IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 27, 2018, accepted October 14, 2018, date of publication October 26, 2018,
date of current version November 30, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2878254

Health Big Data Analytics: A Technology Survey

GASPARD HARERIMANA, (Student Member, IEEE), BEAKCHEOL JANG, (Member, IEEE), JONG WOOK KIM, (Member, IEEE), AND HUNG KOOK PARK
Department of Computer Science, Sangmyung University, Seoul 03016, South Korea

Corresponding author: Beakcheol Jang (bjang@smu.ac.kr)

This work was supported by the National Research Foundation of Korea funded by the Korea Government under Grant NRF-2016R1D1A1B03930815.

ABSTRACT Because of the vast availability of data, there has been an additional focus on the health industry and an increasing number of studies that aim to leverage the data to improve healthcare have been conducted. The health data are growing increasingly large, more complex, and its sources have increased tremendously to include computerized physician order entry, electronic medical records, clinical notes, medical images, cyber-physical systems, medical Internet of Things, genomic data, and clinical decision support systems. New types of data from sources like social network services and genomic data are used to build personalized healthcare systems, hence health data are obtained in various forms, from varied sources, contexts, technologies, and their nature can impede a proper analysis. Any analytical research must overcome these obstacles to mine data and produce meaningful insights to save lives. In this paper, we investigate the key challenges, data sources, techniques, technologies, as well as future directions in the field of big data analytics in healthcare. We provide a do-it-yourself review that delivers a holistic, simplified, and easily understandable view of various technologies that are used to develop an integrated health analytic application.

INDEX TERMS Big data, cyber-physical systems, health analytics, machine learning, social networks analysis.

I. INTRODUCTION

Key attributes of volume, velocity, and variety [1], [5] make big data a favorite among the latest innovations. It is estimated that the sheer volume of health data is expected to skyrocket to approximately 25,000 Pb by 2020 [2]. Healthcare is a data-intensive field; hence the data cannot be handled by traditional electronic medical records (EMR)-based software. Moreover, health data has become very ubiquitous because of improvements to recording systems in healthcare, the participation of patients in their treatment using social networks [3], as well as the introduction of cyber-physical systems (CPS) in health care [4], [6], [7]. Hence, the field of big data and computational intelligence promises a bright prospect in building state-of-the-art health systems.

As sources of big data are ubiquitous in nature it is hard to predict how the future might look like because sources of data will continue to increase exponentially. Hence, existing hospital informatics systems will not be adequate for performing data analysis. Currently, big data which is being used in the healthcare sector come in prevalent forms and swiftly emerge from divergent platforms. As there is an urgency to take prompt action in case of a medical emergency, the analytics system must be able to aggregate all data and provide insights to the physician in real-time.

A. MOTIVATION FOR THIS SURVEY

Heretofore, health data sources were limited to classical testing equipment and techniques such as Electrocardiogram (ECG) mammography, Magnetic Resonance Imaging (MRI), ultrasound, CT scanners, and many other testing equipment. The physician must perform all the analytical tasks manually, and while legacy software systems were used to perform tasks such as patient logging, billing, transfer, admission, and assets management, diagnostic tasks were usually performed by abductive reasoning. However, with the advent of ubiquitous cyber-physical systems in the industry, the physician now needs to handle a huge amount of data manually. Hence now, more than ever, computational intelligence will have to play an active role. It is now very arduous to fully leverage and effectively aggregate all health data in a unified manner owing to its varied nature and description. Furthermore, the interpretation of health data and the process of deriving inferences from it require additional analytic solutions for a specific dataset. For example, to retrieve meaning from

EMR entries, such as doctors' notes, another solution that can analyze the SNOMED-CT database (which is a nomenclature database) is required. As these two databases are different, a fit-all analytic solution that can handle these datasets (which diverge in terms of data types, speeds, naming standards) cannot be determined.

Without the assistance of an integrated healthcare solution, physicians might use the traditional EMR or they might use it to a varying degree depending on how conversant the user is [8], [10]. To consolidate data from EMR, additional data sources that are in use today need special analytics that should be integrated with EMR analytics. For example, to retrieve a meaningful insight from the patient's social network or that of patients with similar cases, we must integrate Social Networks Analysis (SNA) with EMR-based analytics and other cyber-physical systems.

In recent years, several studies [138], [139] have attempted to integrate some of the sources of health data to develop integrated solutions. Specialized health solutions such as COHESY [11], CARE (Collaborative Assessment Recommendation Engine) [140], which relies only on medical record data, and AEGLE [12], have attempted to integrate health analytics at a certain degree; however, they have not been able to cater to each aspect of healthcare data.

This survey serves as a guide to accommodate all or most of the critical data sources and produce an integrated analytic solution that can save lives and prevent unnecessary spending on health care. Fig.1 shows a blueprint of how

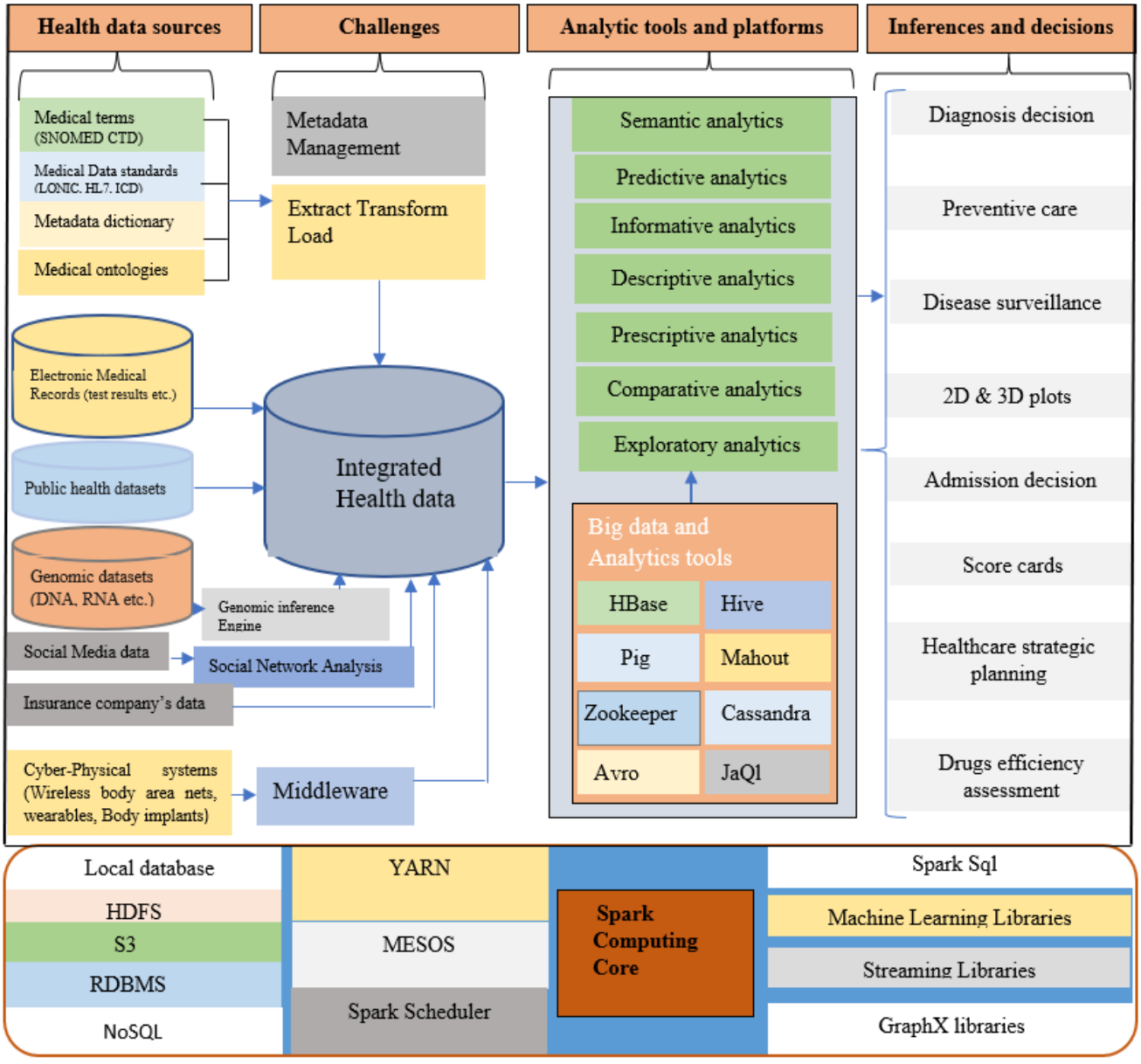

FIGURE 1. Conceptual overview of health big data analytics technologies. Spark is used as an example of a big data platform.

integrated healthcare analytics would appear. Health data comes from both structured data sources such as EMR [9] and insurance providers' databases as well as unstructured forms such as doctors' notes, prescriptions, IoT devices (such as wearables), medical sensors, electronic monitors, mobile applications, social media, and research registries.

While straightforward algorithms such as decision trees or k-means can solve forthright tasks such as deciding if a given patient should be hospitalized or not, certain tasks like those involving the application of MRI scans for differentiating a cancerous tumor from a benign one can require sophisticated algorithms such as Convolutional Neural Networks (CNNs). Hence, the choice of a technique for a given health instance requires a thorough analysis. In this review, we provide options to choose from, depending on the health predicament.

To scale up, the health analytic solution must rely on a big data platform or any other parallel computing hardware. The choice of the platform depends on the data. While some static data such as EMR records can be handled effectively by batch computing platforms such as Hadoop MapReduce, some other real-time data that are critical for the patient's survival such as real-time ECG readings or social network analysis will require stream computing platforms such as Apache Spark or Apache Flink. The platform must also accommodate support libraries such as machine learning libraries and graph libraries for relationship analysis.

B. ORGANIZATION OF THE PAPER

This paper is organized as follows: In section II we discuss the related healthcare analytics surveys. We analyze their key aspects and we cover their limitations as well as the added value of the current study. In section III we discuss key challenges that a healthcare analytics solution must address. In section IV we discuss key sources of healthcare data. In this section, we investigate sources that are usually not considered while building health informatics like social networks data and mIoT data sources. In section V we thoroughly cover the data mining techniques that are used in health analytics. We first provide a brief overview of some algorithms by recalling their general uses and then looking into how they can be tailored for health analytics. We provide an overall summary of disparate analytic techniques as well as selected use cases in health analytics.

In section VI we deeply analyze diverse Big data platforms that can be used as foundations upon which an inclusive analytic solution can be built. As in section V we choose some of the most popular platforms and discuss how they are generally used and their specific uses in health analytics then finally we provide an elaborate summary of trendy platforms.

II. RELATED WORK

The field of health analytics and big data has recently gained a big deal of attention. Table 1 summarizes the key related reviews. On each review, we describe its contribution and the topic covered.

At the end of the table, we provide a hasty comparison of our work with these others that are covered. Among all the papers in Table 1 none that instills a do-it-yourself motive by covering all the key problems from the data sources to insightful visualizations.

III. HEALTH BIG DATA ANALYTICS CHALLENGES

The goals of designing an integrated health analytics span a whole patient's ecosystem. The system should not only be able to help to the provision of a successful and timely care by recommending a practical diagnosis which worked well against similar cases but also is able to predict possible medical aggravations that might occur. This includes using the Complex Events Processing (CEP) for the determination of a disease progression which can help to stop it at earlier stages of development. Kuo *et al.* [16] consider the big data challenges per stages of big data pipeline, considering four distinct stages here each stage has its own challenges to overcome. The following challenges at each stage are considered:

A. DATA AGGREGATION

Health big data comes from various sources and sometimes these sources might be large repositories of data which have to be brought into a common platform for a unified analysis. The aggregation challenge is related to high volume and variety of data that needs to be brought together from divergent data warehouses and real-time data. The solution can be the use of high-speed file transfer technologies. From Fig.1 we can observe that the warehouses that host human genomes are completely different from another that hosts the SNOMED-CT nomenclature data. Some studies have been performed to deal with this aggregation hustle. One example is the use of EasyGenomics (BGI) a technology that is used by the Beijing Genomic Institute to transfer large genomic data. Another solution that is presented to deal with data aggregation challenge is data compression. Cox *et al.* [17] proposed a solution that is based on the Burrows-Wheeler Transform (BWT) to compress many DNA sequences. The BWT is a string compression algorithm that compresses the data by grouping similar characters in a series of strings.

B. DATA MAINTAINANCE/STORAGE

Data storage is a key challenge for health big data. The data is ever growing at an exponential rate hence cannot be managed with traditional database management systems. To solve this problem Non-SQL database systems like MongoDB, Cassandra and Hadoop Distributed File System (HDFS)are proposed but cloud computing [18], [137] is argued to be a powerful solution as it can reduce the initial EMR costs.

C. DATA INTEGRATION AND INTEROPERABILITY

The health big data are hugely heterogeneous. They come from many sources with divergent forms and structure hence making interoperability a big challenge. Even the EMR which is the most structured of all data source can present a challenge, especially when more than one EMRs are involved.

TABLE 1. Related papers compared with the current paper.

Study	year	scope	papers reviewed
[13]	2014	This study is one of the pioneer reviews in the field of big data analytics in healthcare. It covers promise and potentials of big data analytics. It does not provide a deep technical aspect of big data analytics in healthcare. The paper also covers key challenges to be overcome to develop a scalable analytic solution.	32
[14]	2015	The study covers various applications of big data in healthcare. It analyses the benefits of big data for the health industry. It thoroughly covers all sources of data that should be leveraged and aggregated for analysis. It also covers many challenges of big data analytics in healthcare.	37
[15]	2018	The survey discusses big data in the healthcare context. It dives into health big data preprocessing and integration as well as analytical and visualization tools. It highlights applications of big data in medical aspects like clinical support and chronical disease monitoring. Finally, the paper discusses matters of security and privacy.	89
[133]	2018	The paper offers a holistic overview of healthcare analytics by focusing on data mining approaches.	176
[134]	2014	This paper focuses on recent research using Big Data tools and approaches for the analysis of Health Informatics data gathered at multiple levels. It performs a review of data mining approaches in healthcare analytics. The paper goes into technical details on how analytics are performed with respect to human-scale biology, clinical-scale, as well as epidemic-scale.	72
[135]	2013	This paper goes deeper into explaining data mining algorithms that are used to mine health data.it provides a concise analysis of how each data mining technique is applied to various diseases. It also covers approximated accuracy for each algorithm	109
[136]	2016	This paper highlights challenges and opportunities for big data in the medical field as well as processing pipeline. It also covers data mining and analysis tools of machine learning.	196
[137]	2015	This paper provides an overview of the processing of medical images as big data. It also covers various challenges that the medical field faces. It goes through signal analytics like ECG and Pulse Oximetry analysis. It covers the role of genomics for diseases diagnoses.	182
Current Study	2018	The current study provides a holistic review of technologies and methods to design and implement a complete healthcare analytic application. It's a do-it-yourself synthesis of all the steps that are involved. We analyze the journey of health big data from the source, we cover various challenges, we analyze most of data mining algorithms by highlighting how each algorithm is used in general and how it is applied to healthcare. We finally discuss platforms and technologies that a health analytic solution should rely upon.	171

In Fig.2, Kuo *et al.* [19] consider interoperability hustles into 3 types; Functional interoperability, metadata interoperability, and data instance interoperability.

Functional interoperability is sensed when two EMRs exchange data while having different naming standards or interpretation. Metadata interoperability can be observed in a relational database whereby a column name which is the metadata of column content has a different naming. As an example, "GENDER" and "SEX" can be used in two EMRs and have the same logical meaning hence it can result in metadata interoperability challenge. Data instance interoperability is when acronyms and other codes do not have the same meaning. For example, a gender might be "M" and "F" in one EMR and "1" and "0" in another EMR.

HL7 (Health level7) is a set of standards that are used to help different EMR to communicate smoothly. It is composed of a set of rules that developers of Hospital Information Systems (HIS) must follow to achieve standardization. With HL7 also medical equipment can share information. However, for a long time, the HL7 adoption has been so lagging and some studies like in [21] and [22] have found this standard to be flawed. Various studies have considered reusing available standards to better deal with interoperability. Lopez and Blobel [20] proposed a semantic interoperability model which consists of trying to implement an existing HL7 standard as a UML profile then applying the profile to system models. Crichton *et al.* [23] and Mudaly *et al.* [24] have considered interoperability problems in the context of low-income countries.

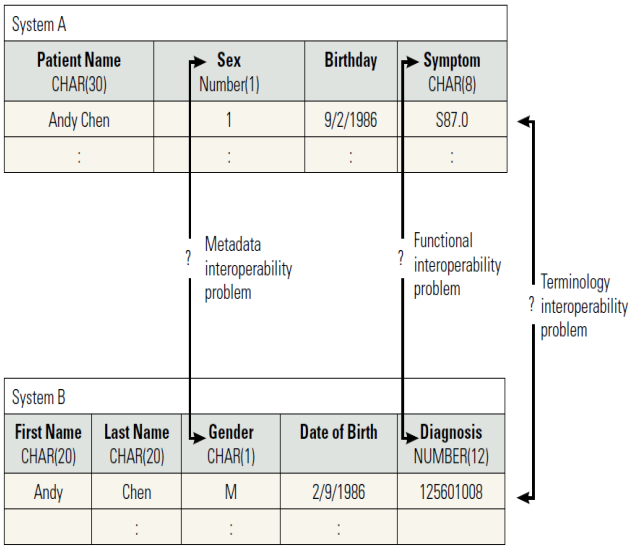

FIGURE 2. Health big data interoperability (source [19]).

D. DATA ANALYSIS

Depending on the complexity of a health problem the traditional SQL based querying time increases exponentially as the number of records increases. The hardware and software needed to analyze data need to be so robust and expensive to get a meaning from the huge dataset. In terms of hardware Supercomputers and cloud computing are the most widely used. The aim of analyzing health data is to apply a predictive model to predict probable occurrences and complications. MapReduce programming model and its Hadoop implementation provide robust analytical tools to do the analysis.

IV. HEALTH BIG DATA SOURCES

Health big data sources are so innumerable and diverse. The sources depend on the level of technology that a health entity uses. Cyber-Physical systems, medical Internet of Things (mIoT), social networks, Electronic Medical Records, and genomic data are the big contributors to health data and are covered in this study.

A. ELECTRONIC MEDICAL RECORDS (EMR)

The primary source for any health analytic solution is the EMR also known as EHR. This is a hospital-based system that combines all the entries that are logged by health practitioners. To obtain an effective health analytic solution, it is paramount that the other sources of data be able to synchronize with the EMR. The EMR is a collection of entries that include doctor notes, diagnosis history, pharmacy data, and the insurance company's data. This aggregation of actors makes its design so complex that its adoption is sometimes hindered.

The challenges faced in designing a proper EMR start at the stage of data entry by the physicians. The traditional method for data entry which is straightforward for practitioners is using an easy to use Word document and a spreadsheet. however, this method exhibit difficulties in providing a meaningful insight through an appropriate analytic algorithm. The analytic solution should be able to process unstructured data while the EMR contains mainly structured data hence it is necessary to transform these unstructured entries.

Yang [38] proposed an XML-based scheme that consists of facilitating both the physicians as well as analytical solution designers. The solution consists of using an XML schema to process entries that were recorded in the usual word processor or spreadsheet fashion and transform it into XML data to be processed as structured data inside the EMR.

Another challenge with the EMR is the interoperability. A patient is treated at different hospitals, which may leave his health data scattered across various EMRs. Interoperability problems between hospitals still pose a big barrier for systems integration. This hindrance is a big obstacle to healthcare as with the absence of prior treatments and the tests, patients undergo different and repeated treatments for the same illness. There are two big questions to consider. The first is the access infrastructure, which is mostly standalone for each hospital and the second is the complexity of the integration of health systems in the context of a country or any other geographical entity.

To solve the access challenge, Wan and Sankaranarayanan [39] proposed a cloud-based EMR that can help all stakeholders to access the EMR by making use of three cloud computing components: Software as a service (SaaS), platform as a service (PaaS), and infrastructure as a service (IaaS). The system allows mobile access for all stakeholders. Additionally, another big challenge to consider is confidentiality as not many hospitals permit their records to be accessed by others.

To solve the compatibility problems, various studies have been conducted and one of the most trusted technologies is Blockchain [169]. The Blockchain technology solves the challenge of personalization and ubiquitous access to records which are the two traits that are required in the healthcare industry. Blockchain ensures data integrity by allowing an immutable way to update records and prevents the tampering with them once they have been logged. Each entry is saved in a block and the content of each block is hashed to constitute the content of the next record.

As depicted in Fig.3, to solve the mistrust between health providers, Azaria *et al.* [40] proposed MedRec a solution that decentralizes the EMR access using Ethereum smart contracts. The smart contract will not only help the patient but can also help other stakeholders. A use case would be to help insurance companies to verify if certain medical treatments like surgeries have occurred before paying for the services.

B. MEDICAL INTERNET OF THINGS(mIoT) AND CYBER-PHYSICAL SYSTEMS (CPA) DATA

With the pervasiveness of wireless sensors, the health industry has revolutionized a lot. With the advances in wireless sensor networks, mobile health and patient remote monitors from home, a huge amount of data is now generated in real time.

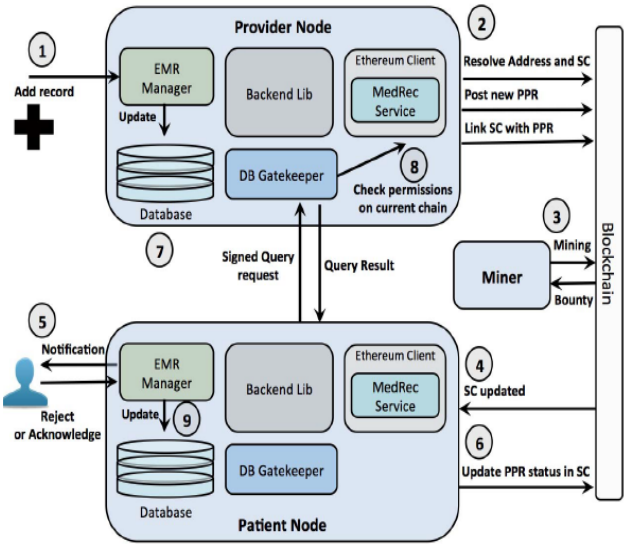

FIGURE 3. Steps for a health provider adding patient records on an Ethereum blockchain (source [40]) .

TABLE 2. Summary of medical IoT devices.

Sensor category	Type	Uses	Example
Pulse Oximetry	Wearable	heart rate (HR) and blood oxygen saturation (SpO2)	Sensor networks for medical care [30]
Monitoring and Alert Systems	Wearable & implant	Monitoring conditions and alerting for high-risk patients	AMON [31] (wearable) Cardiac patients monitoring [32] AlarmNet [33] (Implant)
Tracking systems	Wearables	Used to know the geographical location of a patient	Health and Usage Monitoring System (HUMS [34])
Community health monitoring	Community-based systems	A networked system to monitor key health cases in a community	chitsMS in Philippines [35]
Movement detection	Wearable and implants	used to monitor any movement of a patient	[36], [37]

For a long time, EMR was believed to be the sole provider of patient-related information. However, with the introduction of real-time health systems (RTHS), the EMR is just a part of the overall health IT ecosystem. Dimitrov [25] has proposed a conceptual structure of an RTHS system that includes mIoT devices. Patient-based wearables, such as bio-shirts, and body implants (smart capsules) feed data into the EMR, which contains clinical data of the patient. A health analytic solution can then help a physician to treat an emergency or obtain an insight regarding an impending one. Various wearables that perform certain tasks like fall detection, position detection, glucose level monitoring, location tracking, have been developed. To integrate various sensors with different standards, customized middleware must be developed. Lu and Chen [27] proposed a middleware design for Tele-homecare applications. It consists of two layers, a device management layer as well as the data management layer. This Middleware which is an interface between vital signs sensors and tele-homecare applications receives vital signs data from the sensors and channels them up to the tele-homecare applications. Various other studies such as those in [26], [28], and [29] have focused on hardships to consider when designing a middleware for various remote sensing devices in healthcare. Medical IoT devices are used to monitor the physiological vital signs of inpatients as well as outpatients with temporary or chronic diseases. The main categories of vital signs these devices measure are the respiratory rate (RR), body temperature (BT), blood pressure (BP), oxygen saturation (SO2), blood glucose (BG), heart rate (HR), and so on. Table 2 summarizes the IoT sensors that are used in healthcare as well as their properties.

The advances in mIoT permit a remote diagnosis for patients who cannot arrive at hospitals. They also help to supervise the patients who are incapacitated and monitor them from the comfort of their home beds. This pervasiveness also can permit self-treatments especially in the earlier development of sicknesses. However, such easy availability would require appropriate solutions that can integrate IoT data with the EMR data once the patient is hospitalized and clinical tests are conducted.

C. SOCIAL NETWORK DATA

Traditionally, it is assumed that the physicians and their measuring equipment are the only sources of a patient's data. With the advent of behavioral informatics and their applications to personalized healthcare systems, it is now possible for the patients to be collaborators in their treatments. For example, consider a teen who visits his hospital for treatment and the physician performs various measurements such as ECG or MRI and feeds the readings into EMR to perform the diagnosis. But suddenly he checks the teen's Facebook page and finds out that the subject posted a suicide message last night. Treating the patient by omitting this critical information would be inappropriate. Using his own intuition without being able to incorporate the information with other readings would also not be elaborate enough. Hence, there is a need for the health analytic solution to accommodate the EMR data with information on social networks to perform a complete diagnosis.

In their study regarding human social influence, Zheng *et al.* [120] found out that social contagion alters behaviors such as the decision to go for treatment or adoption and compliance with medical prescriptions. They also concluded that social influence is apparently channeled through interactions with friends rather than in a professional context. This explains that a patient might not disclose fully his situation to a physician rather he can opt to convey his feelings to his close friends on social networks.

Social Networks Analysis(SNA) [121]–[123] uses networks and graph techniques to mine social network content. Social network data can be used as a social health support tool in a community for purposes such as raising local health awareness or crisis communication. Social media data can be integrated into clinical decision support systems and be a constituent of the diagnosis decision [124]. To obtain a unified analytics solution, it is necessary for the SNA to be integrated with other analytical tools. However, there can be possible resistance within the health institutions as data from social networks, such as self-reporting, are not sufficiently fine-grained to be incorporated with other standards that are tested. Thus, a hierarchically integrated EMR with a patient portal that contains fine-grained or censored social media data that allows a patient to collaborate with peers is necessary.

D. GENOMIC DATA

The human genome is a chain of all genes that make up a human cell. The genes are made of DNA. The Human Genome Project (HGP) performed a full sequencing of the human genome to determine patterns that might be sources of some diseases [125]. Though this process is expensive, complex, and took more than 10 years to complete, a full sequencing of the human genome is a novelty and can save lives. A sequence of DNA proteins (ATCG) in the individual's genome can decide all the human traits like the color of eyes, skin color, vulnerability to a certain disease etc. Hence, if we apply analytical techniques such as machine learning to a human genome, we can obtain significant information and insights about the subject's health condition. We can study how certain drugs work and know specifically they work on an individual. Genomic data is a backbone to the personalized healthcare [127] whereby specific treatments for a patient can be considered.

As it is for other sources of health data, an analytical solution must fully integrate the genomic data with data obtained from other sources to provide an effective diagnosis. To do so, the system must integrate molecular pathology with clinical pathology. Even with these recent developments in genome sequencing, designing and implementing a genome-enabled electronic medical recording system is challenging [128].

The big challenge is taxonomy and vocabularies necessary to transform genomic sequences into clinically meaningful descriptions that can support existing diagnoses that are deduced from EMR data. While clinical pathology uses a standard naming system, such as the SNOMED-CT, molecular pathology still falls short of a concrete standardization. Several recent studies have attempted to address this drawback. Hoffman *et al.* [127] proposed the clinical Bioinformatics Ontology (CBO), which is a semantic resource that can describe clinically meaningful genomics concepts.

Green *et al.* [129] concluded that owing to the lack of a standardized matching between phenotypes and genomic data, concrete integration and utilization of genomes for diagnosis will not be achieved until the year 2020. One of the successful studies was the eMERGE (Electronic Medical Records and Genomics) [130]. This study involved Phenome-Wide Association Studies [131], which is a process of using EMR and analyzing various phenotypes with reference to one genetic variant.

V. DATA MINING TECHNICS

Machine learning techniques are vital for predicting disease occurrences or their complications. Though all machine learning algorithms and practices can be applied to healthcare problems, each illness and its complications are best described by a single algorithm or a combination of some of them. Hence, it is necessary to inspect the algorithms closely and apply them where they are appropriate. The following algorithms are widely used in health informatics. we briefly described them as well as their specific use in diseases diagnosis.

A. K NEAREST NEIGHBOR ALGORITHM

K-nearest neighbor (KNN) [166] is among the simplest and most classical machine learning techniques used. Its use can be described as ''Tell me your close friends and I will tell you who you are''.

In building healthcare analytics, this algorithm is effective in classifying a disease by matching it with the already known cases and the resulted complications.

In healthcare, to determine if intestinal cells are cancerous, we need to classify intestinal cells in five possible classes of tumor cells: adenocarcinoma, sarcoma, carcinoid tumor, gastrointestinal tumor, and lymphoma. To classify cells under test, we can use Principal Components Analysis, a statistical feature that has more variance on data, and plots the known cell classes.

In the example shown in Fig.4, we associate the undetermined cell T with one of the classes of our training data. After testing, and assuming the cells as per their principal

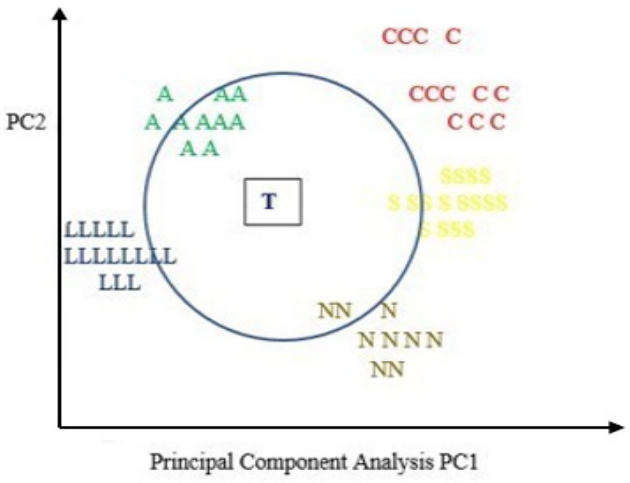

FIGURE 4. K Nearest neighbor algorithm.

components P1 and P2, we would like to find several k cells for which the sum of the distances to T is minimum (nearest neighbors of T). After calculating the sum, we conclude that the cells under inquiry are classified as Adenocarcinoma cell types. The value of k must be selected appropriately to avoid overfitting.

Shouman *et al.* [41] have proposed a method to use this algorithm to diagnose a heart condition. During their study, due to its complicity and its convergence, KNN has outperformed other classification techniques.

B. SUPPORT VECTOR MACHINES (SVM)

SVM [167] is an effective classifications algorithm that is used to classify data. As depicted in Fig. 5, the SVM splits data using a hyperplane, which is a plane that separates the nearest points that are known as support vectors. Choosing the hyperplane is a constrained optimization problem as we must optimally choose the margin between classes that is wide enough.

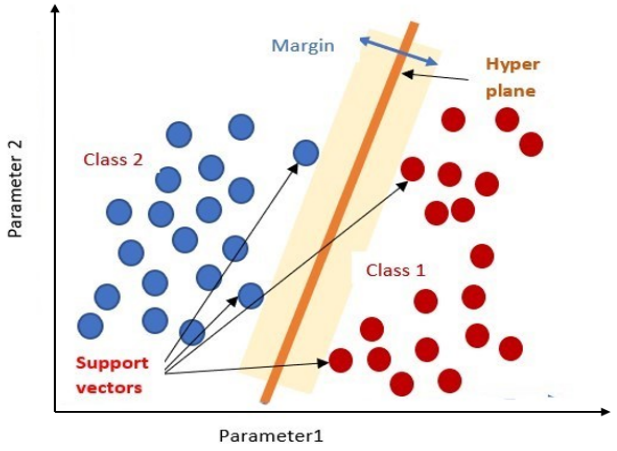

FIGURE 5. A two-dimensional SVM model.

A hyperplane is a plane that is chosen to divide the classes, and sometimes, the data cannot be separated linearly. If such problems arise we can combine the SVM with additional kernel techniques like Radial Basis Functions. Though SVM is an old machine learning technique, recent studies have resulted in robust classification capabilities.

Brown *et al.* [47] introduced a method to classify genes using gene expression data by applying SVM techniques. Furthermore, Guyon *et al.* [48] have used SVM based on Recursive Feature Elimination (RFE) and performed gene selection for cancer classification. The aim of the experiment was to determine if certain genes are active, silent, or hyperactive.

Khedher *et al.* [49] have proposed a method for early diagnosis of Alzheimer's disease using SVMs on segmented MRI images. In this study, they used Principal Component Analysis to perform feature selection using an ADNI (Alzheimer's Disease Neuroimaging Initiative) dataset and SVM to determine if a certain patient is suffering from cognitive problems of old age or is developing the Alzheimer's disease.

C. NEURAL NETWORKS

Neural networks [168] are algorithms that perform exceptionally well in grasping insights from unstructured data. Health analytics require algorithms that can receive inputs from disparate sources and extract a meaning. Medical unstructured data comprise doctors' notes, radiology scans, MRI images, microscopy, CT scans, ultrasound images, and so on. Interpreting these data using neural networks is an excellent technique.

Neural networks are a series of neuron-like layers of computation that apply a chain of computing algorithms to the input data computation cell to produce outputs. Fig.6 depicts a basic structure of a neural network. A Convolutional Neural Network(CNN) is comprised of many layers that are capable of transforming inputs by applying convolutional filters to produce a clear output.

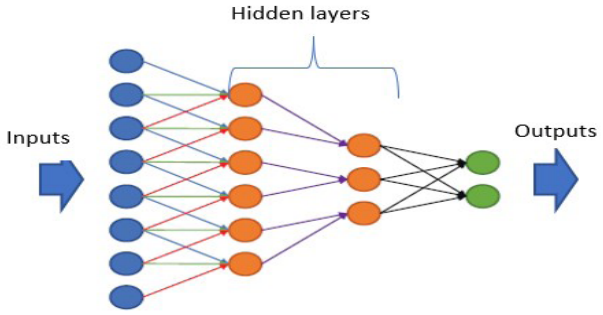

FIGURE 6. Basic structure of a neural network.

Neural networks are currently used for medical imaging, such as brain lesions analysis, fetal imaging, and cardiac analysis. To understand neural networks in the context of healthcare, consider its application in treating or predicting if a given breast tumor is malignant or benign. In this use case, inputs features can be various values of the most critical biopsies for the breast cancer which can be obtained by medical imaging. From the Breast Cancer Wisconsin Data Set [42], the breast cancer biopsies that can be inputs of neural networks are clump thickness, cell size, cell Shape, Marg adhesion cell size, bare nuclei, bland chromatin, normal nuclei, and mitosis. The practitioner can input values of these biopsies and the neural network is able to conclude if the tumors are benign or malignant.

Currently, another key application of neural networks is the analysis of brain lesions. Kamnitsas *et al.* [43] were able to use 3D CNNs to analyze MRI brain scans. The process involved capturing each brain 3D voxel, obtained from a 3D scan and applying it to a network of convolutional layers and classification layers that produce a clear inference. Another key area of the neural network application is matching medical text reports with medical images, which is critical as we

can interpret one using another. Kooi *et al.* [44] compared the traditional CAD-based mammography system, which relies on manual features, and a CNN-based mammogram.

Another interesting application of neural networks and brain activity scans is the prediction of survivability using MRI images. In Fig.7, van der Burgh *et al.* [45] combined MRI images with clinical data obtained from Amyotrophic Lateral Sclerosis (a disease that causes the death of neurons that control voluntary muscles) patients. Using this data, they were able to predict survivability by applying deep neural networks. The system was comprised of four neural networks: a network made of Clinical data, Structural connectivity from MRI, brain morphology MRI, and a combination of these three.

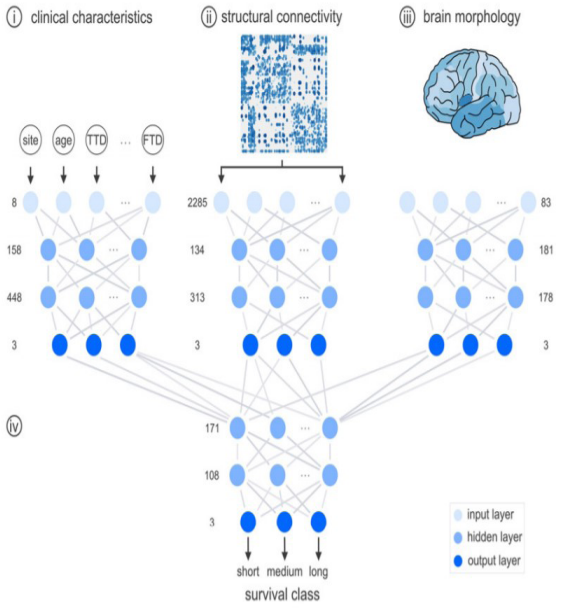

FIGURE 7. Classifying survivability using neural networks. Combining data from various source and predicting survivability as short, medium, and long (Source [45]).

By combining these networks, the system was able to predict the survivability of a patient as short, medium or long-term. In this process, the number of input nodes depended to several characteristics of each neural network. For example, the input vectors for the clinical characteristic neural network were the age at onset and time for diagnosis. The prediction accuracy was improved by 84% compared to the results that were obtained without using Neural Networks [46].

D. K-MEANS CLUSTERING TECHNIQUES

The K-means clustering technique is one of the most popular unsupervised learning techniques [81]–[83]. It is a clustering algorithm that is used in classifying points on a Euclidian plane into categories. It is an iterative technique that takes several n points and classifies them into k possible clusters around k centroids.

There are so many concrete applications of the K-means in healthcare. Gash and Eisen [84] used fuzzy k-means clustering and identified overlapping clusters of yeast genes. Fuzzy k-means allows a given point to belong to every cluster; however, with varying degree of membership. Ng *et al.* [85] applied k-means algorithm to perform medical images segmentation. Their method combined k-means with an improved watershed algorithm with k-means taking 2D MRIs as inputs and producing clustered images. The clustered images were then segmented using the watershed techniques. Zheng *et al.* [86] applied hybrid k-means and SVMs for breast cancer diagnosis. During the experiment, k-means was used for identifying various hidden patterns of benign and malignant tumors.

E. ENSEMBLE LEARNING

Medical problems are usually complex and one learning algorithm might fail to yield an informative result. Hence, we can combine the learning techniques into a more effective and robust learning technique better than the constituent algorithms. An ensemble [87], [88], [91] is a collection of various learning algorithm working together in parallel or in sequence to produce better results. It is mainly used in classification problems like sentiment classification [89]. Ensemble learning has shown to provide very efficient and favorable outcome compared to individual learning systems as its usually perceived as a process of consulting multiple experts before deciding [90]. In this method which combines more than one classifier, the outcome depends on a set of rules that are used during combination.

The starting point of an ensemble process is feeding the training data to a base algorithm then voting is used to make an accurate decision from classification outcomes of each classifier in the ensemble. The most popular ensemble learning techniques are; Bagging and Boosting [92]. during the bagging process we take the training data and divide it into bags (or subsets).We then apply an individual classifier on each subset and finally, we apply voting to produce a final prediction. Boosting is rather dependent on refining wrong predictions. We divide the training set as we do for bagging, we apply a learning algorithm to a subset, we test the classifier and check the wrong predictions. We take these cases and we combine them with a new subset and we apply another classifier and so on until we finish all the subsets. Boosting produces extremely refined decisions.

Ensemble learning has many applications in healthcare and is the best fit for most of the health problems. Gu *et al.* [93] used the FKNN ensemble learning method to classify membrane proteins using a hybrid approach of predicted secondary structural features (PSSF) as well as approximate entropy (ApEn) to predict the G-protein coupled receptors in low homology. Savio *et al.* [94] use the combination of Voxel-based Morphometry and used Support Vector Machine and neural networks to analyze structural MRI images for an earlier detection of Alzheimer's disease (AD) and myotonic dystrophy of type 1 (MD1)

F. MARKOV DECISION PROCESS (MDP)

Markov Decision Process (MDP) [170] is a powerful stochastic control algorithm used in decision support. Even though the physician is presented with appropriate algorithms for a certain health case, there are some stochastic problems that are dependent on critical operations that need a decision support system. Some decisions such as a transplant or even the diagnosis itself will need mathematical models for accurate decisions to be made. Sonnenberg and Beck [171] proposed a methodology that can be used as a practical guide for using the MDP for clinical diagnosis. At each stage of the treatment, a patient's health state is analyzed as a Markov state and death or other complications are considered as terminal nodes in this MDP. The events that influence a patient's health are modeled as transitions between nodes. For example, bleeding (an event) can cause a patient's transition from a coma (state) to death (final state). By building a Markov state diagram, critical decisions will be made and decisions which lead to catastrophic results will be omitted. Hence, MDP can be incorporated into a health analytic solution for decision-making purposes. As a synthesis, Fig.8 provides a detailed view of use cases of various data mining techniques in healthcare analytics.

VI. HEALTH BIG DATA PLATFORMS AND TOOLS

Building scalable health analytics is complex because data varies in speed, size, urgency, availability etc. To scale and accommodating the data, the analytic system must be integrated into a parallel and distributed computing framework. The most challenge with healthcare data is the diversity in volume. Analytics must be performed proactively and reactively; hence, the architecture must be sufficiently extensible to support all the analytics. The big challenge faced while selecting an appropriate platform to use for the analytic solution is to accommodate the urgency of action and required insights. While some data such as ECG readings and other mIoT data might be urgent and require in-memory streaming and immediate analysis, some other data, such as HER records, might require batch processing. Though these platforms exhibit different architecture any parallel computing platforms share components to consider while designing scalable big data analytics. Fig. 9 depicts a conceptual architecture of a big data platform. The following big data platforms are popular in health informatics:

A. HADOOP

Hadoop [58] is a parallel computing platform that stores and processes very large computing clusters on its core architecture and is empowered by three main components.

HDFS [59], which is the underlying file system distributed on clusters, has storage devices arranged in various racks. This file system is distributed, and each data block is duplicated as copies across clusters. Another key component in Hadoop is YARN [60], which is a resource management and job scheduling tool. Its role is mainly to manage the extensive storage resource and keep track of the computing workload across clusters.

The other component of Hadoop is the MapReduce [61]. Its role is to process the data stored on HDFS clusters and that role is accomplished in two functions of Map and Reduce. During a Map step, the master node will divide the job into smaller tasks and distribute the resources based on the task. After computations, the Reduce function aggregates all results to produce a solution to the original problem. Fig. 10 provides the detailed operational steps of the Hadoop framework.

1) HADOOP FOR HEALTH BIG DATA ANALYTICS

Hadoop has the potential to be used in building a healthcare analytics solution. However as discussed, it is a batch-only big data platform; hence, it cannot leverage fully the potential of real-time emergencies like ECG reading, whereby a patient may need an immediate attention as per the alarm. Various researches have introduced medical products that are built on Hadoop. Lijun *et al.* [66] have proposed Medoop, a Hadoop-based medical platform. This system leverages the attributes of scalability, high reliability and high throughput of Hadoop. Sweeney *et al.* [67] developed Hadoop Image Processing Interface (HIFI) which is used for Image-based MapReduce activities. Studies in [68] and [69] have also attempted to leverage the attributes of Hadoop and MapReduce for healthcare.

2) PROBLEMS WITH HADOOP MAPREDUCE

Though Hadoop is widely used it has some flaws [62]. The first flaw is that it is strictly a batch computing platform. Hadoop is mainly designed to perform computing loads in batches hence not appropriate for real-time streaming applications where immediate insights are required.

Another flaw with Hadoop is the Skew problem. This problem is observed during the Map and Reduce operation. After a Map step, the Reduce function must be notified regarding the availability of data before a Reduce operation. This elapsed time is called a shuffle. When there is an imbalance in computational loads between the two steps it can cause the execution time of one of them to delay and causing the skew problem [63]. However, various studies have tried to address this problem [64], [65].

B. HIGH PERFORMANCE COMPUTING CLUSTER

High performance computing cluster (HPC) [160] is a high-speed computing paradigm made by a group of servers connected with a dedicated high-speed network. These individual servers are in some cases powered by arrays of GPUs (Graphical Processing Units). Each node in these clusters must solve a computing task. The cluster management is performed by a master node which also ensures proper parallelization using specific tools like OpenMP. Healthcare analytics need a high-speed robust computing paradigm hence HPD is highly regarded in building up health informatics.

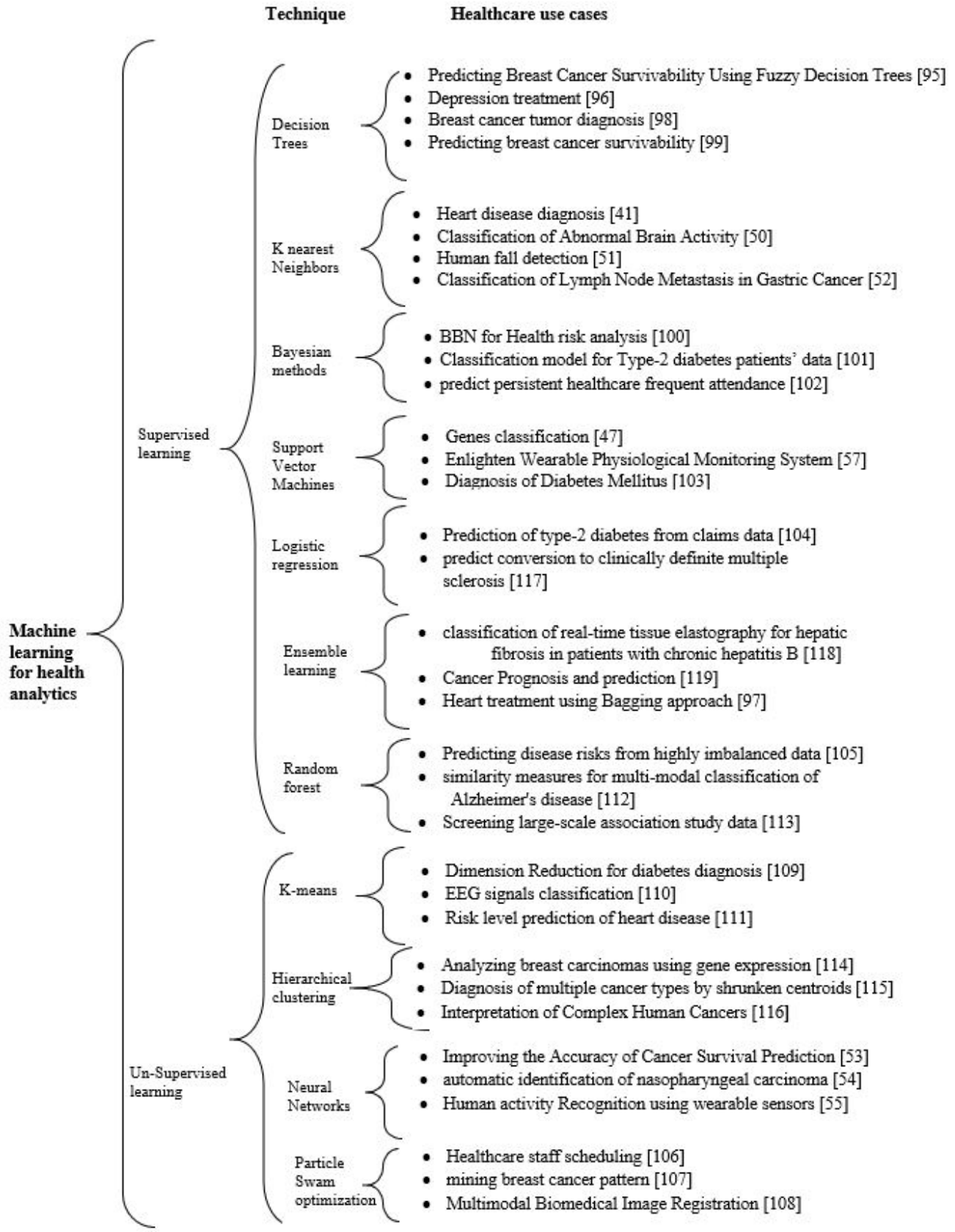

FIGURE 8. Popular Machine learning techniques and their use cases in health analytics.

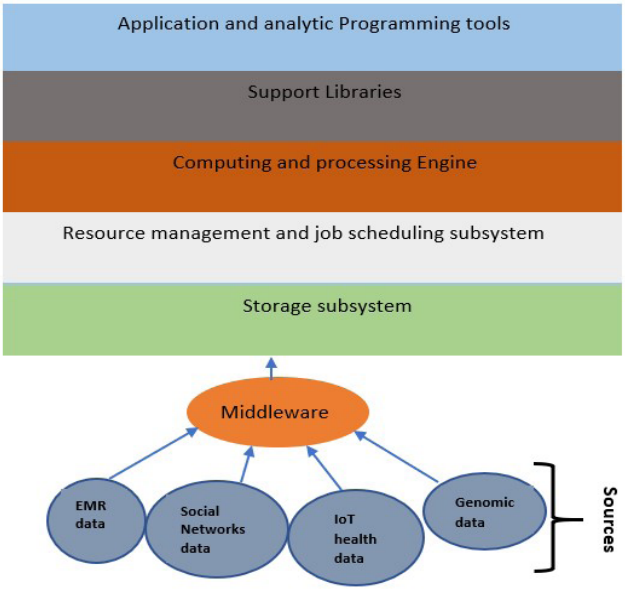

FIGURE 9. A conceptual architecture of a health big data platform .

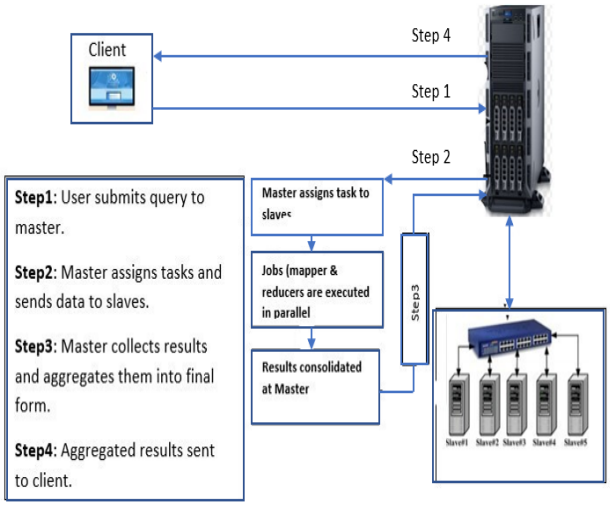

FIGURE 10. The architecture of the Hadoop cluster and a simplified MapReduce Operation.

Samant *et al.* [157] have used HPC for the analysis of Deformable Image Registration(DIR). As DIR is a process that requires fast near real-time online analysis, the use of HPC computing offered significant performance acceleration for all the algorithms that were implemented.

C. SPARK

Spark was introduced as an improvement of Hadoop. It is a fast, general purpose, cluster computing platform for big data with a high-level API in Java, Scala, R, and Python. Spark can capture batch, streaming, and interactive jobs in a solution that combines some or all of them [70]. Perhaps the main capability of spark is the capability for the integration of a powerful machine learning library (MLlib) [56]. The platform extends the MapReduce programming model with a data-sharing abstraction named "Resilient Distributed Datasets," RDD [71].

The importance of RDD provides one advantage. With its immutability, a single data object might be accessed in parallel and remain unaltered. Spark provides a faster parallel computing method than MapReduce. An example is the execution of the gradient descent for the minimization of the cost function in the logistic regression algorithm. To minimize the cost function, the gradient descent must be executed in a multitude of iterations, hence requiring more computing power.

Because Spark uses in-memory computing, it performs extremely faster. In the gradient descent execution, MapReduce takes around 110 s for each iteration as the data must be loaded from the disks. However, for the same operation, Sparks uses only one second for one iteration as it only needs to perform the first loading and the remaining iterations must be completed without memory loads. The big advantage that Spark demonstrates over MapReduce is that while MapReduce requires Impala [141] for querying operations and Mahout [146] for machine learning, for Spark you can develop an application and access all engines using a unified API. As an example, you would want to develop an application which needs to process user queries (SQL engine), predicts outcomes (Machine Learning) and maps user relationships (GraphX library) by using a single engine and in the main memory.

Spark is a heavily used platform for healthcare big data analytics. It leverages its stream computing capabilities to perform faster analysis without the need to use other supportive frameworks. Wiewiórka *et al.* [72] proposed SparkSeq [73], a cloud-based genomic data analysis with nucleotide precision built on Spark. MacDonald [74] implemented the COPA (Cancer Outlier Profile Analysis) a system that analyze genes expression to detect repeated translocations for a given cancer type. Freeman *et al.* [75] built an analytical tool called Thunderbuilt that uses Apache Spark to analyze large-scale neural data obtained from a larval zebrafish brain. Apache Spark was used in this study to account for the high volume of data generated by neural recordings.

D. FLINK

Most of the big data platforms were designed for batch processing. However, data that needs real-time processing is increasing tremendously. The number of applications such as Twitter analytics, weblogs, and fraud detection has increased, and they need a real-time processing analysis. Even so, batch streaming is not dropped. Modern analytics application must encompass both batch and real-time data. Apache Flink [80] can process continuous data streams at the same time acknowledging that there is also a need to process historical batch data. From Fig.11, we can see that even if Flink is built on a streaming processing engine, it has a Dataset API that processes batch datasets. Flink also offers a rich library

TABLE 3. Summary of big data platforms, their properties and ther use cases in healthcare analytics.

Platform/tool	Uses	Batch/stream	advantages	Where it is better used	Health analytics uses cases
Apache Spark [70]	platform	Stream +batch interactive	Excellent as it relies on HDFS and avoids writing on disks as much as possible	Uses a unified API. better for applications requiring high speed	Analysis of large-scale functional MRI data [164]
MapReduce [58]	Platform	Only Batch	Excellent as it relies on HDFS	Excellent for Java-based solutions and Good for ETL tasks	Large-scale Chinese traditional medicine data processing [163]
Apache Storm [79]	Platform	Only Stream	Fault tolerance is excellent as it relies on Nimbus and zookeeper	Better used where real-time is the only concern.	urgent decision making in cardiological ambulance control [165]
Apache Impala [141]	SQL Engine for Hadoop	Only interactive	Built on top HDFS which is a fault tolerant system	RDBMS-like experience with Hadoop flexibility and scalability	OHDSI [142]
Apache Flink [80]	framework	Stream +batch	Fault tolerance is excellent via checkpointing and partial re-execution	Applications where stream data can be batched like Logs, Sensor data.	Genomic analysis [143]
Apache Mahout [146]	Machine learning engine for Hadoop	Batch	Rich libraries but it's slow	Good for recommender systems as well as classification systems	Predicting risk of readmission for heart patients [145]
Hive [147]	Warehousing over MapReduce	Batch	Runs on HDFS hence enjoy features like scalability, fault tolerance and redundancy	Used when we want to have an SQL experience while analyzing large datasets	I2b2 hive for Predicting Asthma Exacerbations [148]
Cassandra [149]	No-SQL database	Batch +stream	Elastic scalability and no single point of failure	Better used for applications that have heavy data	Ubiquitous sensor-based care [144]
MongoDB [150]	No-SQL database	Batch+ stream	Schema-less does not use tables it stores data as documents in JSON style.	Good for Mobile and social networking applications.	BIGNASim: analysis portal for nucleic acids simulation data [151]
Nvidia CUDA [159]	Parallel computing platform	Batch + stream	Massively parallel hardware on a single unit. High speed	Better used for image processing applications	positron emission tomography (PET for medical image reconstruction [155]
HPC [160]	Parallel computing platform	Batch Stream	Aggregation of computing power. High speed	better in modeling physical phenomena.	Deformable image registration (DIR) in medicine [157]
GPU [161]	Parallel programmable processor	Batch/+stream	High-speed parallel computing	Better used in Graphic based applications	*Proton computed tomography [154] *intensity-modulated radiation therapy (IMRT) [156]
Field Programmable Gate Array [162]	An integrated circuit to be reprogrammed	Batch+ stream	High performance and versatile as it can be reprogrammed as per the application	Better used for customizable embedded systems.	Wearable Electrocardiogram (ECG) acquisition [158]
Apache Kafka [152]	Stream processing software platform	Stream	Handle high-speed streams without big hardware. It's also highly fault-tolerant.	Better for real-time data like Tweets, Log aggregation and instant messaging.	storing and retrieving bioinformatic data [153]

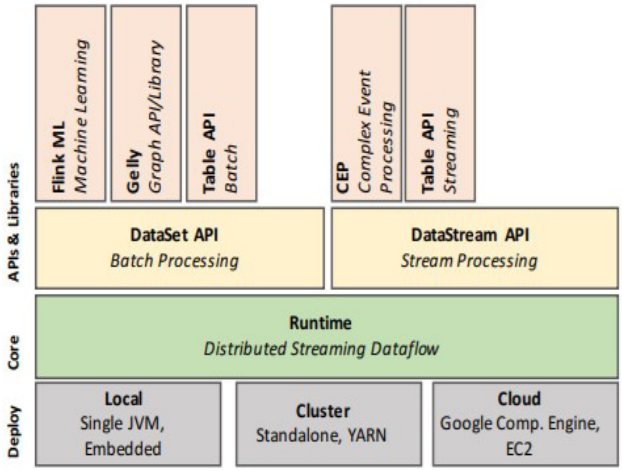

FIGURE 11. Flink architecture (source [80]).

support including machine learning, Graph API, and table API to process SQL like operations.

FLINK USE CASES IN HEALTHCARE

Apache Flink is an excellent platform of choice for event-driven applications. Moreover, health analytics need to move away from a traditional reactive approach to a more proactive approach. Hence, a streaming platform is a choice for real-time health monitoring aspects of a health analytic application such as ECG monitoring, MRI readings, wearables monitoring, and other cyber-physical systems.

E. STORM

Apache Storm [79] is an alternative for Hadoop MapReduce when we need heavy real-time processing. Hence, when our analytic solution is expected to process huge data at a rapid rate, Apache Storm is the best performer for building such solutions. The key advantages of Storm are that it can work well for small and large-scale implementations and it is fault-tolerant, scalable, and exhibits a higher reliability. Apache storm is superficially like Hadoop and while jobs are run in Hadoop, topologies are run in Storm. However, the key difference between jobs and topologies is that finally, a MapReduce job will finish while a Storm topology continues to handle incoming messages until the user terminates the process. The most basic data structure is a tuple made of a pair of elements and a given number of these tuples make a stream.

CRITERIA FOR CHOOSING A BIG DATA PLATFORM FOR A GIVEN HEALTHCARE APPLICATION

The choice of a big data platform for your health solution depends on many factors: Real-time needs, data size, speed, scalability, throughput etc.... Some healthcare applications like EMR records might not need real-time processing, hence for such application, a non-streaming platform like Hadoop MapReduce is enough. For others like ECG analysis will need a real-time intervention hence streaming will be paramount but scalability will not be a problem to consider for such applications.

Some recommendation applications like diagnosis suggestion support will require a platform which can scale and accommodate the huge amount of data, in that case, a vertical scaling platform like CUDA [76] or High-Performance Computing (HPC) [77] will be of no use but horizontal scaling systems like Spark will be extremely useful. However, a health analytic solution encompasses many aspects of health analytic requirements hence the need to choose a platform that can be suitable for all these requirements. Another aspect to consider is fault-tolerance. This is the platform's ability to continue operating even after failure. To this aspect, all though not equally the horizontal scaling platforms like Hadoop and Spark wins as they distribute the work across many clusters whereas vertical scaling platforms have one point of failure. For a thorough comparison, Singh and Reddy [78] have compared the platforms with respect to their capabilities to perform a K-means machine learning algorithm. In this study though many criteria to consider for choosing an appropriate platform for a healthcare application are presented, few criteria are considered as paramount in processing healthcare big data. In table 3 we cover the most popular big data platforms and their use cases in healthcare analytics.

VII. CONCLUSION

The development of a scalable healthcare analytic application requires the amalgamation of various technologies whose choice requires a thorough scrutiny as a successful diagnosis and disease deterrence reckons on the incorporation of as many data sources as possible. In this work, we have highlighted most of the technologies to choose from to do so. The very first challenge is to choose from Big data platforms on which the application must rely upon. The platform should have all necessary libraries including the machine learning libraries. Spark provides integrated models for effective data ingestion, data processing as well as effective Machine learning libraries. Effective diagnosis requires medical images to be closely analyzed hence deep learning algorithms are better feet for identifying malignant spots on these images. Convolutional Neural Networks(CNN) which applies optimization algorithms like the Gradient Descent provides an effective analysis of life-threatening spots on images that are obtained from measurements like ECG, MRI, etc. As medical application needs enormous precision where a slight error can result in fatalities, utmost precision is required. The most challenging task in developing a robust solution is the aggregation of all data from divergent sources. As the semantics and context of data vary in a healthcare setting, the development of an inclusive middleware is the key to ingestion of all data. For data mining, no single algorithm provides a fit-all solution to health data. Hence an ensemble learning which includes the use of many machine learning algorithms can provide a better analysis. Social networks data's role into healthcare will continue to grow as more

patient's health problems can be revealed on his social circles than to the physicians. This coupled with advances in Social Network Analysis as well as sentiment analysis tools will help practitioners to gain more insights than before. Genetics role in healthcare will be more evident. With the advances in human genomic researches, personalized healthcare can now be a reality. However, matching the genome sequences with medically related phenotypes is still an area which needs further researches.

REFERENCES

[1] P. Russom, "Big data analytics," *TDWI Best Pract. Rep., Fourth Quart.*, vol. 19, no. 4, pp. 1–34, 2011.

[2] J. Sun and C. K. Reddy, "Big data analytics for healthcare," in *Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2013, p. 1525.

[3] A. Kotov, "Social media analytics for healthcare," *Healthcare Data Anal.*, pp. 309–340, 2015. [Online]. Available: http://www.crcnetbase.com/doi/abs/10.1201/b18588-11

[4] S. A. Haque, S. M. Aziz, and M. Rahman, "Review of cyber-physical system in healthcare," *Int. J. Distrib. Sensor Netw.*, vol. 10, no. 4, p. 217415, 2014.

[5] Y. Demchenko, C. De Laat, and P. Membrey, "Defining architecture components of the big data ecosystem," in *Proc. Int. Conf. Collaboration Technol. Syst. (CTS)*, May 2014, pp. 104–112.

[6] R. Baheti and H. Gill, "Cyber-physical systems," *Impact Control Technol.*, vol. 12, pp. 161–166, Mar. 2011.

[7] Y. Zhang, M. Qiu, C.-W. Tsai, M. M. Hassan, and A. Alamri, "Health-CPS: Healthcare cyber-physical system assisted by cloud and big data," *IEEE Syst. J.*, vol. 11, no. 1, pp. 88–95, Mar. 2017.

[8] R. H. Miller and I. Sim, "Physicians' use of electronic medical records: Barriers and solutions," *Health Affairs*, vol. 23, no. 2, pp. 116–126, 2004.

[9] T. J. Hannan, "Electronic medical records," in *Health Informatics—An Overview*, E. Hovenga, M. Kidd, and B. Cesnik, Eds. Melbourne, VIC, Australia: Churchill Livingstone, 1996, p. 133.

[10] D. A. Ludwick and J. Doucette, "Adopting electronic medical records in primary care: Lessons learned from health information systems implementation experience in seven countries," *Int. J. Med. Inform.*, vol. 78, no. 1, pp. 22–31, 2009.

[11] E. Vlahu-Gjorgievska and V. Trajkovik, "Towards collaborative health care system model-COHESY," in *Proc. IEEE Int. Symp. World Wireless, Mobile Multimedia Netw. (WoWMoM)*, Jun. 2011, pp. 1–6.

[12] D. Soudris *et al.*, "AEGLE: A big bio-data analytics framework for integrated health-care services," in *Proc. Int. Conf. Embedded Comput. Syst., Archit., Modeling, Simulation (SAMOS)*, Jul. 2015, pp. 246–253.

[13] W. Raghupathi and V. Raghupathi, "Big data analytics in healthcare: Promise and potential," *Health Inf. Sci. Syst.*, vol. 2, no. 1, p. 3, 2014.

[14] L. Wang and C. A. Alexander, "Big data in medical applications and healthcare," *Amer. Med. J.*, vol. 6, no. 1, p. 1, 2015.

[15] R. Lin, Z. Ye, H. Wang, and B. Wu, "Chronic diseases and health monitoring big data: A survey," *IEEE Rev. Biomed. Eng.*, vol. 11, pp. 275–288, 2018.

[16] M.-H. Kuo, T. Sahama, A. W. Kushniruk, E. M. Borycki, and D. K. Grunwell, "Health big data analytics: Current perspectives, challenges and potential solutions," *Int. J. Big Data Intell.*, vol. 1, nos. 1–2, pp. 114–126, 2014.

[17] A. J. Cox, M. J. Bauer, T. Jakobi, and G. Rosone, "Large-scale compression of genomic sequence databases with the Burrows–Wheeler transform," *Bioinformatics*, vol. 28, no. 11, pp. 1415–1419, 2012.

[18] X. Wang and Y. Tan, "Application of cloud computing in the health information system," in *Proc. Int. Conf. Comput. Appl. Syst. Modeling (ICCASM)*, vol. 1, Oct. 2010, p. V1-179.

[19] M.-H. Kuo, A. Kushniruk, and E. Borycki, "A comparison of national health data interoperability approaches in Taiwan, Denmark and Canada," *Electron. Healthcare*, vol. 10, no. 2, pp. 14–25, 2011.

[20] A. Moreno-Conde *et al.*, "Clinical information modeling processes for semantic interoperability of electronic health records: Systematic review and inductive analysis," *J. Amer. Med. Inform. Assoc.*, vol. 22, no. 4, pp. 925–934, 2015.

[21] E. Fernandez and T. Sorgente, "An analysis of modeling flaws in HL7 and JAHIS," in *Proc. ACM Symp. Appl. Comput.*, 2005, pp. 216–223.

[22] A. Hasman, "HL7 RIM: An incoherent standard," in *Proc. Ubiquity, Technol. Better Health Aging Soc. (MIE)*, vol. 124, 2006, p. 133.

[23] R. Crichton, D. Moodley, A. Pillay, R. Gakuba, and C. J. Seebregts, "An architecture and reference implementation of an open health information mediator: Enabling interoperability in the Rwandan health information exchange," in *Proc. Int. Symp. Found. Health Informat. Eng. Syst.*, 2012, pp. 87–104.

[24] T. Mudaly, D. Moodley, A. Pillay, and C. J. Seebregts, "Architectural frameworks for developing national health information systems in low and middle income countries," in *Proc. Enterprise Syst. Conf. (ES)*, Nov. 2013, pp. 1–9.

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

[26] A. B. Waluyo, S. Ying, I. Pek, and J. K. Wu, "Middleware for wireless medical body area network," in *Proc. IEEE Biomed. Circuits Syst. Conf. (BIOCAS)*, Nov. 2007, pp. 183–186.

[27] H.-F. Lu and J.-L. Chen, "Design of middleware for tele-homecare systems," *Wireless Commun. Mobile Comput.*, vol. 9, no. 12, pp. 1553–1564, 2009.

[28] C. R. Leite, B. G. De Araújo, R. A. M. de Valentim, G. B. Brandão, and A. M. Gueirreiro, "Middleware for remote healthcare monitoring," in *Proc. Int. Conf. Innov. Inf. Technol. (IIT)*, Dec. 2009, pp. 185–189.

[29] S. Spahni, J.-R. Scherrer, D. Sauquet, and P.-A. Sottile, "Middleware for healthcare information systems," *Stud. Health Technol. Inform.*, vol. 52, no. 1, pp. 212–216, Nov. 1998.

[30] V. Shnayder, B. R. Chen, K. Lorincz, T. R. F. Fulford-Jones, and M. Welsh, "Sensor networks for medical care," Division Eng. Appl. Sci., Harvard Univ., Cambridge, MA, USA, Tech. Rep. TR-08-05, 2005

[31] U. Anliker *et al.*, "AMON: A wearable multiparameter medical monitoring and alert system," *IEEE Trans. Inf. Technol. Biomed.*, vol. 8, no. 4, pp. 415–427, Dec. 2004.

[32] P. Kakria, N. K. Tripathi, and P. Kitipawang, "A real-time health monitoring system for remote cardiac patients using smartphone and wearable sensors," *Int. J. Telemed. Appl.*, vol. 2015, p. 8, Jan. 2015.

[33] A. D. Wood *et al.*, "Context-aware wireless sensor networks for assisted living and residential monitoring," *IEEE Netw.*, vol. 22, no. 4, pp. 26–33, Jul./Aug. 2008.

[34] V. Rouet and J. Venet, "Low power tracking system for advanced health monitoring," in *Proc. Integr. Miniaturized Syst.-MOMS, MOEMS, ICS Electron. Compon. (SSI)*, 2008, pp. 1–3.

[35] R. Manguni, Jr., M. L. Navarro, K. Rosario, and C. A. Festin, "chitSMS: Community health information tracking system using short message service," in *Proc. 3rd Int. Conf. Hum.-Centric Comput. (HumanCom)*, Aug. 2010, pp. 1–6.

[36] E. Hanada, T. Seo, and H. Hata, "An activity monitoring system for detecting movement by a person lying on a bed," in *Proc. IEEE 3rd Int. Conf. Consum. Electron. Berlin (ICCE-Berlin)*, Sep. 2013, pp. 1–3.

[37] S. Oniga, A. Tisan, and R. Bólyi, "Activity and health status monitoring system," in *Proc. IEEE 26th Int. Symp. Ind. Electron. (ISIE)*, Jun. 2017, pp. 2027–2031.

[38] H. Yang, "Design and implementation of electronic medical record template based on XML schema," in *Proc. 2nd World Congr. Softw. Eng. (WCSE)*, vol. 1, Dec. 2010, pp. 225–228.

[39] A. T. Wan and S. Sankaranarayanan, "Development of a Health Information System in the Mobile Cloud Environment," in *Proc. IEEE Int. Conf. Embedded Ubiquitous Comput. High-Perform. Comput. Commun. (HPCC_EUC)*, Nov. 2013, pp. 2187–2193.

[40] Azaria, A. Ekblaw, T. Vieira, and A. Lippman, "Medrec: Using blockchain for medical data access and permission management," in *Proc. Int. Conf. Open Big Data (OBD)*, Aug. 2016, pp. 25–30.

[41] M. Shouman, T. Turner, and R. Stocker, "Applying k-nearest neighbour in diagnosing heart disease patients," *Int. J. Inf. Educ. Technol.*, vol. 2, no. 3, pp. 220–223, 2012.

[42] Accessed: Aug. 18, 2018. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original)

[43] K. Kamnitsas *et al.*, "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation," *Med. Image Anal.*, vol. 36, pp. 61–78, Feb. 2017.

[44] T. Kooi *et al.*, "Large scale deep learning for computer aided detection of mammographic lesions," *Med. Image Anal.*, vol. 35, pp. 303–312, Jan. 2017.

[45] H. K. van der Burgh, R. Schmidt, H.-J. Westeneng, M. A. de Reus, L. H. van den Berg, and M. P. van den Heuvel, "Deep learning predictions of survival based on MRI in amyotrophic lateral sclerosis," *NeuroImage, Clin.*, vol. 13, pp. 361–369, Jan. 2017.
[46] C. Schuster, O. Hardiman, and P. Bede, "Survival prediction in Amyotrophic lateral sclerosis based on MRI measures and clinical characteristics," *BMC Neurol.*, vol. 17, no. 1, p. 73, 2017.
[47] M. P. Brown *et al.*, "Knowledge-based analysis of microarray gene expression data by using support vector machines," *Proc. Nat. Acad. Sci. USA*, vol. 97, no. 1, pp. 262–267, 2000.
[48] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Mach. Learn.*, vol. 46, nos. 1–3, pp. 389–422, 2002.
[49] L. Khedher *et al.*, "Early diagnosis of Alzheimer's disease based on partial least squares, principal component analysis and support vector machine using segmented MRI images," *Neurocomputing*, vol. 151, pp. 139–150, Mar. 2015.
[50] W. A. Chaovalitwongse, Y. J. Fan, and R. C. Sachdeo, "On the time series k-nearest neighbor classification of abnormal brain activity," *IEEE Trans. Syst., Man, Cybern. A, Syst. Humans*, vol. 37, no. 6, pp. 1005–1016, Nov. 2007.
[51] C.-L. Liu, C.-H. Lee, and P.-M. Lin, "A fall detection system using k-nearest neighbor classifier," *Expert Syst. Appl.*, vol. 37, no. 10, pp. 7174–7181, 2010.
[52] C. Li *et al.*, "Using the K-nearest neighbor algorithm for the classification of lymph node metastasis in gastric cancer," *Comput. Math. Methods Med.*, vol. 2012, 2012, Art. no. 876545.
[53] H. B. Burke *et al.*, "Artificial neural networks improve the accuracy of cancer survival prediction," *Cancer*, vol. 79, no. 4, pp. 857–862, 1997.
[54] M. A. Mohammed, M. K. A. Ghani, R. I. Hamed, D. A. Ibrahim, and M. K. Abdullah, "Artificial neural networks for automatic segmentation and identification of nasopharyngeal carcinoma," *J. Comput. Sci.*, vol. 21, pp. 263–274, Jul. 2017.
[55] W. Jiang and Z. Yin, "Human activity recognition using wearable sensors by deep convolutional neural networks," in *Proc. 23rd ACM Int. Conf. Multimedia*, 2015, pp. 1307–1310.
[56] X. Meng *et al.*, "Mllib: Machine learning in apache spark," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 1235–1241, 2016.
[57] Y. Geng, J. Chen, R. Fu, G. Bao, and K. Pahlavan, "Enlighten wearable physiological monitoring systems: On-body RF characteristics based human motion classification using a support vector machine," *IEEE Trans. Mobile Comput.*, vol. 15, no. 3, pp. 656–671, Mar. 2016.
[58] D. Borthakur, "The Hadoop distributed file system: Architecture and design," *Hadoop Project Website*, vol. 11, p. 21, Aug. 2007.
[59] Borthakur, "HDFS architecture guide," *Hadoop Apache Project*, vol. 53, pp. 1–13, 2008.
[60] V. K. Vavilapalli *et al.*, "Apache Hadoop yarn: Yet another resource negotiator," in *Proc. 4th Annu. Symp. Cloud Comput.*, 2013, p. 5.
[61] J. Dean and S. Ghemawat, "MapReduce: Simplified data processing on large clusters," *Commun. ACM*, vol. 51, no. 1, pp. 107–113, 2008.
[62] A. Alam and J. Ahmed, "Hadoop architecture and its issues," in *Proc. Int. Conf. Comput. Sci. Comput. Intell. (CSCI)*, vol. 2, Mar. 2014, pp. 288–291.
[63] Y. Kwon, M. Balazinska, B. Howe, and J. Rolia, "A study of skew in MapReduce applications," in *Proc. Open Cirrus Summit*, vol. 11, 2011, pp. 1–5.
[64] Y. Kwon, K. Ren, M. Balazinska, B. Howe, and J. Rolia, "Managing skew in Hadoop," *IEEE Data Eng. Bull.*, vol. 36, no. 1, pp. 24–33, Mar. 2013.
[65] Y. Kwon, M. Balazinska, B. Howe, and J. Rolia, "Skewtune: Mitigating skew in MapReduce applications," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2012, pp. 25–36.
[66] W. Lijun, H. Yongfeng, C. Ji, Z. Ke, and L. Chunhua, "Medoop: A medical information platform based on Hadoop," in *Proc. IEEE 15th Int. Conf. e-Health Netw., Appl. Services (Healthcom)*, Oct. 2013, pp. 1–6.
[67] C. Sweeney, L. Liu, S. Arietta, and J. Lawrence, "HIPI: A Hadoop image processing interface for image-based MapReduce tasks," *Chris. Univ. Virginia*, vol. 2, no. 1, pp. 1–5, 2011.
[68] M.-H. Kuo, D. Chrimes, B. Moa, and W. Hu, "Design and construction of a big data analytics framework for health applications," in *Proc. IEEE Int. Conf. Smart City/SocialCom/SustainCom (SmartCity)*, Dec. 2015, pp. 631–636.
[69] Q. Yao, Y. Tian, P.-F. Li, L.-L. Tian, Y.-M. Qian, and J.-S. Li, "Design and development of a medical big data processing system based on Hadoop," *J. Med. Syst.*, vol. 39, no. 3, no. 3, p. 23, 2015.
[70] M. Zaharia *et al.*, "Apache spark: A unified engine for big data processing," *Commun. ACM*, vol. 59, no. 11, pp. 56–65, 2016.
[71] M. Zaharia *et al.*, "Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing," in *Proc. 9th USENIX Conf. Netw. Syst. Design Implement.*, 2012, pp. 1–2.
[72] M. S. Wiewiórka, A. Messina, A. Pacholewska, S. Maffioletti, P. Gawrysiak, and M. J. Okoniewski, "SparkSeq: Fast, scalable and cloud-ready tool for the interactive genomic data analysis with nucleotide precision," *Bioinformatics*, vol. 30, no. 18, pp. 2652–2653, 2014.
[73] Accessed: Aug. 17, 2018. [Online]. Available: https://bitbucket.org/mwiewiorka/sparkseq/overview
[74] J. W. Macdonald and D. Ghosh, "COPA—Cancer outlier profile analysis," *Bioinformatics*, vol. 22, no. 23, pp. 2950–2951, 2006.
[75] J. Freeman *et al.*, "Mapping brain activity at scale with cluster computing," *Nature Methods*, vol. 11, no. 9, p. 941, 2014.
[76] H. Scherl, B. Keck, M. Kowarschik, and J. Hornegger, "Fast GPU-based CT reconstruction using the common unified device architecture (CUDA)," in *Proc. IEEE Nucl. Sci. Symp. Conf. Rec. (NSS)*, vol. 6, Oct. 2007, pp. 4464–4466.
[77] R. Buyya, *High-Performance Cluster Computing: Architectures and Systems*, vol. 1. Upper Saddle River, NJ, USA: Prentice-Hall, 1999, p. 999.
[78] D. Singh and C. K. Reddy, "A survey on platforms for big data analytics," *J. Big Data*, vol. 2, no. 1, p. 8, 2015.
[79] T. Jones, "Process real-time big data with Twitter Storm," IBM Tech. Library, New York, NY, USA, Apr. 2013. [Online]. Available: https://www.ibm.com/developerworks/opensource/library/os-twitterstorm/index.html
[80] P. Carbone, A. Katsifodimos, S. Ewen, V. Markl, S. Haridi, and K. Tzoumas, "Apache flink: Stream and batch processing in a single engine," *Bull. IEEE Comput. Soc. Tech. Committee Data Eng.*, vol. 38, no. 4, pp. 28–38, Dec. 2015.
[81] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An efficient k-means clustering algorithm: Analysis and implementation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 881–892, Jul. 2002.
[82] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A k-means clustering algorithm," *J. Roy. Stat. Soc. C Appl. Stat.*, vol. 28, no. 1, pp. 100–108, 1979.
[83] A. Likas, N. Vlassis, and J. J. Verbeek, "The global k-means clustering algorithm," *Pattern Recognit.*, vol. 36, no. 2, pp. 451–461, Feb. 2003.
[84] P. Gasch and M. B. Eisen, "Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering," *Genome Biol.*, vol. 3, no. 11, p. research0059-1, 2002.
[85] H. P. Ng, S. H. Ong, K. W. C. Foong, P. S. Goh, and W. L. Nowinski, "Medical image segmentation using k-means clustering and improved watershed algorithm," in *Proc. IEEE Southwest Symp. Image Anal. Interpretation*, Mar. 2006, pp. 61–65.
[86] B. Zheng, S. W. Yoon, and S. S. Lam, "Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms," *Expert Syst. Appl.*, vol. 41, no. 4, pp. 1476–1482, 2014.
[87] T. G. Dietterich, "Ensemble methods in machine learning," in *Proc. Int. Workshop Multiple Classifier Syst.*, 2000, pp. 1–15.
[88] D. J. C. MacKay, "Ensemble learning for hidden Markov models," Cavendish Lab., Univ. Cambridge, Cambridge, MA, USA, Tech. Rep., 1997.
[89] G. Wang, J. Sun, J. Ma, K. Xu, and J. Gu, "Sentiment classification: The contribution of ensemble learning," *Decis. Support Syst.*, vol. 57, pp. 77–93, Jan. 2014.
[90] R. Polikar, "Ensemble based systems in decision making," *IEEE Circuits Syst. Mag.*, vol. 6, no. 3, pp. 21–45, Sep. 2006.
[91] Z.-H. Zhou, *Ensemble Methods: Foundations and Algorithms*. London, U.K.: Chapman & Hall, 2012.
[92] D. Opitz and R. Maclin, "Popular ensemble methods: An empirical study," *J. Artif. Intell. Res.*, vol. 11, pp. 169–198, Aug. 1999.
[93] Q. Gu, Y.-S. Ding, and T.-L. Zhang, "An ensemble classifier based prediction of G-protein-coupled receptor classes in low homology," *Neurocomputing*, vol. 154, pp. 110–118, Apr. 2015.
[94] A. Savio *et al.*, "Neurocognitive disorder detection based on feature vectors extracted from VBM analysis of structural MRI," *Comput. Biol. Med.*, vol. 41, no. 8, pp. 600–610, 2011.
[95] U. Khan, J. P. Choi, H. Shin, and M. Kim, "Predicting breast cancer survivability using fuzzy decision trees for personalized healthcare," in *Proc. 30th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBS)*, Aug. 2008, pp. 5148–5151.

[96] C. Andreescu *et al.*, "Empirically derived decision trees for the treatment of late-life depression," *Amer. J. Psychiatry*, vol. 165, no. 7, pp. 855–862, 2008.

[97] C. Tu, D. Shin, and D. Shin, "Effective diagnosis of heart disease through bagging approach," in *Proc. 2nd Int. Conf. Biomed. Eng. Informat. (BMEI)*, Oct. 2009, pp. 1–4.

[98] W.-J. Kuo, R.-F. Chang, D.-R. Chen, and C. C. Lee, "Data mining with decision trees for diagnosis of breast tumor in medical ultrasonic images," *Breast Cancer Res. Treat.*, vol. 66, no. 1, pp. 51–57, 2001.

[99] D. Delen, G. Walker, and A. Kadam, "Predicting breast cancer survivability: A comparison of three data mining methods," *Artif. Intell. Med.*, vol. 34, no. 2, pp. 113–127, 2005.

[100] K. F. Liu and C.-F. Lu, "BBN-based decision support for health risk analysis," in *Proc. 5th Int. Joint Conf. INC, IMS IDC (NCM)*, Aug. 2009, pp. 696–702.

[101] Y. Huang, P. McCullagh, N. Black, and R. Harper, "Feature selection and classification model construction on type 2 diabetic patients' data," *Artif. Intell. Med.*, vol. 41, no. 3, pp. 251–262, 2007.

[102] T.-H. Koskela, O.-P. Ryynanen, and E. J. Soini, "Risk factors for persistent frequent use of the primary health care services among frequent attenders: A Bayesian approach," *Scandin. J. Primary Health Care*, vol. 28, no. 1, pp. 55–61, 2010.

[103] N. Barakat, A. P. Bradley, and M. N. H. Barakat, "Intelligible support vector machines for diagnosis of diabetes mellitus," *IEEE Trans. Inf. Technol. Biomed.*, vol. 14, no. 4, pp. 1114–1120, Jul. 2010.

[104] N. Razavian, S. Blecker, A. M. Schmidt, A. Smith-McLallen, S. Nigam, and D. Sontag, "Population-level prediction of type 2 diabetes from claims data and analysis of risk factors," *Big Data*, vol. 3, no. 4, pp. 277–287, 2015.

[105] M. Khalilia, S. Chakraborty, and M. Popescu, "Predicting disease risks from highly imbalanced data using random forest," *BMC Med. Inform. Decis. Making*, vol. 11, no. 1, p. 51, 2011.

[106] M. Mutingi and C. Mbohwa, "Home healthcare staff scheduling: A clustering particle swarm optimization approach," in *Proc. Int. Conf. Ind. Eng. Oper. Manage.*, Indonesia, Jan. 2014, pp. 303–312.

[107] W.-C. Yeh, W.-W. Chang, and Y. Y. Chung, "A new hybrid approach for mining breast cancer pattern using discrete particle swarm optimization and statistical method," *Expert Syst. Appl.*, vol. 36, no. 4, pp. 8204–8211, 2009.

[108] M. P. Wachowiak, R. Smolikova, Y. Zheng, J. M. Zurada, and A. S. Elmaghraby, "An approach to multimodal biomedical image registration utilizing particle swarm optimization," *IEEE Trans. Evol. Comput.*, vol. 8, no. 3, pp. 289–301, Jun. 2004.

[109] T. Santhanam and M. S. Padmavathi, "Application of K-means and genetic algorithms for dimension reduction by integrating SVM for diabetes diagnosis," *Procedia Comput. Sci.*, vol. 47, pp. 76–83, Jan. 2015.

[110] U. Orhan, M. Hekim, and M. Ozer, "EEG signals classification using the K-means clustering and a multilayer perceptron neural network model," *Expert Syst. Appl.*, vol. 38, no. 10, pp. 13475–13481, 2011.

[111] P. K. Anooj, "Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 24, no. 1, pp. 27–40, 2012.

[112] K. R. Gray, P. Aljabar, R. A. Heckemann, A. Hammers, D. Rueckert, and A. D. N. Initiative, "Random forest-based similarity measures for multi-modal classification of Alzheimer's disease," *NeuroImage*, vol. 65, pp. 167–175, Jan. 2013.

[113] K. L. Lunetta, L. B. Hayward, J. Segal, and P. van Eerdewegh, "Screening large-scale association study data: Exploiting interactions using random forests," *BMC Genet.*, vol. 5, no. 1, p. 32, 2004.

[114] T. Sørlie *et al.*, "Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications," *Proc. Nat. Acad. Sci. USA*, vol. 98, no. 19, pp. 10869–10874, 2001.

[115] Tibshirani, T. Hastie, B. Narasimhan, and G. Chu, "Diagnosis of multiple cancer types by shrunken centroids of gene expression," *Proc. Nat. Acad. Sci. USA*, vol. 99, no. 10, pp. 6567–6572, 2002.

[116] S.-O. Deininger, M. P. Ebert, A. Futterer, M. Gerhard, and C. Rocken, "MALDI imaging combined with hierarchical clustering as a new tool for the interpretation of complex human cancers," *J. Proteome Res.*, vol. 7, no. 12, pp. 5230–5236, 2008.

[117] F. Barkhof *et al.*, "Comparison of MRI criteria at first presentation to predict conversion to clinically definite multiple sclerosis," *Brain, J. Neurol.*, vol. 120, no. 11, pp. 2059–2069, 1997.

[118] Y. Chen *et al.*, "Machine-learning-based classification of real-time tissue elastography for hepatic fibrosis in patients with chronic hepatitis B," *Comput. Biol. Med.*, vol. 89, pp. 18–23, Jan. 2017.

[119] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Comput. Struct. Biotechnol. J.*, vol. 13, pp. 8–17, Dec. 2015.

[120] K. Zheng, R. Padman, and M. P. Johnson, "Social contagion and technology adoption: A study in healthcare professionals," in *Proc. AMIA Annu. Symp.*, vol. 11, Oct. 2007, p. 1175.

[121] K. Blanchet and P. James, "How to do (or not to do). . . a social network analysis in health systems research," *Health Policy Planning*, vol. 27, no. 5, pp. 438–446, 2011.

[122] P. Appel, V. F. de Santana, L. G. Moyano, M. Ito, and C. S. Pinhanez. (2018). "A Social Network Analysis Framework for Modeling Health Insurance Claims Data." [Online]. Available: https://arxiv.org/abs/1802.07116

[123] J. Scott, "Social network analysis," *Sociology*, vol. 22, no. 1, pp. 109–127, 1988.

[124] K. Denecke, "Integrating social media and mobile sensor data for clinical decision support: Concept and requirements," in *Studies in Health Technology and Informatics*, vol. 225. Amsterdam, The Netherlands: IOS Press, 2016, pp. 562–566.

[125] J. C. Venter *et al.*, "The sequence of the human genome," *Science*, vol. 291, no. 5507, pp. 1304–1351, 2001.

[126] S. Ginsburg and H. F. Willard, "Genomic and personalized medicine: Foundations and applications," *Transl. Res.*, vol. 154, no. 6, pp. 277–287, 2009.

[127] M. Hoffman, C. Arnoldi, and I. Chuang, "The clinical bioinformatics ontology: A curated semantic network utilizing RefSeq information," in *Biocomputing*. Singapore: World Scientific, 2005, pp. 139–150.

[128] M. A. Hoffman, "The genome-enabled electronic medical record," *J. Biomed. Inform.*, vol. 40, no. 1, pp. 44–46, 2007.

[129] E. D. Green, M. S. Guyer, and N. H. Genome, "Charting a course for genomic medicine from base pairs to bedside," *Nature*, vol. 470, no. 7333, p. 204, 2011.

[130] J. L. Kannry and M. S. Williams, *Integration of Genomics into the Electronic Health Record: Mapping Terra Incognita*. London, U.K.: Nature Publishing Group, 2013.

[131] J. C. Denny *et al.*, "Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data," *Nature Biotechnol.*, vol. 31, no. 12, p. 1102, 2013.

[132] M. S. Islam, M. M. Hasan, X. Wang, and H. D. Germack, "A systematic review on healthcare analytics: Application and theoretical perspective of data mining," *Healthcare*, vol. 6, no. 2, p. 54, 2018.

[133] M. Herland, T. M. Khoshgoftaar, and R. Wald, "A review of data mining using big data in health informatics," *J. Big Data*, vol. 1, no. 1, p. 2, 2014.

[134] D. Tomar and S. Agarwal, "A survey on data mining approaches for healthcare," *Int. J. Bio-Sci. Bio-Technol.*, vol. 5, no. 5, pp. 241–266, 2013.

[135] Fang, S. Pouyanfar, Y. Yang, S.-C. Chen, and S. S. Iyengar, "Computational health informatics in the big data age: A survey," *ACM Comput. Surv.*, vol. 49, no. 1, p. 12, 2016.

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

[137] A. He, X. Jin, Z. Zhao, and T. Xiang, "A cloud computing solution for hospital information system," in *Proc. IEEE Int. Conf. Intell. Comput. Intell. Syst. (ICIS)*, vol. 2, Oct. 2010, pp. 517–520.

[138] S. Cha, A. Abusharekh, and S. S. Abidi, "Towards a'big'health data analytics platform," in *Proc. IEEE 1st Int. Conf. Big Data Comput. Service Appl. (BigDataService)*, Mar. 2015, pp. 233–241.

[139] Y. Shi, X. Liu, Y. Xu, and Z. Ji, "Semantic-based data integration model applied to heterogeneous medical information system," in *Proc. 2nd Int. Conf. Comput. Autom. Eng. (ICCAE)*, vol. 2, Feb. 2010, pp. 624–628.

[140] D. A. Davis, N. V. Chawla, N. A. Christakis, and A.-L. Barabási, "Time to CARE: A collaborative engine for practical disease prediction," *Data Mining Knowl. Discovery*, vol. 20, no. 3, pp. 388–415, 2010.

[141] M. Bittorf *et al.*, "Impala: A modern, open-source SQL engine for Hadoop," in *Proc. 7th Biennial Conf. Innov. Data Syst. Res.*, 2015, pp. 1–10.

[142] Accessed: Aug. 18, 2018. [Online]. Available: https://www.ohdsi.org/

[143] F. Versaci, L. Pireddu, and G. Zanetti, "Scalable genomics: From raw data to aligned reads on Apache YARN," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2016, pp. 1232–1241.

[144] A. Mukherjee, A. Pal, and P. Misra, "Data analytics in ubiquitous sensor-based health information systems," in *Proc. 6th Int. Conf. Next Gener. Mobile Appl., Services Technol. (NGMAST)*, Sep. 2012, pp. 193–198.
[145] K. Zolfaghar, N. Meadem, A. Teredesai, S. B. Roy, S.-C. Chin, and B. Muckian, "Big data solutions for predicting risk-of-readmission for congestive heart failure patients," in *Proc. IEEE Int. Conf. Big Data*, Oct. 2013, pp. 64–71.
[146] D. Lyubimov and A. Palumbo, *Apache Mahout: Beyond MapReduce*. Scotts Valley, CA, USA: CreateSpace Independent Publishing Platform, 2016.
[147] A. Thusoo *et al.*, "Hive: A warehousing solution over a map-reduce framework," *Proc. VLDB Endowment*, vol. 2, no. 2, pp. 1626–1629, 2009.
[148] M. Meystre, V. G. Deshmukh, and J. Mitchell, "A clinical use case to evaluate the i2b2 Hive: Predicting asthma exacerbations," in *Proc. AMIA Annu. Symp.*, 2009, p. 442.
[149] G. Wang and J. Tang, "The NoSQL principles and basic application of Cassandra model," in *Proc. Int. Conf. Comput. Sci. Service Syst. (CSSS)*, Aug. 2012, pp. 1332–1335.
[150] K. Banker, *MongoDB in Action*. Shelter Island, NY, USA: Manning Publications, 2011.
[151] A. Hospital *et al.*, "BIGNASim: A NoSQL database structure and analysis portal for nucleic acids simulation data," *Nucleic Acids Res.*, vol. 44, no. D1, pp. D272–D278, 2015.
[152] J. Kreps, N. Narkhede, and J. Rao, "Kafka: A distributed messaging system for log processing," in *Proc. NetDB*, 2011, pp. 1–7.
[153] A. Lawlor, R. Lynch, M. M. Aogáin, and P. Walsh, "Field of genes: Using Apache Kafka as a bioinformatic data repository," *GigaScience*, vol. 7, no. 4, p. giy036, 2018.
[154] N. T. Karonis *et al.*, "Distributed and hardware accelerated computing for clinical medical imaging using proton computed tomography (pCT)," *J. Parallel Distrib. Comput.*, vol. 73, no. 12, pp. 1605–1612, 2013.
[155] J.-Y. Cui, G. Pratx, S. Prevrhal, and C. S. Levin, "Fully 3D list-mode time-of-flight PET image reconstruction on GPUs using CUDA," *Med. Phys.*, vol. 38, no. 12, pp. 6775–6786, Dec. 2011.
[156] C. Men *et al.*, "GPU-based ultrafast IMRT plan optimization," *Phys. Med. Biol.*, vol. 54, no. 21, p. 6565, 2009.
[157] S. S. Samant, J. Xia, P. Muyan-Özçelik, and J. D. Owens, "High performance computing for deformable image registration: Towards a new paradigm in adaptive radiotherapy," *Med. Phys.*, vol. 35, no. 8, pp. 3546–3553, 2008.
[158] S. Borromeo, C. Rodriguez-Sanchez, F. Machado, J. A. Hernandez-Tamames, and R. de la Prieta, "A reconfigurable, wearable, wireless ECG system," in *Proc. 29th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBS)*, Aug. 2007, pp. 1659–1662.
[159] E. Lindholm, J. Nickolls, S. Oberman, and J. Montrym, "NVIDIA tesla: A unified graphics and computing architecture," *IEEE Micro*, vol. 28, no. 2, pp. 39–55, Apr. 2008.
[160] R. Buyya, *High Performance Cluster Computing: Architectures and Systems*, vol. 1. Upper Saddle River, NJ, USA: Prentice-Hall, 1999, p. 999.
[161] J. D. Owens, M. Houston, D. Luebke, S. Green, J. E. Stone, and J. C. Phillips, "GPU computing," *Proc. IEEE*, vol. 96, no. 5, pp. 879–899, May 2008.
[162] S. D. Brown, R. J. Francis, J. Rose, and Z. G. Vranesic, *Field-Programmable Gate Arrays*, vol. 180. Berlin, Germany: Springer, 2012.
[163] X. Fei, X. Li, and C. Shen, "Parallelized text classification algorithm for processing large scale TCM clinical data with MapReduce," in *Proc. IEEE Int. Conf. Inf. Autom.*, Aug. 2015, pp. 1983–1986.
[164] R. N. Boubela, K. Kalcher, W. Huf, C. Našel, and E. Moser, "Big data approaches for the analysis of large-scale fMRI data using apache spark and GPU processing: A demonstration on resting-state fMRI data from the human connectome project," *Frontiers Neurosci.*, vol. 9, p. 492, Jan. 2016.
[165] S. V. Kovalchuk, E. Krotov, P. A. Smirnov, D. A. Nasonov, and A. N. Yakovlev, "Distributed data-driven platform for urgent decision making in cardiological ambulance control," *Future Gener. Comput. Syst.*, vol. 79, pp. 144–154, Feb. 2018.
[166] J. M. Keller, M. R. Gray, and J. A. Givens, "A fuzzy k-nearest neighbor algorithm," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-15, no. 4, pp. 580–585, Jul./Aug. 1985.
[167] M. A. Hearst, S. T. Dumais, E. Osman, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intell. Syst. Appl.*, vol. 13, no. 4, pp. 18–28, Jul./Aug. 2008.
[168] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Netw.*, vol. 61, pp. 85–117, Jan. 2015.
[169] M. Swan, *Blockchain: Blueprint for a New Economy*. Newton, MA, USA: O'Reilly Media, 2015.
[170] A. J. Schaefer, M. D. Bailey, S. M. Shechter, and M. S. Roberts, "Modeling medical treatment using Markov decision processes," in *Operations Research and Health Care*. Boston, MA, USA: Springer, 2005, pp. 593–612.
[171] F. A. Sonnenberg and J. R. Beck, "Markov models in medical decision making: A practical guide," *Med. Decis. Making*, vol. 13, no. 4, pp. 322–338, 1993.

GASPARD HARERIMANA (S'13) received the B.S. degree in computer engineering from Ethiopian Defense University in 2008 and the M.S. degree in information technology from Carnegie Mellon University in 2015. He is currently pursuing the Ph.D. degree with the Department of Computer Science, Sangmyung University, Seoul, South Korea. He was a Staff and a Researcher with the Rwanda's Ministry of Defense, Kigali, Rwanda, and a Visiting Lecturer with the Adventist University of Central Africa, Kigali. He is a Research Assistant with the Department of Computer Science, Sangmyung University. His research interests involve computer networks, big data, and machine learning with emphasis on deep learning.

BEAKCHEOL JANG (M'17) received the B.S. degree in computer science from Yonsei University in 2001, the M.S. degree in computer science from the Korea Advanced Institute of Science and Technology in 2002, and the Ph.D. degree in computer science from North Carolina State University in 2009. He is currently an Associate Professor with the Department of Computer Science, Sangmyung University. His primary research interests are wireless networking with an emphasis on ad hoc networking, wireless local area networks, mobile network technologies, and machine learning. He is a member of ACM.

JONG WOOK KIM (M'17) received the Ph.D. degree from the Computer Science Department, Arizona State University, in 2009. He was a Software Engineer with the Query Optimization Group, Teradata, from 2010 to 2013. He is currently an Assistant Professor of computer science with Sangmyung University. His primary research interests are in the areas of data privacy, distributed databases, query optimization, and artificial intelligence. He is a member of ACM.

HUNG KOOK PARK received the B.A. degree in business administration from Seoul National University, South Korea, and the M.B.A. and Ph.D. degrees in management of information systems from the Claremont Graduate School, California. He is currently a Professor of computer science and the Head of the Center for Global Creation and Collaboration, Sangmyung University, Seoul, South Korea. He is also a Senior Advisor and a Specialist of the Enterprise Growth Program, EBRD, London, U.K., and a Consultant of the African Development Bank, Abidjan, Côte d'Ivoire. He is a member of the Joint Advisory Board, Carnegie Mellon University Africa, Kigali, Rwanda. His papers have appeared in the *European Journal of Operational Research* and the *Journal of International Information Management*. His research interests involve big data, machine learning, computer networks, and database.