International Journal of E-Health and Medical Communications
Volume 11 • Issue 2 • April-June 2020

Big Data Classification and Internet of Things in Healthcare

Amine Rghioui, Research Team in Smart Communications-ERSC–Research Centre E3S, EMI, Mohamed V University, Rabat, Morocco

Jaime Lloret, Integrated Management Coastal Research Institute, Universitat Politecnica de Valencia, 46370 Valencia, Spain

Abedlmajid Oumnad, Research Team in Smart Communications-ERSC–Research Centre E3S, EMI, Mohamed V University, Rabat, Morocco

ABSTRACT

Every single day, a massive amount of data is generated by different medical data sources. Processing this wealth of data is indeed a daunting task, and it forces us to adopt smart and scalable computational strategies, including machine intelligence, big data analytics, and data classification. The authors can use the Big Data analysis for effective decision making in healthcare domain using the existing machine learning algorithms with some modification to it. The fundamental purpose of this article is to summarize the role of Big Data analysis in healthcare, and to provide a comprehensive analysis of the various techniques involved in mining big data. This article provides an overview of Big Data, applicability of it in healthcare, some of the work in progress and a future works. Therefore, in this article, the use of machine learning techniques is proposed for real-time diabetic patient data analysis from IoT devices and gateways.

KEYWORDS

Big Data, Healthcare, Internet of Things, Machine Learning

1. INTRODUCTION

The Internet of Things (IoT) is a computing concept that describes a future where every day physical objects will be connected to the Internet and be able to identify themselves to other devices. This paper presents a review of literature on the subject of the IoT technologies and their applications domains and the futuristic research areas. Several research studies have addressed and developed this topic with detailed studies synthesis about the fields of application of internet of things, and general visions (Gubbi, Buyya, Marusic, & Palaniswami, 2013). Other papers summarize the applications of IoT in the healthcare industry and identify the intelligentization trend and directions of future research in this field (Yin, 2016).

Over the last two decades, we have seen an enormous amount of growth in data. The data has been doubling every two years since 2011. As a result of this technological revolution, big data is becoming an important issue in the sciences, governments, and enterprises increasingly. Big Data

DOI: 10.4018/IJEHMC.2020040102

is a data set, which is difficult to capture, store, filter, share, analyze and visualize on it with current technologies (Young, Min, Wenixa, & Depeng, 2015).

By understanding, processing and utilizing the knowledge and information hidden in Big Data concerning health issues and disease trends in certain population, we can find solutions, with which, we can live longer and healthier (Lloret, Parra, Taha, & Tomás, 2017, 2017). Big data analytics improve health care insights in many aspects shown in Figure 1.

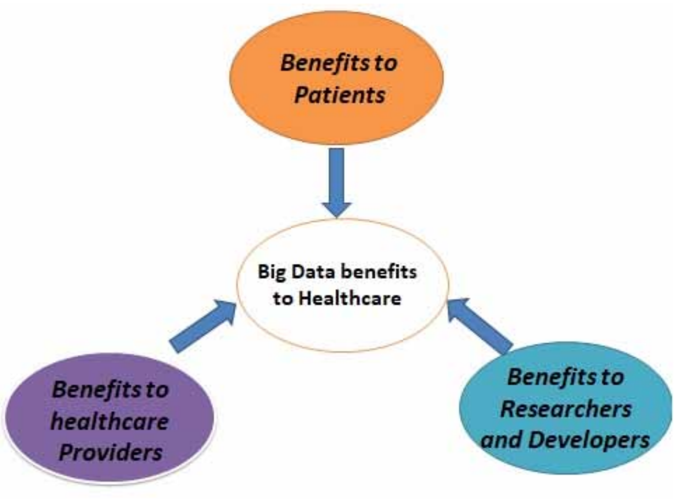
Benefits to Patients: Big Data in healthcare is being used to predict epidemics, cure disease, improve quality of life and avoid preventable deaths. With the world's population increasing and everyone living longer, models of treatment delivery are rapidly changing, and data are driving many of the decisions behind those changes. Big data can help patients make the right decision in a timely manner. From patient data, analytics can be applied to identify individuals that need "proactive care" or need a change in their lifestyle to avoid health condition degradation. For example, patients in early stages of some diseases (e.g., heart failure often caused by some risk factors such as hypertension or diabetes) should be able to benefit from preventive care thanks to Big data.

Benefits to Researchers and Developers (R & D): R & D contribute to new algorithms and tools, such as the algorithms by Google, Facebook, and Twitter that define what we find about the health system. Google, for instance, has applied algorithms of data mining and machine learning to detect influenza epidemics through search queries. R & D can also: Enhance predictive models to produce more devices and treatment for the market, and can give statistical tools and algorithms to improve the clinical trial design and patient recruitment to better match treatments to individual patients, thus reducing trial failures and speeding new treatments to market.

Benefits to healthcare Providers: can analyze disease patterns and tracking disease outbreaks and transmission to improve public health surveillance and speed response, Turning large amounts of data into actionable information that can be used to identify needs, provide services, and predict and prevent crises, Capture and analyze in real-time large volumes of fast-moving data from in-hospital and in-home devices, for safety monitoring and adverse event prediction, also providers can Apply advanced analytics to patient profiles to identify individuals who would benefit from proactive care or lifestyle changes, for example, those patients at risk of developing a specific disease (e.g., diabetes) who would benefit from preventive care (Sendra, Parra, Lloret, & Tomás, 2018).

The rising costs of health care and the increasing availability of new personal health devices are the ingredients of the vision of the IoT in the connected healthcare. The vision of connected healthcare

Figure 1. Benefits in healthcare


Benefits to Patients
Big Data benefits to Healthcare
Benefits to healthcare Providers
Benefits to Researchers and Developers

is growing because of the availability of new technological tools. By the application of the IoT and new technologies, it is possible to create a health application that appears every morning to request reading the level of glucose in the blood and collects data from the patient automatically (Bennett, Savaglio, Lu et al., 2014). In the vision of connected healthcare, patients are those who take control of their health and being in good physical and mental health due to this application. In addition, this leads to a good responsibility and control of heath by allowing a real scenario for the IoT in healthcare. IoT will help doctors to respond quickly in emergencies and allow them to cooperate with international hospitals to track the status of a patient. There are also other applications of IoT such as patient identification; this application aims to reduce adverse events for patients, maintenance of comprehensive electronic medical records (Rghioui, Sendra, Lloret, & Oumnad, 2016).

The contour of the contributions of this paper compared to other documents from the field survey, this survey provides a deeper summary of the Big Data in Healthcare, which allows us to know what the value of Big Data analytics in Healthcare is in details. We also present the different algorithms of data classification in healthcare fields.

The remainder of this article is structured as follows: Section II presents and explaining other surveys on big data in Health care. Section III contains concept of Big Data in healthcare, their technologies, architecture, and applications. The Big Data tools and platforms are discussed in Section IV, section V present the Benefits of Big Data to Healthcare. HealthCare Big Data Sources are described in Section VI, and Section VII explains Big Data Initiatives in Healthcare. We will then present our use case in Data Classification with Weka tool. Finally, some remarks conclude the paper.

2. RELATED WORK

In this part, we are going to show the main research areas considered by most surveys published in the field of the Internet of Things and Big Data in Health-care.

There are many related works in the literature about Big Data and IoT and their useful applications in many life aspects including healthcare. Not neglecting the important issue of classification data, a predictive Big Data analytics in Healthcare is proposed in (Reddy & Kumar, 2016). This paper gives an overview of storing and retrieval methods, Big Data tools, and techniques used in healthcare clouds.

There are several documents published survey covering different aspects of Big Data technology. For example, the survey by (Thara, Premasudha, Ram, & Suma, 2016) presents the review of various research efforts made in healthcare domain using Big Data concepts and methodologies. In (Wang, Kung, & Byrd, 2018), the authors examine the historical development, architectural design and component functionalities of big data analytics. In (Dogaru & Dumitrache, 2017), the authors present big data from the perspective of improving healthcare services and offers a holistic view of system security and factors determining security breaches. A baseline for assessing the rapid growth of the implementation of big data analytics into healthcare and life science aspects that assists in the understanding the big data applications and its impacts presented in (Zayeri, 2017). In (Moreira, Rodrigues, Furtado et al., 2018), the authors use of a machine learning technique, known as averaged one-dependence estimators, is proposed for real-time pregnancy data analysis from IoT devices and gateways. The authors (Sterling, 2017) provide an overview of big data analytics for healthcare.

The IoT big-data management and knowledge discovery is a key research challenge for the real-time industrial automation applications. Therefore, we study some existing system, models, or frameworks that are implemented in IoT big-data management and knowledge discovery perspective. The IoT big-data management includes several managerial activities such as data collection, integrations, cleaning, storage, processing, analysis, and visualizations that have been implemented through various systems, models, and frameworks (Zhou, Hu, Wang, Lu, & Zhao, 2013) (Mozumdar, Shahbazian, & Ton, 2014).

In addition, (Lee, 2017) share some successes in healthcare, defense, and service sector applications through innovation in predictive and big data analytics through the modeling and computational advances in integer programming. A framework to help organize and guide the understanding of the application of big data technologies for processing health and healthcare data presented in (Sheeran & Steele, 2017). Examination of the concept of big data in healthcare, its benefits, and attendant challenges and implementation of big data in healthcare presented in (Olaronke & Oluwaseun, 2016). A probabilistic data collection mechanism and the correlation analysis of those collected data is given by (Sahoo, Mohapatra, & Wu, 2016). The use of Artificial Neural Networks (ANN) model based on clinical and biochemical variables in patients with moderate to severe traumatic injury. Is presented in (Gholipour, Rahim, Fakhree, & Ziapour, 2015). In (Moudani, Hussein, AbdelRazzak, & Mora-Camino, 2014), a proficient methodology for the extraction of significant patterns from the Coronary Heart Disease warehouses for heart attack prediction has been presented, and they propose to validate the classification using Multi-classifier decision tree to identify the risky heart disease cases.

The work in (Gazal & Kaur, 2015) gives a brief research direction in data management and knowledge discovery prospective in IoT big-data management platform, in which three activities are mainly associated, that is, data association, inference, and knowledge discovery.

3. BIG DATA IN HEALTHCARE

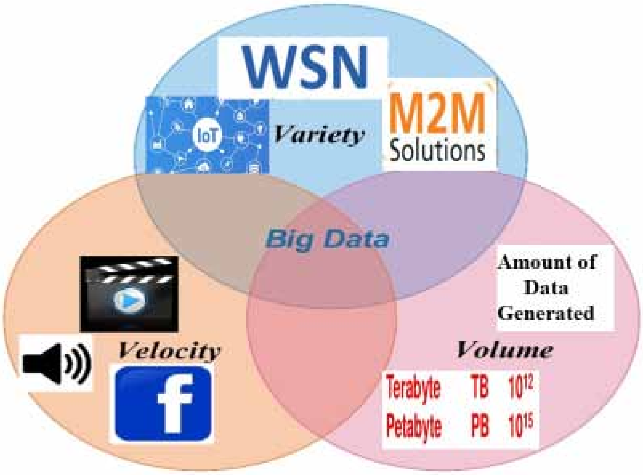
Doug Laney, an analyst of META (presently Gartner), defined challenges and opportunities brought about by increased data with a 3Vs model. (Gartner, 2012) (Figure 2).

In the "3Vs" model, Volume means, with the generation and collection of masses of data, data scale becomes increasingly big; Velocity means the timeliness of big data, specifically, data collection and analysis, must be rapidly and timely conducted; Variety indicates the various types of data, which include semi-structured and unstructured data such as audio, video, webpage, and text, as well as traditional structured data.

We explain the meanings of the 3 Vs that characterize Big Data and motivate their relevance to health care data:

- **Volume:** Health care data grows dramatically. Health care systems data require terabytes and petabytes. These systems include information such as personal information, radiology images, personal medical records, 3D imaging, genomics, and biometric sensor readings. Healthcare systems can now have the potential to manage and analyse this complex data structure. Thanks to

Figure 2. Big Data: The three V's



the use of cloud computing, manipulation, storage and use of such a complex data is now made possible. According to KPMG report, the volume of healthcare data reached 150 Exabyte's in 2013, and it is increasing at a prominent rate of 1.2 - 2-4 Exabyte's a year;
- **Variety:** Health care data sources and complexity, this dimension represents a big challenge because of the variety of data: structured, unstructured and semi-structured. Structured information, such as clinical data, are easy to manipulate, store and analyse by machine. However, most of health care data, such as office medical records, doctor notes, paper prescriptions, images, and radiograph films, are unstructured or semi-structured. The most challenging aspect of big data in health care consists-of-combining traditional data with new forms of data to get the closer to the right solution for a specific patient;
- **Velocity:** Big data analytics the information stored in health care systems is often correct, but not always even if it is updated on a regular basis. Thus, big data must be retrieved, analysed and compared to make time and accurate decisions based on real-time data processing. Life or death of patients can rely on real time data. Therefore, big data analytics must be done to prevent and detect infections as early as possible to make better decisions and consequently save lives.

Big data has advanced not only the size of data but also creating value from it. In other words, big data, that becomes a synonymous of data mining, business analytics and business intelligence, has made a big change in BI from reporting and decision to prediction of results. In healthcare, this value can be translated into understanding new diseases and therapies, predicting outcomes at earlier stages, making real-time decisions, promoting patients' health, enhancing medicine, reducing cost and improving healthcare value and quality (Stankovic, 2014).

3.1. Existing Technologies

The Big Data technologies involve commercial and open source platforms and services for storage, security, access and processing of data, many of them are based on the widely used open-source Hadoop framework. It is an open-source framework designed to deal with large-scale data using clusters of commodity hardware. It consists of a distributed storage component: Hadoop Distributed File System (HDFS) and a processing component: MapReduce programming model. Hadoop Distributed File System (HDFS) is a distributed file system and data engine designed to handle extremely high volumes of data in any structure. Hadoop is an independent Java-based programming framework that enhances the computation of large data sets in a distributed computing environment. Hadoop has two components (Narayan, Bailey, & Daga, 2012):

- Hadoop distributed file system
- Map Reduce

HDFS is a distributed file system that provides high-performance access to data distributed in Hadoop clusters. Like other technologies related to Hadoop, HDFS has become a key tool for managing Big Data pools and supporting analytic applications. HDFS is usually deployed on low-cost so-called convenience servers. Breakdowns are frequent. The file system is therefore designed to be extremely fault-tolerant, while facilitating fast data transfer between system nodes. When HDFS collects data, the system segments the information into several bricks and distributes them on several nodes of the cluster, which allows parallel processing. The file system copies each data brick several times and distributes the copies on each of the nodes, placing at least one copy on a separate server in the cluster. As a result, the data, stored on failed nodes, can be found elsewhere in the cluster. Treatment can continue despite the breakdown. HDFS is developed to support applications with large volumes of data, such as individual files whose quantity can be counted in terabytes (Ilakiyaa & Nalini, 2017).

MapReduce is one of the most adopted frameworks in the field of batch processing, is a programming model in which a MapReduce program can have two functions: the map and the reduction, which requires moving data across the nodes. The map and the reduction, each defining a mapping of one set of key-value pairs to another. MapReduce is an efficient solution for one-pass computation but when it comes to multi-pass computation, Map Reduce is inefficient due to the high latency of disk operations. MapReduce programs can be written in various languages; Java, Ruby, Python, and C++. These functions work the same for any data size, but the execution time depends on the data size and the cluster size. Increasing data size causes increased execution and increasing cluster size decreases the execution time (Shah, Shukla, & Pandey, 2016).

3.2. Big Data Architecture

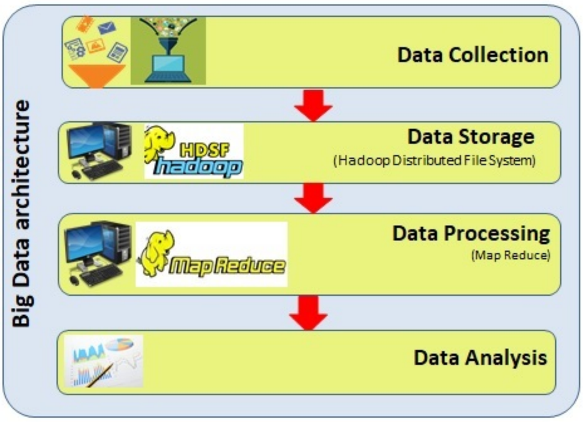
In this section, we present the proposed Big Data architecture as shown in Figure 3. Given the challenges of Big Data analysis, there is a clear need to aggregate and organize this data to facilitate analysis. To this end, the proposed architecture is designed to make this analysis simpler, allowing for focus on analytics rather than data management, storage and collection (Din, Ghayvat, Paul et al., 2015):

- **Data collection:** Data collection is to utilize special data collection techniques including data sensing using medical sensors, data acquisition, and data buffering.to acquire raw data from a specific data generation environment;
- **Data storage:** The storage system functions as a decision model in the proposed scheme, because the storage system checks whether the data is real-time data or offline data.

In the case of real-time data, the data is transmitted to a filtration system. Filtration requires a special algorithm that filters the data, so it is necessary to pre-process the data since the real-time data arrives quickly in the system. Therefore, the filtration system helps eliminate unwanted data. Apparently, if the data is offline data, the data is sent to the storage server. The storage server is used to store a massive volume of data. The storage server performs the following tasks:

- Provide server storage capabilities;
- Sharing the massive amount of data;
- Equal distribution of data between different processing servers;
- Queuing of processing efficiency improving data.

Figure 3. Big data architecture



- **Data processing:** Is a fundamental component that receives sequence files from the collection unit. It processes the data while performing necessary statistical calculations based on the nature of the data. For Big Data, to accomplish high efficiency, the overall data are disassembled into small pieces, and each of the pieces is separately processed in parallel using HDFS and MapReduce. Therefore, in order to enable effective data analysis, we shall pre-process data under many circumstances to integrate the data from different sources, which cannot only reduce storage expense but also improve analysis accuracy;
- **Data analysis:** Data analysis is the final and the most important phase in the value chain of big data, with the purpose of extracting useful values, providing suggestions or decisions. This layer delivers connectivity to the end-user to access various facilities, such as hospitals, emergency treatment. Furthermore, doctors can also monitor the patient by continuous analysis of his or her medical history. These services enable a doctor to connect to a facility to obtain a patient's present health status.

Different levels of potential values can be generated through the analysis of datasets in different fields.

4. BIG DATA PLATFORMS

There are thousands of Big Data tools out there, we have compiled a list of a few of data tools in the areas of extraction, storage, cleaning, mining, visualizing, analysing and integrating. Here are the top open source tools for big data.

Table 1 gives the different platforms and corresponding tools in handling the big data (import.io, 2015).

There are other tools used in large data, namely Hortonworks, hypertable, CouchDB, Grid Gain. It is used to improve the various factors in the development of large data and functionality of a computer system.

Below is comparison Table 2, consisting of various Big Data tools and the key features or facilities they support. We note that all Big Data tools are open source except Cloudera, as all these tools are scalable, so it differs just with the programming language and data storage format.

5. HEALTHCARE BIG DATA SOURCES

Healthcare big data is a revolutionary tool in the healthcare industry and is becoming vital in current patient-centric care. Diverse data sources have been aggregated into the healthcare big data ecosystem. These data sources are discussed below:

1. **Physiological data:** These data are huge in terms of volume and velocity:
   a. **Volume:** A variety of signals is collected from heterogeneous sources to monitor patient characteristics, including blood pressure, blood glucose, and heart rate;
   b. **Velocity:** The growth rate of data generation from continuous monitoring requires these data to be processed in real-time, for decision-making. Efficient and comprehensive methods are also required to analyze and process the collected signals to provide useable data to the healthcare professionals and other related stakeholders;
2. **EHRs:** Or electronic medical records (EMRs) are digitized structured healthcare data from a patient. The EHRs are collected from and shared among hospitals, and insurance companies. Security, integrity and privacy violations of these data can cause irremediable damage to the health, or even death, of the individual and loss to society. Thus, big healthcare data security is now a key topic of research;

Table 1. Platforms and tools for big data

Platform/Tool	Description
Hadoop Distributed File System (HDFS)	Hadoop has the potential to process extremely large amounts of data mainly by allocating partitioned data sets to numerous servers (nodes)
MapReduce	Originally developed by Google, a software framework for distributed processing of large data set on computing clusters of commodity hardware. MapReduce provides the interface for the distribution of sub-tasks and the gathering of outputs, the sequence of the name MapReduce implies, the reduce job is always performed after the map job.
Cassandra	Developed by Facebook, and built on Amazon Dynamo and Google BigTable, it's designed to handle large amounts of data across many commodity servers. It works on distributed servers where it requires reliable service and no failure.
MongoDB	A cross-platform document-oriented database that supports dynamic schema design. It's a NoSQL database with document-oriented storage and full index support. MongoDB can be used as a file system. It's good for managing data that changes frequently or data that is unstructured or semi-structured.
Zookeeper	Open source service for maintaining and configuration service for large distributed systems. ZooKeeper provides an infrastructure for cross-node synchronization and can be used by applications to ensure that tasks across the Hadoop cluster are serialized or synchronized. A ZooKeeper server is a machine that keeps a copy of the state of the entiresystem and preserves this information in local log files.
Hive	Hive was initially developed by Facebook but is now being used and developed by other companies like Netflix and Amazon. It is a query language it runs on Hadoop architecture, their creation, called Apache® Hive™, allows SQL developers to write Hive Query Language (HQL) statements that are similar to standard SQL statements, but it is built on top of Hadoop and MapReduce for providing data summarization, query, and analysis operations with several key differences.
Cloudera	Cloudera creates a commercial version of Hadoop with some additional services. They can help your business create a corporate data center so that people in your organization can access the data you store. Cloudera is mainly and a business solution to help companies manage their Hadoop ecosystem. Although Hadoop is a free and open source project for storing large amounts of data on inexpensive computing servers, the free version of Hadoop is not easy to use.
HBase	Traditional databases are row-oriented database management systems but HBase is a column-oriented. HBase provides random, real time access to your data in Hadoop. It is known for providing strong data consistency on reads and writes, which distinguishes it from other NoSQL databases. It combines the scalability of Hadoop by running on the Hadoop Distributed File System (HDFS), with real-time data access as a key/value store and deep analytic capabilities of Map Reduce.

3. **Medical images:** These images generate a huge volume of data to assist healthcare professionals for identifying or detecting disease. Medical imaging techniques such as X-ray scan play a crucial role in diagnosis. Owing to the complication, dimensionality and noise of the collected images, efficient image processing methods are required to provide clinically suitable data for patient care;
4. **Sensed data:** Collected from patients using different wearable or implantable devices. Sensed data must be collected, pre-processed, stored, shared and delivered correctly in a reasonable time to be of use to healthcare providers when making clinical decisions. It is a challenge to collect and collate multimodal sensed data from multiple sources at the same time;
5. **Clinical notes:** Its claims, recommendations, and decisions constitute one of the largest unstructured sources of healthcare big data. Owing to the variety in format, reliability, completeness, and accuracy of the clinical notes, it is challenging to ensure the health care provider has the correct information. Efficient data mining and natural language processing techniques are required to provide meaningful data.

Table 2. Big data tools comparison table

Big Data Tool	Open Source	Programming Language	Scalable	Data Storage Format
Hadoop Distributed File System (HDFS)	Yes	Java	Yes	Structured/Semi-Structured
MapReduce	Yes	Java/C#/C++	Yes	Structured
Cassandra	Yes	Java	Yes	Structured/Semi-Structured/ Unstructured
MongoDB	Yes	C++	Yes	semi-structured/ unstructured
Zookeeper	Yes	Java	Yes	Structured
Hive	Yes	SQL	Yes	Structured/Unstructured
Cloudera	No	Python	Yes	Semi-Structured
HBase	Yes	Many Language	Yes	Structured

6. BIG DATA INITIATIVES IN HEALTHCARE

There are several initiatives utilizing the potential of Big Data in healthcare. Some of the examples are listed below:

- **Asthmapolis:** Launched in 2010 to help find a solution by leveraging the advances in sensor technology (and the reduced costs of producing said sensors) and mobile data monitoring to help people manage their asthma more effectively, in turn reducing the costs both for those suffering from asthma and for the U.S. healthcare system itself.

When a patient is suffering from an asthma attack and is required to use his or her inhaler, the little device records the time and place that the inhaler was used and transmits this information to a web site. This data is then combined with information available through the Center for Disease Control (CDC).

- **Battling the Flu:** The Big Data analysis has become a weapon for the CDC to fight the flu, which claims millions of lives a year. Each week, the CDC receives over 700,000 reports on influenza. These reports include details about the disease, the treatment that was given and whether the treatment was unsuccessful. The CDC has made this information available to the general public under the name Flu View, an application that organizes and analyzes this vast amount of data to give doctors a clearer picture of the spread of the disease in near real-time. In addition to providing the precise location of patients who are struggling with the flu (Marjani, Nasaruddin, Gani et al., 2017);
- **GNS Healthcare and Aetna:** GNS Healthcare is a privately held data analytics company based in Cambridge. Has come together with the health insurance company Aetna to help combat people at risk or already with metabolic syndromes. Founded in 2000 by Cornell physicists Colin Hill and Iya Khalil, GNS Healthcare uses proprietary causal Bayesian network modeling and simulation software to analyze data for clients in the pharmaceutical, biotechnology, healthcare provider, health insurance, pharmacy benefit management and health informatics industries;
- **Diabetes and Big Data:** Diabetes is a major public health problem affecting more than 400 million people worldwide and causes 1.5 million deaths each year, according to the World Health Organization (WHO). With more than 9% of the adult population now living with diabetes, the recent WHO World Diabetes Report calls for more initiatives to improve the management and

treatment of this disease. Diabetes patients can also benefit from the Big Data revolution in health care. A company named Common Sensing has produced GoCap, a cap for prefilled insulin pens that can not only record the amount of insulin administered daily but also the specific times the dosages were administered. This information is then transmitted either to a mobile phone where an application records this data or to a connected glucometer. This data is then easily available to healthcare professionals and allows them to identify problems before they become severe and to tweak dosages if required;

- **USC Medical Monitor:** Computer scientists at the University of Southern California (USC) are teaming up with neurologists, kinesiologists and public health experts to fight against Parkinson's disease. The team uses various devices to track the movement of the patient and gather large amounts of data including data from 3D sensors of Microsoft Kinect, patient's smartphone, and additional body sensors. Then the data is fed into an algorithm that analyzes the data to identify any significant changes in movement and monitor disease progression and the effectiveness of treatments in real-time." The team hopes to extend this technology beyond Parkinson's disease in the future.

7. CLASSIFICATION DATA IN HEALTHCARE

Data classification is a process with many types of existing data sets for analysis, the development of classification technology has made great achievements. In general, we can classify them into two categories: one is the use of statistical principles, such as KNN, support vector machine, regression model, maximum entropy model Bayesian networks and other methods; Another is based on the principle of classifying certain rules, such as rough set theory, association rules, and decision trees and so on.

Diabetes comes from no communicable diseases (NCDs), and many people suffer from them. Nowadays, for developing countries like Morocco, diabetes has become a big health problem. Diabetes is one of the critical diseases that has associated long-term complications and also follows with various health problems. With the help of technology, it is necessary to build a system that stores and analyzes diabetes data and to predict possible risks accordingly. This work will be able to predict what types of diabetes are prevalent, future risks related and depending on the level of risk of the patient, the type of treatment can be provided.

There are six main classification models integrated in recent Weka tools; namely, decision tree, ripper rule, neural networks, naive Bayes, k-nearest neighbor and support vector machine:

1. Decision Tree (J48) is one of the tree classification techniques in which a particular tree will be generated as attributes, leaves as classes and edges as test results;
2. Ripper Rule (JRIP) is used to generate various rules by adding repetitive datasets until the rules cover all data configurations according to the set of learning data. In addition, once all the rules are generated, some of them will be merged to reduce the size;
3. Neural Networks (MLPs - Multilayer Perceptron) have a distinctive feature as a three-layer feed-forward neural network: an input layer, a hidden layer, and an output layer. In order to link each node in each level, it may include additional weight to properly adjust the traversing path selection process;
4. Naïve Bayes is derived from Bayes' theorem by applying the probabilistic learning knowledge for classification, assuming that the predictive attribute is conditionally independent according to each individual class;
5. k-Nearest-Neighbor (IBK) is used to perform the classification considering k subsets of data, each of them has similar characteristics by applying the Euclidean distance to understand the group, and here, IBK is the one of the k-Nearest-Simplified-Neighbor Classifiers;

6. Vector Support (Sequential Minimal Optimization (SMO)) is basically a linear classifier (two classes) used to determine the largest distance between two sets, and SMO is the minimum sequential optimization algorithm for SVM training using polynomial or Gaussian kernels.

7.1. Use Case

Continuous glucose monitors (CGMs) generate data streams that have the potential to revolutionize the possibilities of reducing extreme blood glucose levels (BGs) that characterize blood sugar levels in diabetes mellitus. These data, however, are both large and complex, and their analysis requires an understanding of the physical and physical, biochemical and mathematical principles involved in this new technology. According to the International Diabetes Federation (IDF), in 2015, 418 million people had Diabetes Mellitus (DM) in the world.

As shown in Figure 4, the Glucose sensor is used to sense the glucose values in the blood of the diabetic patient, and transfer the sensed data over short-range wireless communication to the patient's Android smartphone. The smartphone then aggregates and stores the sensed data, provides the healthcare monitoring interface to the patients for logging and also sends the physiological data to the medical server at a specified time interval whereby the physicians can directly have access for further analysis, diagnosis and intervention.

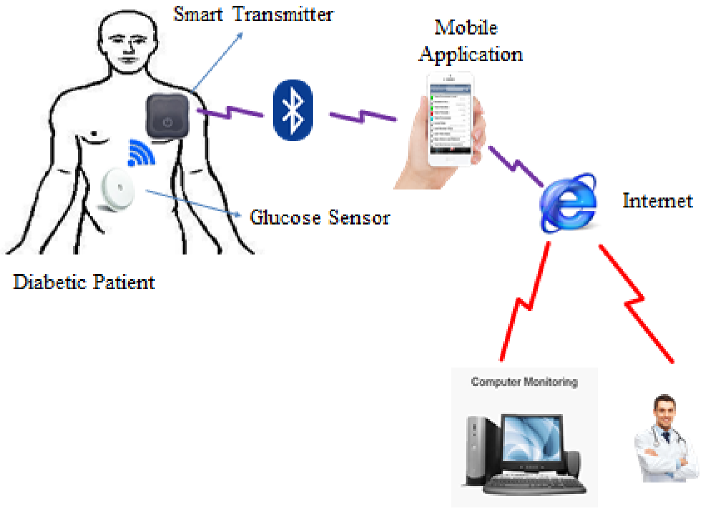
A glucose concentration value ≤70 mg/dl is defined as hypoglycemic and a glucose concentration value ≥180 mg/dl is defined as hyperglycemic. If ever the threshold set is reached, an alarm is triggered whereby the patient will receive a warning message on his mobile phone and similarly the physician will receive a warning message on the remote server.

5.2. Tool Used

WEKA 3.6.12 (Hall, Frank, Holmes et al., 2009) is used for analysis purpose; this tool provides a large range of classifiers like Bayes, Functions, Meta, and Tree-based. It can be also used for clustering and association of data. WEKA processes ".arff" files in its explorer to perform classification. It also provides different test options such as percentage split, cross-validation. and it also generates the graphical results like visualized classifiers errors, visualized margin curve.

The main limitation of this study is the data source is mostly here in Morocco, the national hospitals do not contain any database relating to the diabetic patient. Also, the challenges of the

Figure 4. Illustration of blood glucose measurement



health care industry is that uses information technology in a manner generally inferior to that of other industries, making it difficult to find these data. The dataset measured with CGM is used as data to test the individual and ensemble classifier. This study enrolled 20 diabetes men who were at least 45 years old with 3 measurements per day, for 20 days. The format of this dataset contains five columns designated by Date, Day, Glucose Level, and Request. The Glucose dataset is implemented and tested on four classifiers. Three out of four classifiers are individual classifiers, namely: the NaiveBayes classifier, the J48 classifier, ZeroR classifier and the BayesNet classifier.

These results prove that the used classifiers give good results of classification as well as attest the beneficial use of Trees Random Forest compared to Random Trees. The ROC curve is often used to determine the optimal threshold in classification problems. The area under the ROC curve (AUC) gives a good estimate of the system's rejection capability, Figure 5 and 6 show the ROC curve (%) for the True and False Classifiers.

The ROC which is receiver operating characteristic curve for the too high which indicate good performance of the techniques, therefore, it can be used for prediction, Figure 6 shows the comparison of False Classifiers Based on Area under ROC, PRC Area, and Precision of ROC Area, for J48, NaivBayes, RandomTree, and ZeroR algorithms.

When the dataset is tested on four classifiers, we obtain the precision and F-measures summarized in Table 3 using the "recall" approach. The accuracy of the NaiveBayes classifier is 88.78%, the accuracy of the J48 classifier is 99.21%, the accuracy of the BayesNet classifier is 93.48%, and the accuracy of the ZeroR classifier is 69.6078%.

The precision corresponds to the average success percentage for k iterations (Moreira, Rodrigues, Kumar et al., 2018). In practice, cross-validation with k equal to ten is the most commonly used method. This study also used the F-measure for indicating the imbalance among the classes. Table 4 compares the performance of the proposed method with similar works in literature. All these studies used the same database, but with different treatments for the data. The results show that for the leading indicator, namely the F-measure, the method proposed in this work provides excellent performance in comparison with other methods in literature. Regarding precision, the BayesNet algorithm has a performance that is very close to that of the J48 tree-based classifier.

Figure 5. Comparison of true classifiers based on area under ROC, PRC area, and precision

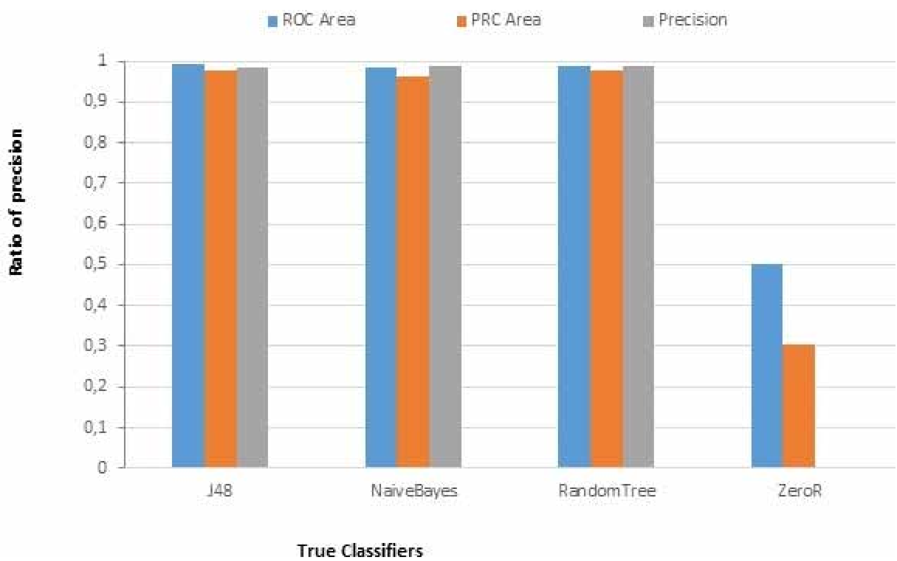


Figure 6. Comparison of false classifiers based on area under ROC, PRC area, and precision

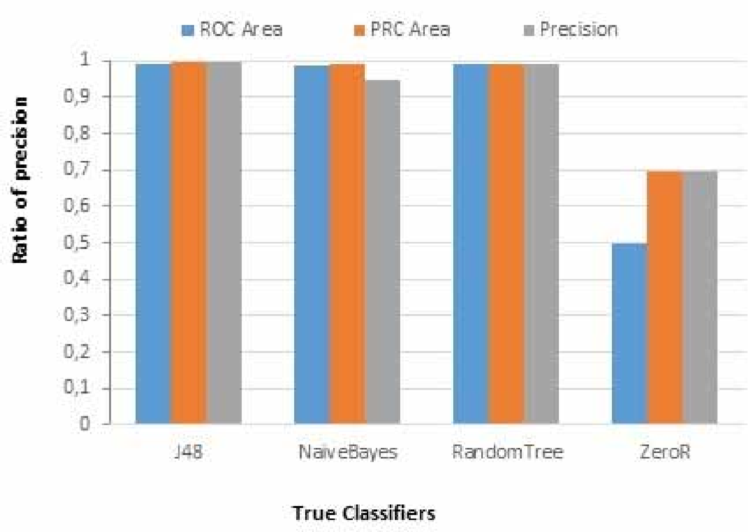


Table 3. Accuracy of classifications algorithms

Algorithms	Correctly Classified Instances	Incorrectly Classified Instances	F-Measure	Precision
Naïve Bayes	88.78%	11.22%	0,970	0,946
BayesNet	93.48%	6.52%	0,994	0,993
ZeroR	69.60%	30.40%	0,821	0,696
J48	99.21%	0.79%	0,994	0,992

Table 4. Precision and F-measure values in recent research using the diabetes database

Authors	Method	F-Measure	Precision
Habibi, Ahmadi, & Alizadeh, 2015	J48	0,705	0,717
Sa’di, Maleki, Hashemi et al., 2015	BayesNet	0,767	0,768
	J48	0,786	0,771
Rghioui, Sendra, Lloret, & Oumnad, 2016	BayesNet	0,994	0,993
	J48	0,994	0,992

The graphs representation in Figure7 shows the difference of correctly and incorrectly classified instances using four algorithms. The four algorithms are Naïve Bayes, BayesNet, ZeroR, and J48. We found that the J48 algorithm was the best as it had 99.21% correctly classified instances and only 0,79% were incorrectly classified instances.

Figure 8 shows the time graph of various classification algorithms. The longest time is taken by ZeroR consuming a time of 0.05 seconds and the shortest time is taken by RandomTree consuming 0.01 seconds only.

Figure 7. The graph of correctly and incorrectly classified instances of algorithms

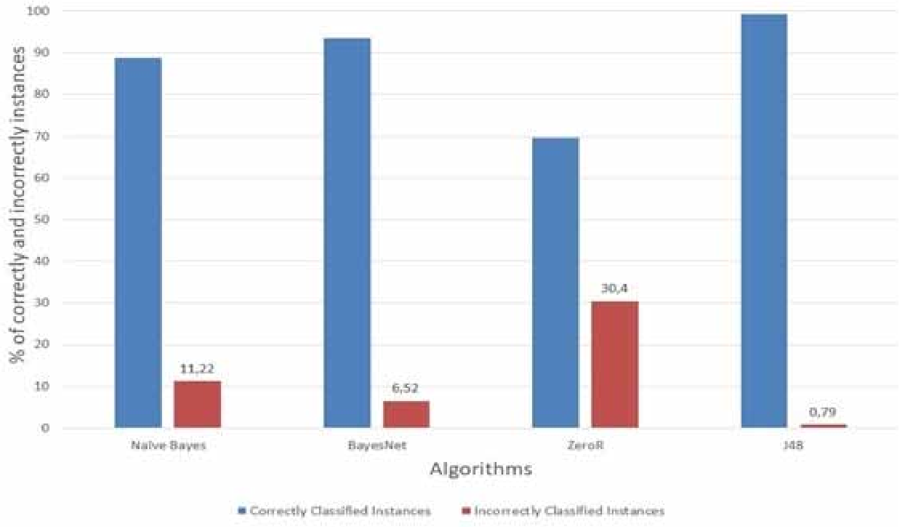


Figure 8. The graph for training time for top four classifiers

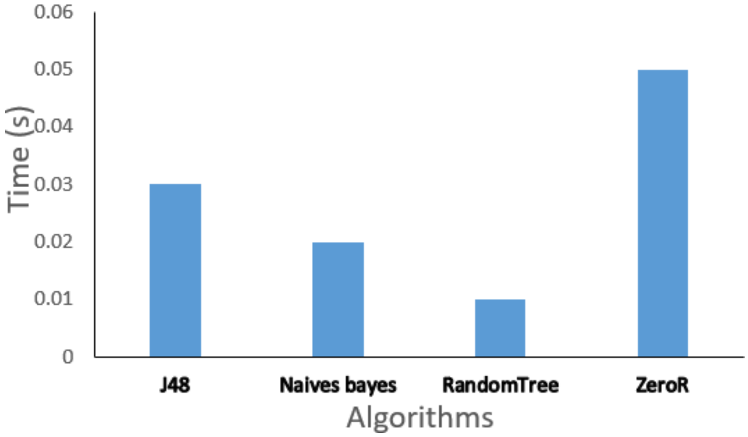
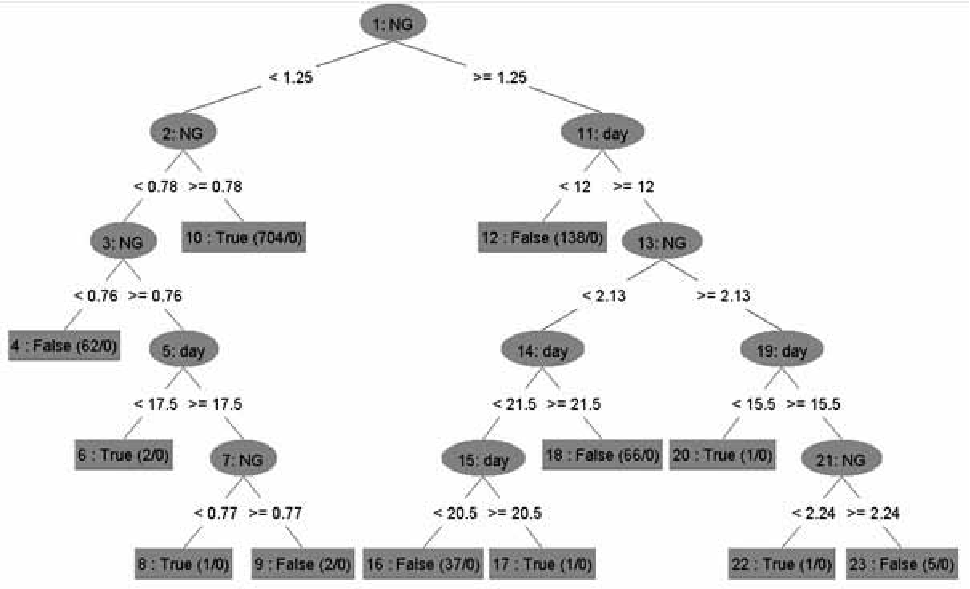


Figure 9 shows an example of Weka's visualization of a decision tree regarding the seed-size attribute from the Glucose Level.arff file. Once the information has been classified with the Weka algorithm J4.8, the decision tree shown in Figure 9 was obtained at the output.

6. CONCLUSION

Today the healthcare industry is just beginning to understand all the innovative things that can be done with Big Data. The intersection of data from multiple sources, tools, and technologies will promote informative extrapolations of Big Data, allowing the information to generate new and innovative solutions to healthcare. While Big Data technologies are improving day by day this also means that, the volume of data along with the rate at which data is flowing into enterprises today is increasing. The healthcare system today is on a trajectory that is unsustainable. In this paper, classification techniques are used for prediction on the dataset of patient's data, to analyze overall diabetic performance and

Figure 9. Visualize tree with random tree



predict some relative's disease. In this study, among all data, mining classifiers J48 performs best with 99.21% accuracy and therefore J48 proves to be potentially effective and efficient classifier algorithm. The main contribution of this study is that it provides the possibility of handling a large amount of data to find useful results that support health experts in the decision-making process. In the future, some new factors can be applied to improve the patient's performance and results can be obtained based on station data. In addition, more data mining techniques such as k-means, clustering algorithms and other classification algorithms can be applied to this data. Thus, a focus in the future should be on preventive care as well as population health management and overall wellness. With Big Data, health management of a population can be understood better.

REFERENCES

Bennett, T. R., Savaglio, C., Lu, D., Massey, H., Wang, X., Wu, J., & Jafari, R. (2014, August). Motionsynthesis toolset (most): a toolset for human motion data synthesis and validation. In *Proceedings of the 4th ACM MobiHoc workshop on Pervasive wireless healthcare* (pp. 25-30). ACM.

Dogaru, D. I., & Dumitrache, I. (2017, June). Holistic perspective of big data in healthcare. In *Proceedings of the 2017 E-Health and Bioengineering Conference (EHB)* (pp. 418-421). IEEE.

Gartner. (2012). 3D Data management controlling data volume velocity and variety. Retrieved from https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Gazal, & Kaur, P.D. (2015). A survey on big data storage strategies. In *Proceedings of the 2015 International Conference on Green Computing and Internet of Things (ICGCIoT)* (pp. 280-284). IEEE.

Din, SGhayvat, H., Paul, A., Ahmad, A., Rathore, M. M., & Shafi, I. (2015). An architecture to analyze big data in the Internet of Things. *9th International Conference on Sensing Technology (ICST).*

Gholipour, C., Rahim, F., Fakhree, A., & Ziapour, B. (2015). Using an artificial neural networks (ANNs) model for prediction of intensive care unit (ICU) outcome and length of stay at hospital in traumatic patients. *Journal of Clinical and Diagnostic Research: JCDR*, *9*(4), OC19.

Gubbi, J., Buyya, R., Marusic, S., & Palaniswami, M. (2013). Internet of Things (IoT): A vision, architectural elements, and future directions. *Future Generation Computer Systems*, *29*(7), 1645–1660.

IHTT. (2013). *Transforming Health Care through Big Data Strategies for leveraging big data in the health care industry.*

Ilakiyaa, R. N., & Nalini, N. J. (2017). Supervised Learning Based HDFS Replication Management System. In *Proceedings of the International Conference on Technical Advancements in Computers and Communications (ICTACC)* (pp. 116-120). Academic Press. doi:10.1109/ICTACC.2017.38

import.io. (2015). All the best big data tools and how to use them. Retrieved from https://www.import.io/post/all-the-best-big-data-tools-and-how-to-use-them

Lee, E. K. (2017, December). Innovation in big data analytics: Applications of mathematical programming in medicine and healthcare. In *Proceedings of the 2017 IEEE International Conference on Big Data (Big Data)* (pp. 3586-3595). IEEE.

Liu, J., Li, Y., Chen, M., Dong, W., & Jin, D. (2015). Software-defined internet of things for smart urban sensing. *IEEE Communications Magazine*, *53*(9), 55–63.

Lloret, J., Parra, L., Taha, M., & Tomás, J. (2017). An architecture and protocol for smart continuous eHealth monitoring using 5G. *Computer Networks*, *129*, 340–351.

Marjani, M., Nasaruddin, F., Gani, A., Karim, A., Hashem, I. A. T., Siddiqa, A., & Yaqoob, I. (2017). Big IoT data analytics: Architecture, opportunities, and open research challenges. *IEEE Access: Practical Innovations, Open Solutions*, *5*, 5247–5261.

Moreira, M. W., Rodrigues, J. J., Furtado, V., Kumar, N., & Korotaev, V. V. (2018). Averaged one-dependence estimators on edge devices for smart pregnancy data analysis. *Computers & Electrical Engineering*.

Moreira, M. W., Rodrigues, J. J., Kumar, N., Al-Muhtadi, J., & Korotaev, V. (2018). Evolutionary radial basis function network for gestational diabetes data analytics. *Journal of Computational Science*, *27*, 410–417.

Moudani, W., Hussein, M., AbdelRazzak, M. & Mora-Camino, F. (2014). Heart disease diagnosis using fuzzy supervised learning based on dynamic reduced features. International Journal of E-Health and Medical Communications, 5(3), 78-101. doi:10.4018/ijehmc.2014070106

Mozumdar, M., Shahbazian, A., & Ton, Q. (2014). A big data correlation orchestrator for Internet of Things. In *Proceeding of the IEEE World Forum on Internet of Things (WF-IoT '14)*, 304–308. doi:10.1109/WF-IoT.2014.6803177

Narayan, S., Bailey, S., & Daga, A. (2012, November). Hadoop acceleration in an openflow-based cluster. In *Proceedings of the 2012 SC Companion: High Performance Computing, Networking Storage and Analysis* (pp. 535-538). IEEE.

Olaronke, I., & Oluwaseun, O. (2016). Big Data in Healthcare: Prospects, challenges and resolutions. In *Proceedings of the Future Technologies Conference (FTC)*. Academic Press. doi:10.1109/FTC.2016.7821747

Reddy, A. R., & Kumar, P. S. (2016, February). Predictive big data analytics in healthcare. In *Proceedings of the 2016 Second International Conference on Computational Intelligence & Communication Technology (CICT)* (pp. 623-626). IEEE.

Rghioui, A., Sendra, S., Lloret, J., & Oumnad, A. (2016). Internet of things for measuring human activities in ambient assisted living and e-health. *Network Protocols and Algorithms*, *8*(3), 15–28.

Sa'di, S., Maleki, A., Hashemi, R., Panbechi, Z., & Chalabi, K. (2015). Comparison of data mining algorithms in the diagnosis of type II diabetes. *International Journal on Computational Science & Applications*, *5*(5), 1–12.

Sahoo, P. K., Mohapatra, S. K., & Wu, S. L. (2016). Analyzing Healthcare Big Data with prediction for future Health condition. *IEEE Access*, *4*, 9786–9799. doi:10.1109/ACCESS.2016.2647619

Sendra, S., Parra, L., Lloret, J., & Tomás, J. (2018). Smart system for children's chronic illness monitoring. Information Fusion, 40, 76-86.

Shah, M., Shukla, P. K., & Pandey, R. (2016). Phase level energy aware map reduce scheduling for big data applications. In *Proceedings of the International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES)*, pp. 532-535. doi:10.1109/SCOPES.2016.7955884

Sheeran, M., & Steele, R. (2017). A framework for big data technology in health and healthcare. *IEEE 8th Annual Conference Ubiquitous Computing, Electronics and Mobile Communication (UEMCON).*

Stankovic, A. J. (2014). Research directions for the internet of things. *IEEE Internet of Things Journal*, *1*(1), 3–9.

Sterling, M. (2017, October). Situated big data and big data analytics for healthcare. In *Proceedings of the 2017 IEEE Global Humanitarian Technology Conference (GHTC)*. IEEE.

Thara, D. K., Premasudha, B. G., Ram, V. R., & Suma, R. (2016, December). Impact of big data in healthcare: A survey. In *Proceedings of the 2016 2nd International Conference on Contemporary Computing and Informatics (IC3I)* (pp. 729-735). IEEE.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, *11*(1), 10–18. doi:10.1145/1656274.1656278

Habibi, S., Ahmadi, M., & Alizadeh, S. (2015). Type 2 diabetes mellitus screening and risk factors using decision tree: Results of data mining. *Global Journal of Health Science*, 7(5), 304–310. PMID:26156928

Wang, Y., Kung, L., & Byrd, T. A. (2018). Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. *Technological Forecasting and Social Change*, *126*, 3–13.

Yuehong, Y. I. N., Zeng, Y., Chen, X., & Fan, Y. (2016). The internet of things in healthcare: An overview. *Journal of Industrial Information Integration*, *1*, 3–13.

Zaveri, C. (2017, February). Use of big-data in healthcare and lifescience using hadoop technologies. In *Proceedings of the 2017 Second International Conference on Electrical, Computer and Communication Technologies (ICECCT)* (pp. 1-5). IEEE.

Zhou, J., Hu, L., Wang, F., Lu, H., & Zhao, K. (2013). An efficient multidimensional fusion algorithm for IoT data based on partitioning. *Tsinghua Science and Technology*, *18*(4), 369–378.

Amine Rghioui was born in Morocco, in 1989. He received a B.S. degree from Faculty of Sciences, Fez, Morocco in 2012 and M.S from Faculty of Sciences, Kenitra, Morocco in 2014. He is currently a member of the Electronic and Communication laboratory at Mohammadia School of Engineering. His research interests include wireless sensor network, Internet of Things and connected objects, Big Data.

Jaime Lloret (jlloret@dcom.upv.es) received his B.Sc.+M.Sc. in Physics in 1997, his B.Sc.+M.Sc. in electronic Engineering in 2003 and his Ph.D. in telecommunication engineering (Dr. Ing.) in 2006. He is a Cisco Certified Network Professional Instructor. He worked as a network designer and administrator in several enterprises. He is currently Associate Professor in the Polytechnic University of Valencia. He is the Chair of the Integrated Management Coastal Research Institute (IGIC) and he is the head of the "Active and collaborative techniques and use of technologic resources in the education (EITACURTE)" Innovation Group. He is the director of the University Diploma "Redes y Comunicaciones de Ordenadores" and he has been the director of the University Master "Digital Post Production" for the term 2012-2016. He was Vice-chair for the Europe/Africa Region of Cognitive Networks Technical Committee (IEEE Communications Society) for the term 2010-2012 and Vice-chair of the Internet Technical Committee (IEEE Communications Society and Internet society) for the term 2011-2013. He has been Internet Technical Committee chair (IEEE Communications Society and Internet society) for the term 2013-2015. He has authored 22 book chapters and has more than 450 research papers published in national and international conferences, international journals (more than 200 with ISI Thomson JCR). He has been the co-editor of 40 conference proceedings and guest editor of several international books and journals. He is editor-in-chief of the "Ad Hoc and Sensor Wireless Networks" (with ISI Thomson Impact Factor), the international journal "Networks Protocols and Algorithms", and the International Journal of Multimedia Communications. Moreover, he is Associate Editor-in-Chief of "Sensors" in the Section sensor Networks, he is advisory board member of the "International Journal of Distributed Sensor Networks" (both with ISI Thomson Impact factor), and he is IARIA Journals Board Chair (8 Journals). Moreover, he is (or has been) associate editor of 46 international journals (16 of them with ISI Thomson Impact Factor). He has been involved in more than 450 Program committees of international conferences, and more than 150 organization and steering committees. He has led many local, regional, national and European projects. He is currently the chair of the Working Group of the Standard IEEE 1907.1. He has been general chair (or co-chair) of 40 International workshops and conferences. He is IEEE Senior, ACM Senior and IARIA Fellow.

Abdelmajid Oumnad Professor in the Department of Electrical Engineering at Mohammadia School of engineering (EMI) and member of the Laboratory of Electronics and Communication (EMI). He received his Ph.D. in electronics of the Claude Bernard University in Lyon, He is a professor in Electronic and Communication Technology at the Department of Electrical Engineering in Mohammadia School of engineering in Rabat.