



A Conception of a Predictive Analytics Platform in Healthcare Sector by Using Data Mining Techniques and Hadoop

Basma BOUKENZE*

CNMM laboratory

University, Settat, Morocco

Hajar MOUSANNIF

LISI Laboratory, FSSM Cadi Ayyad

University, Marrakesh, Morocco

Abdelkrim HAQIQ

CNMM laboratory

University, Settat, Morocco

Abstract— *The emergence of new information technology such as mobile application, cloud computing, big data analytics impacted all sectors. This is particularly true for the healthcare system as an important sector, Nowadays Healthcare industry depends mainly on Information technology to provide best services. And that promise the healthcare area a big change especially in front of the explosion of medical data sources: data analysis and the extraction of knowledge from medical data using data mining techniques. We can say that a big-data revolution is under way in health care and Start with the vastly increased supply of health data. And that push us to apply these new technologies to get off their advantages and improve the medical sector. This paper will show the importance of applying predictive analytics techniques in medical platforms , and give Architecture design which combines big data analysis , data-mining and the mobile healthcare for self-monitoring. This system will be able to exploit the healthcare data through an intelligent process analysis and big data processing; in order to extract useful knowledge to helping in decision making and ensure a medical monitoring in real-time.*

Keywords— *a platform of healthcare data analysis, healthcare data, big data analytics; data mining, predictive analytics.*

I. INTRODUCTION

Today digitized information is omnipresent everywhere because Data is growing and moving faster than healthcare organizations can consume it. This is due mainly to the efforts of researchers in the medical field and their discoveries take as an example human DNA. Widespread use of the electronic medical records wish totally transforms medical care [7]. the latest innovations concerning genetics and smart home or smart places enables patient self-monitoring and treatment by using simpler devices[15]. The appearance of sensing technology like M-health [11]; healthcare data with all of this becomes voluminous and appears like a digital flood creating puddles and lakes, creeks and torrents, of data: numbers, words, voices, images, video; and this increases in parallel with the rapid growth in the use of mobile devices like smartphones, laptops, tablets, personal sensors that generating a data deluge.

Large data volumes at high velocities were originally an option that characterizes supercomputers, nuclear physics, military simulations and space travel. Late in the 20th century, bigger and faster data appeared in airline and bank operations, particularly with the growth of credit cards. Starting in 1990, The Human Genome Project was the launch of Big Data in healthcare [21], and this was due to a statistic that showed that 80% of medical data is unstructured and is clinically relevant and much significant. This data resides in multiple places like individual EMRs, lab and imaging systems, physician notes, medical correspondence, claims, CRM systems, and finance. For that a data-intensive research effort that pushed the limits of available data processing technology.

The potential of Big Data analytics allows us to hope to slow the ever-increasing costs of care, help providers to practice more effective medicine, empower patients and healthcare providers, support fitness and preventive self-care, and to dream about more personalized and predictive medicine. Yet, as with the Internet, social media, cloud computing, and using the intelligent procedure for managing analyzing and extracting information from Data; all that will transform healthcare system and gives the power to explore , predict and why not anticipate the cure. Big-data analysis promises and affirms that future is no longer mysterious.

As follows we will discuss the great role played by new technology in the field of health like healthcare analysis, and then we will present our proposed architecture and its contribution in healthcare sector.

The rest of this paper will present as follow: in section II, we present related works concerning technologies applied in the healthcare system and research's work in this field. Section III is reserved for description of our proposed platform. And the last section gives conclusions and perspectives.

II. RELATED WORKS

If we talk about Cloud computing as new technology applied in the healthcare system ,it brings many benefits ; by creating a network of doctors; patients and healthcare institutes and facilitates access to medical information anywhere and anytime [1],

Cloud computing provides health care much-appreciated services concerning data handling by ensuring [2,3] :

- Resiliency: cloud service providers offer platforms with a very powerful infrastructure that provides redundancy and storage of any data quantity to ensuring high availability anytime and anywhere.
- Privacy: cloud computing infrastructure ensures a high level of security than local IT department in a hospital can ensure.
- Speed of innovation: everything is handled in the cloud data redundancy and the update. By cloud provider don't need doing updates or installing the certificates or repairing blocking systems
- Mobile applications: while the mobile applications used are stored in the cloud and the data is also stored in the cloud; the communication will be done in an easier and more flexible way given that the facility of access will be the same to one patient or several at the same time .
- Developing trend: cloud adapts to all situations to ensure ease of access at a high level.

A lot of researchers are focused in this regard [14,10] and it cited the big role played by cloud computing in the stage of managing healthcare data that are becoming increasingly large. More than, that some of them give design of a cloud computing-based Healthcare SaaS Platform (HSP) to deliver healthcare information services with low cost, high clinical value and high usability with a high level of security [8, 6].

Big data analysis especially in healthcare area has been considered as a revolutionary approach to improving the quality of healthcare service [4,9], because analytics figures to play a pivotal role in the future of healthcare system and as a result of research to develop healthcare sector [21,22] systems found obliged to receive a new form of data such as human DNA , data genetics; hence the necessity of leveraging all these resources and embitter human health. Analytics can also be applied in healthcare to compare the cost and effectiveness of interventions, treatments, public health policies, or medical devices to reduce failed investments.

In fact, this kind of analysis could give the best solution to prevent medical disasters. For example, infectious diseases could be predicted by data healthcare analysis and the health authority could manage this situation and save the human.

Will soon be awash in genomic data [5,24]. Given the incredible size and dimensionality of these datasets, the field of analytics will need to borrow techniques to make it useful.

In addition to that , some predictive analytics platforms for disease targets across varying patient cohorts using electronic health records (EHRs) are created to facilitate specific biomedical research workflows, such as refinement of hypotheses or data semantics [13].

About tools used in that The most important platform for big data analytics platform is the open-source distributed data processing platform Hadoop (Apache platform) [20]. It belongs to the class of technologies "NoSQL" that have evolved to managing data at high volume. Hadoop has the potential to process extremely large amounts of data mainly by allocating partitioned data sets to numerous servers (nodes), each of which solves different parts of the larger problem and then integrates them for the final result [29-30]. Hadoop can serve both roles of organizing and data analyzing tool .Hadoop can handle very large volumes of data with different structures or no structure at all. But Hadoop is a little difficult to install, configure and manage, and people with Hadoop skills are not easily found. In addition, for these reasons, it appears organizations are not quite ready to embrace Hadoop completely.

knowing that the adoption of EHRs and electronics data, prepared a submitted base for applying analysis and become the norm in healthcare, it enables the building of predictive analytic solutions. These predictive models, as we know have the potential to lower cost and improve the overall health of the population. As predictive models become more pervasive, some standards appears to be used by all the parties involved in the modeling process: like The Predictive Model Markup Language (PMML) [19].It allows for predictive solutions to be easily shared between applications and systems. and it can be used to expedite the adoption and use of predictive solutions in the healthcare industry.

According to our research, we found that there are many efforts to creating platforms based on cloud computing for managing medical records and simplify access to data. The patient does not care about the way with his doctor manages his medical data. But which is more important to him is what is the positive impact of this on his health situation on one hand , and on the other hand and become involved in the treatment process.

What we propose is a platform that combines the benefits of mobile healthcare and big data analysis. Making as the primary objective _after the data analysis_ the exploration and extraction of information. And monitoring in real time patients and include patient as an active player.

III. INTELLIGENT HEALTHCARE SYSTEM ARCHITECTURE

A. System Characterstic

Our proposition is a conception of an intelligent platform wich we called Intelligent Healthcare System (IHCS) that will make analysis of big healthcare data with a quick way and in a real time, data wish coming from various sources and concerns patients, disease (risk factor), treatments, and doctors, after this analysis the system give predicted information that reflects the patient's situation in the future.

1. The system will be hosted in a cloud and can be accessed anytime and anywhere,
2. The system will make a quick analysis in real-time to give accurate future information using intelligent and very specific tools,
3. The doctor can access to the application for consulting and monitoring the health status, also make action if necessary,

4. The patient will still in contact with the intelligent system through an application that will be installed on his Smartphone or on a smart watch with a sensor for detecting each change of his health status (Fig1)

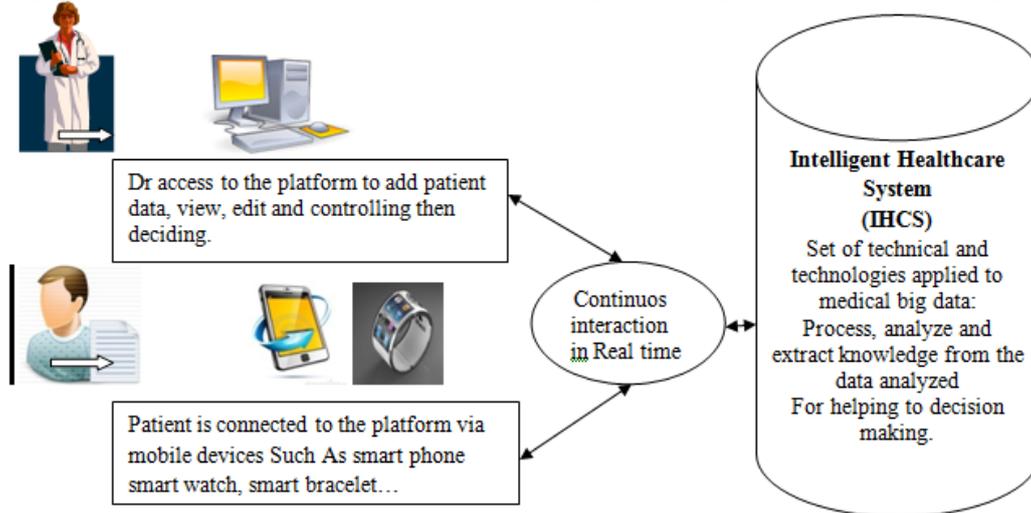


Fig 1. Typical intelligent Healthcare system schema

B. Intelligent healthcare System architecture

IHCS and throughout the medical data will capable to :

- ✓ Analyze a large amount of medical data
- ✓ predict what the patient may have in the future as complexity and pathologies by data mining technics
- ✓ Anticipate the cure and treatment,
- ✓ Monitoring patient in real time ,
- ✓ The patient will have the opportunity to make a self-monitoring in real-time by the use of health mobile devices.

And to ensure that we build our proposed system an architecture that combines several steps:

- ✓ **Data collection:** is the most important and sensitive phase and because the data is the main element and the pivot of the system. We must mention that more data is accurate more the accuracy of the predicted information increases.

The voluminous medical data can be coming from various Electronic Health Record (EHR) / Patient Health Record (PHR), Clinical systems and external sources like government sources, laboratories, pharmacies, insurance companies etc, in various formats (flat files, .csv, tables, ASCII/text, etc.) .

- ✓ **Data Warehousing:** In this phase, massive data coming from various sources warehoused to be cleansed, accumulated and made ready for further processing.

Big data analysis: it is a very important phase seen it demands very powerful techniques and tools to manage and process the voluminous data.

- ✓ **Predictive analysis** is the master step in all this process because it rests on the exploration of analyzed data to extract useful knowledge on the basis of data mining tools and algorithms to find links between the medical data.

- ✓ **Processing analyzed reports:** The results obtained after the predictive analysis process are exploited by:

- the doctor for help in decision making and giving a general view of the patient's status
- the patient will have the results of this process by his doctor but he is always in interaction with the system by a mobile device that he owned.

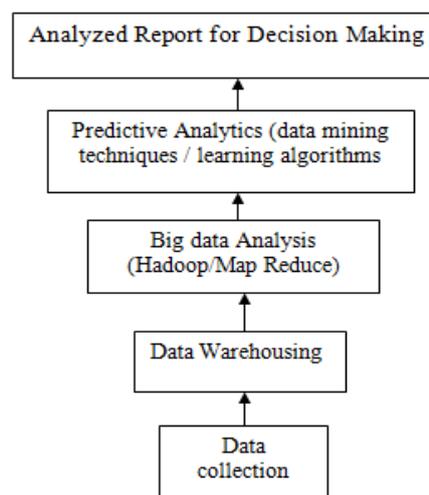


Fig 2. Architecture of the predictive analysis system-Health Care Application

In the first layer Hadoop is used as an open source framework designed to perform processing on massive medical data, The operating principle is as follows, the infrastructure applies the well-known principle of grid computing, of dividing the execution of a process on multiple nodes or clusters of servers.

In Hadoop architecture logic, this list is divided into several parts, each part being stored on a different server cluster. Instead of lean processing in a single cluster, as is the case for traditional architecture, the distribution of information helps distribute the processing across all compute nodes on which the list is distributed.

To implement such a technical process, Hadoop is coupled to a file system called HDFS (Hadoop Distributed File System for). It manages the allocation of storage of user data in blocks of information on different nodes. HDFS was inspired by a technology used by Google to own these cloud services, and known as Google File System (GFS).

Map/Reduce: the distribution and management of the calculations is carried out by Map Reduce. This technology combines two types of function:

- ✓ The Map function: which resides on the master node and then divides the input data or task into smaller subtasks, which it then distributes to worker nodes that process the smaller tasks and pass the answers back to the master node .The subtasks are run in parallel on multiple computers.
- ✓ The Reduce function: collects the results of all the subtasks and combines them to produce an aggregated Final result — which it returns as the answer to the original big query.

The second layer is characterized by the great role of Map-Reduce module for the process of predictive analysis. And to reinforce more and more the system in matters of prediction, it must be equipped by a powerful predictive algorithm or learning algorithm to ensure the important phases of the process and build a suitable model of prediction.

Data mining technology like a delicate process , executed by predictive algorithms, which have shown a strong effectiveness and efficiency in predicting , take as an example supporting victor machine (SVM) [31], decision tree(C4.5) [32] , and Naive Bayes (NB) [33], as They Are Currently classified Among the top 10 classification methods Identified by IEEE Python & Related Resources [34].

For that our system should be equipped with a learning algorithms among the cited ones or a combination of several learning algorithms to benefit from its performances and build a powerful hybrid algorithm that will be apply to all types of medical prediction.

IV. CONCLUSION

Big data analytics in healthcare provides all healthcare delivery system advantages such as explorations of data and knowledge extraction, economical cost reduction and push medical care to the better, but to give the power to predict who will die or who will suffer from diseases and anticipate cure for saving human life, that is a big challenge wish necessitates great efforts to create the necessary tools and platforms. Our paper proposed an architecture of platform based healthcare predictive analytics wish respond to this request. All that push into the integration and development of predictive analysis algorithm for analyzing medical data because nowadays a simple analyze is not sufficient about extracting useful knowledge from it is the biggest challenge .

REFERENCES

- [1] H.Madhusudhana Rao ; H.Madhusudhana Rao ; Dr. B Rambhupal Reddy ; SURVEY OF ADAPTING CLOUD COMPUTING IN HEALTHCARE ; International Journal of Advanced Research in Engineering and Applied Sciences ; 2278-6252
- [2] BARBARA CALABRESE ; MARIO CANNATARO ; CLOUD COMPUTING IN HEALTHCARE AND BIOMEDICIN , SCPE ; Volume 16, Number 1, pp. 1–18. <http://www.scpe.org>
- [3] Sanjay P, Ahuja, Sindhu Mani ,Jesus Zambrano ; A Survey of the State of Cloud Computing in Healthcare ; Canadian Center of Science and Education. Vol. 1, No. 2; 2012 ; September 19, 2012 ; <http://dx.doi.org/10.5539/nct.v1n2p12>
- [4] Michael J. Ward , Keith A. Marsolo , Craig M. Froehle . Applications of business analytics in healthcare.ELSEVIER . Business Horizons (2014) 57, 571—582. Available online at www.sciencedirect.com
- [5] [5] Fabricio F. Costa ; Big data in biomedicine ; ELSEVIER. Drug Discovery Today _ Volume 19, Number 4 _ April 2014 Available online at <http://www.sciencedirect.com/science/article/pii/S1359644613003644>
- [6] Sunkara . Vaishnavi,Akula.Suneetha ; Role of Cloud Computing in Health Monitoring System ; IJSEAT, Vol 2, Issue 10, October – 2014
- [7] Richard Hillestad, James Bigelow, Anthony Bower, Federico Girosi, Robin Meili, Richard Scoville, and Roger Taylor ; Can ElectronicMedical Record Systems Transform Health Care? Potential Health Benefits, Savings, And Costs. *Health Affairs*, 24, no.5 (2005):1103-1117 .Available at <http://content.healthaffairs.org/content/24/5/1103.full.html>
- [8] Sungyoung Oh, Jieun Cha, Architecture Design of Healthcare Software-as-a- Service Platform for Cloud-Based Clinical Decision Support Service ; Healthc Inform Res. 2015 April;21(2):102-110. pISSN 2093-3681 • eISSN 2093-369X. Available at <http://dx.doi.org/10.4258/hir.2015.21.2.102>

- [9] Nura Esfandiari , Mohammad Reza Babavalian ; Knowledge discovery in medicine: Current issue and future trend ;ELSEVIER ; Expert Systems with Applications 41 (2014) 4434–4463. Available at <http://www.sciencedirect.com/science/article/pii/S0957417414000232>
- [10] Jonathan Northover, Brian McVeigh, Sharat Krishnagiri .Healthcare in the cloud: the opportunity and the challenge . MLD . Available at http://www.sunquestinfo.com/images/uploads/CMS/445/mlo_02-12014_-_healthcare_in_the_cloud.pdf
- [11] Santosh Kumar, Wendy J. Nilsen ; Mobile Health Technology Evaluation ; ELSEVIER ; Volume 45, Issue 2, August 2013, Pages 228–236 Available at <http://www.sciencedirect.com/science/article/pii/S0749379713002778>
- [12] Gabriel I. Barbas, Sherry A. Glied, New Technology and Health Care Costs — The Case of Robot-Assisted Surgery ; The NEW ENGLAND JOURNAL of MEDICINE. N Engl J Med 2010; 363:701-704August 19, 2010DOI: 10.1056/NEJMp1006602 .Available at <http://www.nejm.org/doi/full/10.1056/NEJMp1006602>
- [13] Kenney Ng , Amol Ghoting ; PARAMO: A PARAllel predictive MOdeling platform for healthcare analytic research using electronic health records ; ELSEVIER . Journal of Biomedical Informatics 48 (2014) 160–170 . Available at <http://www.sciencedirect.com/science/article/pii/S1532046413002037>
- [14] Lena Griebel, Hans-Ulrich Prokosch. A scoping review of cloud computing in Healthcare ; BMC Medical Informatics and Decision Making (2015); DOI 10.1186/s12911-015-0145-7.Available at <http://www.biomedcentral.com/1472-6947/15/17/>
- [15] Marianthi Theoharidou, Nikos Tsalis, Smart Home Solutions for Healthcare: Privacy in Ubiquitous Computing Infrastructures. Available online at <http://www.cis.aueb.gr/Publications/Smart%20Home%20-%20Site%20TR.pdf>
- [16] Larry J. Kricka , Tracey G. Polsky. The future of laboratory medicine — A 2014 perspective.ELSEVIER. Clinica Chimica Acta 438 (2015) 284–303. Available at <http://www.sciencedirect.com/science/article/pii/S0009898114004008>
- [17] Karthik Kambatlaa, Giorgos Kollias. Trends in big data analytics. J. Parallel Distrib. Comput. 74 (2014) 2561–2573. Available at <http://www.sciencedirect.com/science/article/pii/S0743731514000057>
- [18] Vladimir Stantchev,1 Ricardo Colomo-Palacios, Cloud Computing Based Systems for Healthcare. Volume 2014, Article ID 692619, 2 pages. Available at <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4032744/>
- [19] Alex Guazzeli. Predictive analytics in healthcare The importance of open standards. ZEMENTIS IBM Devellopers Worker . 29 November 2011. Available at <http://zementis.com/knowledge-base/predictive-analytics-in-healthcare-the-importance-of-open-standards/>
- [20] Wullianallur Raghupathi ; Viju Raghupathi ; Big data analytics in healthcare: promise and Potential. Health Information Science and Systems 2014.Available at <http://www.biomedcentral.com/content/pdf/2047-2501-2-3.pdf>
- [21] Bonnie Feldman ;Ellen M. Martin ;Tobi Skotnes ; Big Data in Healthcare Hype and Hope ; Business Devloperment for Digital Health ; October 2012. Availabl at <http://thegrbluebook.com/wp-content/uploads/2013/02/107279699-Big-Data-in-Healthcare-Hype-and-Hope1.pdf>
- [22] Srinath Srinivasa ; Sameep Mehta ; Big Data Analytics. Springer ;December 20-23, 2014 ; Volume 8883 2014 .Available at <http://link.springer.com/book/10.1007/978-3-319-13820-6>
- [23] Dylan Maltby ;Big Data Analytics ; ASIST 2011, October 9-13, 2011, Available at https://www.ischool.utexas.edu/~dmaltby/Big_Data_Analytics.pdf
- [24] Peter Groves ;Basel Kayyali ;The ‘big data’ revolution in healthcare ; McKinsy and Company april 2013. Center for US Health System Reform Business Technology Office. Available at <http://digitalstrategy.nl/wp-content/uploads/E2-2013.04-The-big-data-revolution-in-US-health-care-Accelerating-value-and-innovation.pdf>
- [25] Steven S. Eisenberg, Practical Predictive Analytics for Healthcare 101.Avanade ; 2013 ;Available at <http://www.avanade.com/~media/documents/bi-white-paper-healthcare-analytics-practical-predictive-analytics-101-may-2013.pdf>
- [26] Hsinchun Chen ; BUSINESS INTELLIGENCE AND ANALYTICS: FROM BIG DATA TO BIG IMPACT ; MIS Quarterly Available http://hmchen.shidler.hawaii.edu/Chen_big_data_MISQ_2012.pdf
- [27] Alfredo Cuzzocrea ; Il-Yeol Song ; Analytics over Large-Scale Multidimensional Data: The Big Data Revolution ; DOLAP’11, October 28, 2011, ACM 978-1-4503-0963-9/11/10. Available at <http://dl.acm.org/citation.cfm?id=2064695>
- [28] Rajeev Gupta, Himanshu Gupta, Cloud Computing and Big Data Analytics: What Is New from Databases Perspective?; SPRINGER ; BDA 2012, LNCS 7678, pp. 42–61, 2012. Available at http://link.springer.com/chapter/10.1007/978-3-642-35542-4_5

- [29] Borkar VR, Carey MJ, Chen L: Big data platforms: what's next? ACM Crossroads 2012, 19(1):44–49.
- [30] Zikopoulos PC, Eaton C, DeRoos D, Deutsch T, Lapis G: Understanding Big Data – Analytics for Enterprise Class Hadoop and Streaming Data. McGraw-Hill: Aspen Institute; 2012.
- [31] N., Cristianini., J, Shawe-Taylor.: An Introduction to Support Vector Machines and Other Kernel-based Learning Methods,Cambridge University Press, 2013, 200 pp, available at <http://ebooks.cambridge.org/ebook.jsf?bid=CBO9780511801389>]
- [32] R. M. Rahman, F. Rabbi: Using and comparing different decision tree classification techniques for mining ICDDR, B Hospital Surveillance data’, elsevier, V, 38, pp 11421–11436
- [33] S. L. Ang, H. Ch. Ong and H. Chin Low.: Classification Using the General Bayesian Network ,science and technology, 24 (1) 205 – 211 pp, (2016)]
- [34] Top Data Mining Algorithms Identified by IEEE & Related Python Resources: <http://www.datasciencecentral.com/profiles/blogs/python-resources-for-top-data-mining-algorithms>