Big Data **1**(3), 152–159 (2013)
238. Wang, W., Haerian, K., Salmasian, H., Harpaz, R., Chase, H., Friedman, C.: A drug-adverse event extraction algorithm to support pharmacovigilance knowledge mining from pubmed citations. In: AMIA annual symposium proceedings, vol. 2011, p. 1464. American Medical Informatics Association (2011)
239. Hung, C.L., Lin, Y.L.: Implementation of a parallel protein structure alignment service on cloud. Int. J. Genom., (2013)
240. Wang, L., Chen, D., Ranjan, R., Khan, S.U., KolOdziej, J., Wang, J.: Parallel processing of massive eeg data with mapreduce. In: 2012 IEEE 18th International Conference on Parallel and Distributed Systems, pp. 164–171. Ieee (2012)
241. Meng, B., Pratx, G., Xing, L.: Ultrafast and scalable cone-beam ct reconstruction using mapreduce in a cloud computing environment. Med. Phys. **38**(12), 6603–6609 (2011)
242. Peek, N., Holmes, J., Sun, J.: Technical challenges for big data in biomedicine and health: data sources, infrastructure, and analytics. Yearb. Med. Inf. **23**(01), 42–47 (2014)
243. Maia, A.T., Sammut, S.J., Jacinta-Fernandes, A., Chin, S.F.: Big data in cancer genomics. Curr. Opin. Syst. Biol. **4**, 78–84 (2017)
244. Wong, H.T., Yin, Q., Guo, Y.Q., Murray, K., Zhou, D.H., Slade, D.: Big data as a new approach in emergency medicine research. J. Acute Dis. **4**(3), 178–179 (2015)
245. Viceconti, M., Hunter, P., Hose, R.: Big data, big knowledge: big data for personalized healthcare. IEEE J. Biomed. Health Inf. **19**(4), 1209–1215 (2015)
246. Geerts, H., Dacks, P.A., Devanarayan, V., Haas, M., Khachaturian, Z.S., Gordon, M.F., Maudsley, S., Romero, K., Stephenson, D., Initiative, B.H.M., et al.: Big data to smart data in Alzheimer's disease: the brain health modeling initiative to foster actionable knowledge. Alzheimer's Dement. **12**(9), 1014–1021 (2016)
247. El Naqa, I.: Perspectives on making big data analytics work for oncology. Methods **111**, 32–44 (2016)
248. Lu, J., Xu, Q., Li, B., Yuan, X., Sato, K.: Image processing apparatus, image processing method and medical imaging device (2019). US Patent App. 10/282,631
249. Karmonik, C., Boone, T.B., Khavari, R.: Workflow for visualization of neuroimaging data with an augmented reality device. J. Dig. Imaging **31**(1), 26–31 (2018)
250. Glemser, P.A., Engel, K., Simons, D., Steffens, J., Schlemmer, H.P., Orakcioglu, B.: A new approach for photorealistic visualization of rendered computed tomography images. World Neurosurg. **114**, e283–e292 (2018)
251. Yu, D., Engel, K.: Joint visualization of 3d reconstructed photograph and internal medical scan (2018). US Patent App. 10/092,191
252. Jorge, J.A., Simões Lopes, D.: Challenges and approaches to interactive visualization in healthcare workspaces. Ann. Med. **51**(sup1), 22–22 (2019)
253. Liu, R.W., Ma, Q., Yu, S.C.H., Chui, K.T., Xiong, N.: Variational regularized tree-structured wavelet sparsity for cs-sense parallel imaging. IEEE Access **6**, 61050–61064 (2018)
254. Khan, S., Islam, N., Jan, Z., Din, I.U., Rodrigues, J.J.C.: A novel deep learning based framework for the detection and classification of breast cancer using transfer learning. Pattern Recognit. Lett. **125**, 1–6 (2019)
255. Lakshmanaprabu, S., Mohanty, S.N., Shankar, K., Arunkumar, N., Ramirez, G.: Optimal deep learning model for classification of lung cancer on ct images. Future Gener. Comput. Syst. **92**, 374–382 (2019)
256. Razzak, I., Naz, S., Rehman, A., Khan, A., Zaib, A.: Improving coronavirus (covid-19) diagnosis using deep transfer learning. medRxiv (2020)
257. Grace, R.K., Manimegalai, R., Kumar, S.S.: Medical image retrieval system in grid using hadoop framework. In: 2014 International Conference on Computational Science and Computational Intelligence, vol. 1, pp. 144–148. IEEE (2014)
258. Yang, C.T., Shih, W.C., Chen, L.T., Kuo, C.T., Jiang, F.C., Leu, F.Y.: Accessing medical image file with co-allocation hdfs in cloud. Future Gener. Comput. Syst. **43**, 61–73 (2015)
259. Markonis, D., Schaer, R., Eggel, I., Müller, H., Depeursinge, A.: Using mapreduce for large-scale medical image analysis. In: 2012 IEEE Second International Conference on Healthcare Informatics, Imaging and Systems Biology, pp. 1–1. IEEE (2012)
260. Benjamin, M., Aradi, Y., Shreiber, R.: From shared data to sharing workflow: merging pacs and teleradiology. Eur. J. Radiol. **73**(1), 3–9 (2010)
261. Costa, C., Oliveira, J.L.: Telecardiology through ubiquitous internet services. Int. J. Med. Inf. **81**(9), 612–621 (2012)
262. Ross, P., Pohjonen, H.: Images crossing borders: image and workflow sharing on multiple levels. Insights Imaging **2**(2), 141–148 (2011)
263. Wang, F., Lee, R., Liu, Q., Aji, A., Zhang, X., Saltz, J.: Hadoop-gis: A high performance query system for analytical medical imaging with mapreduce: Technical report. Emory University (2011)
264. Zou, Q., Zeng, J., Cao, L., Ji, R.: A novel features ranking metric with application to scalable visual and bioinformatics data classification. Neurocomputing **173**, 346–354 (2016)
265. Tadist, K., Najah, S., Nikolov, N.S., Mrabti, F., Zahi, A.: Feature selection methods and genomic big data: a systematic review. J. Big Data **6**(1), 79 (2019)
266. Lualdi, M., Fasano, M.: Statistical analysis of proteomics data: a review on feature selection. J. Proteom. **198**, 18–26 (2019)
267. David, S.K., Saeb, A.T., Rafiullah, M., Rubeaan, K.: Classification techniques and data mining tools used in medical bioinformatics. In: Big Data Governance and Perspectives in Knowledge Management, pp. 105–126. IGI Global (2019)
268. Devi, A.S., Maragatham, G.: Big genome data classification with random forests using variantspark. In: International Conference on Computer Networks and Communication Technologies, pp. 599–614. Springer (2019)
269. Patel, D.T.: Big data analytics in bioinformatics. In: Biotechnology: Concepts, Methodologies, Tools, and Applications, pp. 1967–1984. IGI Global (2019)

270. Goli-Malekabadi, Z., Sargolzaei-Javan, M., Akbari, M.K.: An effective model for store and retrieve big health data in cloud computing. Comput. Methods Programs Biomed. **132**, 75–82 (2016)
271. Sultana, S.N., Ramu, G., Reddy, B.E.: Cloud-based development of smart and connected data in healthcare application. Int. J. Distrib. Parallel Syst. **5**(6), 1 (2014)
272. He, C., Fan, X., Li, Y.: Toward ubiquitous healthcare services with a novel efficient cloud platform. IEEE Trans. Biomed. Eng. **60**(1), 230–234 (2012)
273. Wang, Y., Wang, L., Liu, H., Lei, C.: Large-scale clinical data management and analysis system based on cloud computing. In: Frontier and Future Development of Information Technology in Medicine and Education, pp. 1575–1583. Springer (2014)
274. Chen, J., Li, K., Rong, H., Bilal, K., Yang, N., Li, K.: A disease diagnosis and treatment recommendation system based on big data mining and cloud computing. Inf. Sci. **435**, 124–149 (2018)
275. Wang, Y., Kung, L., Byrd, T.A.: Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. Technol. Forecast. Soc. Change **126**, 3–13 (2018)
276. Gupta, M., George, J.F.: Toward the development of a big data analytics capability. Inf. Manag. **53**(8), 1049–1064 (2016)
277. Wang, Y., Byrd, T.A.: Business analytics-enabled decision-making effectiveness through knowledge absorptive capacity in health care. J. Knowl. Manag. **21**(3), 517–539 (2017)
278. Kim, M.K., Park, J.H.: Identifying and prioritizing critical factors for promoting the implementation and usage of big data in healthcare. Inf. Dev. **33**(3), 257–269 (2017)
279. Chui, K.T., Alhalabi, W., Pang, S.S.H., Pablos, P.O.D., Liu, R.W., Zhao, M.: Disease diagnosis in smart healthcare: innovation, technologies and applications. Sustainability **9**(12), 2309 (2017)
280. Schultz, T.: Turning healthcare challenges into big data opportunities: a use-case review across the pharmaceutical development lifecycle. Bull. Am. Soc. Inf. Sci. Technol. **39**(5), 34–40 (2013)
281. Sobhy, D., El-Sonbaty, Y., Elnasr, M.A.: Medcloud: healthcare cloud computing system. In: 2012 International Conference for Internet Technology and Secured Transactions, pp. 161–166. IEEE (2012)
282. Lin, W., Dou, W., Zhou, Z., Liu, C.: A cloud-based framework for home-diagnosis service over big medical data. J. Syst. Softw. **102**, 192–206 (2015)
283. Seth, B., Dalal, S., Kumar, R.: Securing bioinformatics cloud for big data: Budding buzzword or a glance of the future. In: Recent Advances in Computational Intelligence, pp. 121–147. Springer (2019)
284. Garattini, C., Raffle, J., Aisyah, D.N., Sartain, F., Kozlakidis, Z.: Big data analytics, infectious diseases and associated ethical impacts. Philos. Technol. **32**(1), 69–85 (2019)
285. Lamarche-Vadel, A., Pavillon, G., Aouba, A., Johansson, L.A., Meyer, L., Jougla, E., Rey, G.: Automated comparison of last hospital main diagnosis and underlying cause of death icd10 codes, France, 2008–2009. BMC Med. Inf. Decis. Mak. **14**(1), 44 (2014)
286. Cunha, J., Silva, C., Antunes, M.: Health twitter big bata management with hadoop framework. Proc. Comput. Sci. **64**, 425–431 (2015)
287. Gamache, R., Kharrazi, H., Weiner, J.P.: Public and population health informatics: the bridging of big data to benefit communities. Yearb. Med. Inf. **27**(01), 199–206 (2018)
288. Van Schaik, P., Peng, Y., Ojelabi, A., Ling, J.: Explainable statistical learning in public health for policy development: the case of real-world suicide data. BMC Med. Res. Methodol. **19**(1), 152 (2019)
289. Hatef, E., Weiner, J.P., Kharrazi, H.: A public health perspective on using electronic health records to address social determinants of health: the potential for a national system of local community health records in the United States. Int. J. Med. Inf. **124**, 86–89 (2019)
290. Seabrook, E.M., Kern, M.L., Rickard, N.S.: Social networking sites, depression, and anxiety: a systematic review. JMIR Ment. Health **3**(4), e50 (2016)
291. Conway, M., O'Connor, D.: Social media, big data, and mental health: current advances and ethical implications. Curr. Opin. Psychol. **9**, 77–82 (2016)
292. Mohr, D.C., Burns, M.N., Schueller, S.M., Clarke, G., Klinkman, M.: Behavioral intervention technologies: evidence review and recommendations for future research in mental health. Gener. Hosp. Psychiatry **35**(4), 332–338 (2013)
293. Bhardwaj, N., Wodajo, B., Spano, A., Neal, S., Coustasse, A.: The impact of big data on chronic disease management. Health Care Manag **37**(1), 90–98 (2018)
294. Tu, J.V., Chu, A., Donovan, L.R., Ko, D.T., Booth, G.L., Tu, K., Maclagan, L.C., Guo, H., Austin, P.C., Hogg, W., et al.: The cardiovascular health in ambulatory care research team (canheart) using big data to measure and improve cardiovascular health and healthcare services. Circ. Cardiovasc. Qual. Outcomes **8**(2), 204–212 (2015)
295. Kupersmith, J., Francis, J., Kerr, E., Krein, S., Pogach, L., Kolodner, R.M., Perlin, J.B.: Advancing evidence-based care for diabetes: Lessons from the veterans health administration: A highly regarded ehr system is but one contributor to the quality transformation of the vha since the mid-1990s. Health Affairs **26**(Suppl1), w156–w168 (2007)
296. Consortium, I.H.G.S., et al.: Initial sequencing and analysis of the human genome. Nature **409**(6822), 860 (2001)
297. Energy, U.: Insights learned from the human dna sequence, what has been learned from analysis of the working draft sequence of the human genome? what is still unknown? Online. http://www. ornl. gov/hgmis, Accessed 2 May 2011
298. Hey, A.J., Trefethen, A.E.: The data deluge: an e-science perspective. (2003)
299. Ritter, F., Boskamp, T., Homeyer, A., Laue, H., Schwier, M., Link, F., Peitgen, H.O.: Medical image analysis. IEEE Pulse **2**(6), 60–70 (2011)
300. D'Agostino Sr, R.B., Grundy, S., Sullivan, L.M., Wilson, P., Group, C.R.P., et al.: Validation of the framingham coronary heart disease prediction scores: results of a multiple ethnic groups investigation. Jama **286**(2), 180–187 (2001)
301. Waqialla, M., Razzak, M.I.: An ontology-based framework aiming to support cardiac rehabilitation program. Proc. Comput. Sci. **96**, 23–32 (2016)
302. Alexander, C., Wang, L.: Big data analytics in heart attack prediction. J. Nurs. Care **6**(393), 1168–2167 (2017)
303. Palaniappan, S., Awang, R.: Intelligent heart disease prediction system using data mining techniques. In: Computer Systems and Applications, 2008. AICCSA 2008. IEEE/ACS International Conference on, pp. 108–115. IEEE (2008)
304. Shamli, N., Sathiyabhama, B.: Parkinson's brain disease prediction using big data analytics (2016)
305. Razzak, I., Kamran, I., Naz, S.: Deep analysis of handwritten notes for early diagnosis of neurological disorders. In: 2020 International Joint Conference on Neural Networks (IJCNN), pp. 1–6. IEEE (2020)
306. Kamran, I., Naz, S., Razzak, I., Imran, M.: Handwriting dynamics assessment using deep neural network for early identification of Parkinson's disease. Future Gener. Comput. Syst. (2020)
307. Sadhana, S.S., Shetty, S.: Analysis of diabetic data set using hive and r. Int. J. Emerg. Technol. Adv. Eng. **4**(7), 626–9 (2014)
308. Daghistani, T., Al Shammari, R., Razzak, M.I.: Discovering diabetes complications: an ontology based model. Acta Inf. Med. **23**(6), 385 (2015)

309. Panda, M., Ali, S.M., Panda, S.K.: Big data in health care: A mobile based solution. In: Big Data Analytics and Computational Intelligence (ICBDAC), 2017 International Conference on, pp. 149–152. IEEE (2017)
310. Helm-Murtagh, S.C.: Use of big data by blue cross and blue shield of North Carolina. North Carol. Med. J. **75**(3), 195–197 (2014)
311. Wu, X., Zhu, X., Wu, G.Q., Ding, W.: Data mining with big data. IEEE Trans. Knowl. Data Eng. **26**(1), 97–107 (2013)