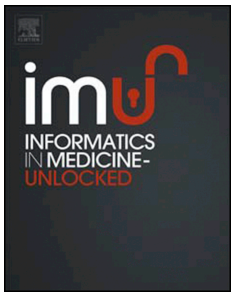
Accepted Manuscript



An optimal big data workflow for biomedical image analysis

Aurelle Tchagna Kouanou, Daniel Tchiotsop, Romanic Kengne, Zephirin Djoufack Tansaa, Ngo Mouelas Adele, René Tchinda

PII: S2352-9148(18)30084-4

DOI: 10.1016/j.imu.2018.05.001

Reference: IMU 109

To appear in: *Informatics in Medicine Unlocked*

Received Date: 5 April 2018

Revised Date: 8 May 2018

Accepted Date: 8 May 2018

Please cite this article as: Tchagna Kouanou A, Tchiotsop D, Kengne R, Tansaa ZD, Adele NM, Tchinda René, An optimal big data workflow for biomedical image analysis, *Informatics in Medicine Unlocked* (2018), doi: 10.1016/j.imu.2018.05.001.

An Optimal Big Data Workflow for Biomedical Image Analysis

Aurelle Tchagna Kouanou[1,*], Daniel Tchiotsop[2], Romanic Kengne[1, 3], Zephirin Djoufack Tansaa[4], Adele Ngo Mouelas[1], René Tchinda[5]

1 Laboratoire de Matière Condensée d'Electronique et de Traitement du Signal (LAMACETS), Faculty of Science,University of Dschang, P.O.Box 67 Dschang, Cameroon.

2 Laboratoire d'Automatique et d'Informatique Appliquée (LAIA), IUT-FV de Bandjoun, Université de Dschang-Cameroun, B.P. 134 Bandjoun.

3 Research Group on Experimental and Applied Physics for Sustainable Development (EAPhySuD), P.O.Box 412 Dschang, Cameroon

4 National Advanced School of Posts, Telecommunications and Information Communication and Technologies (SUP'PTIC), Yaounde- Cameroon, P.O.Box 8950 Yaounde.

5 Laboratoire d'Ingénierie des Systèmes Industriels et de l'Environnement (LISIE), IUT-FV de Bandjoun, Université de Dschang- Cameroun, B.P. 134 Bandjoun.

[*]Corresponding author: Aurelle Tchagna Kouanou

Email: **tkaurelle@gmail.com**

Phone: (+237) 696 263 641

Abstract

Background and Objective: In the medical field, data volume is increasingly growing, and traditional methods cannot manage it efficiently. In biomedical computation, the continuous challenges are: management, analysis, and storage of the biomedical data. Nowadays, big data technology plays a significant role in the management, organization, and analysis of data, using machine learning and artificial intelligence techniques. It also allows a quick access to data using the NoSQL database. Thus, big data technologies include new frameworks to process medical data in a manner similar to biomedical images. It becomes very important to develop methods and/or architectures based on big data technologies, for a complete processing of biomedical image data.

Method: This paper describes big data analytics for biomedical images, shows examples reported in the literature, briefly discusses new methods used in processing, and offers conclusions. We argue for adapting and extending related work methods in the field of big data software, using Hadoop and Spark frameworks. These provide an optimal and efficient architecture for biomedical image analysis. This paper thus gives a broad overview of big data analytics to automate biomedical image diagnosis. A workflow with optimal methods and algorithm for each step is proposed.

Results: Two architectures for image classification are suggested. We use the Hadoop framework to design the first, and the Spark framework for the second. The proposed Spark architecture allows us to develop appropriate and efficient methods to leverage a large number of images for classification, which can be customized with respect to each other.

Conclusions: The proposed architectures are more complete, easier, and are adaptable in all of the steps from conception. The obtained Spark architecture is the most complete, because it facilitates the implementation of algorithms with its embedded libraries.

Keywords: Biomedical images; Big Data; Artificial Intelligent; Machine learning; Hadoop/Spark.

1. Introduction

The term “Big Data” has become a buzzword in recent years, with its usage frequency doubled each year within the last decade according to common search engines [1]. Big data is often defined by three major characteristics called the “3V”: volume (amount of data

generated), variety (data from different categories) and velocity (speed of data generation) [2-8]. Nowadays, we have two more "V": variability (inconsistency of data) and veracity (quality of captured data) [4, 5]. Thus big data problems are now identified by the "5V". Big data is not a new term. The big data application is applied in many fields of science including health [1-4], agriculture [9, 10], internet with social network [11], etc.

Big data in health is concerned with meaningful datasets that are too big, too fast, and too complex for healthcare providers to process and interpret with existing tools [1, 12]. Data are daily generated at unprecedented rates from different heterogeneous sources (e.g., laboratory and clinical data, patients' symptoms uploaded from distant sensors, hospitals operations, and pharmaceutical data) [7]. In biomedical imaging, the techniques that are well established within clinical settings to capture an image are [3]: computed tomography, magnetic resonance imaging, x-ray, molecular imaging, ultra sound, photo-acoustic imaging, fluoroscopy, and positron emission tomography - computed tomography (PET-CT). These techniques take the medical images with high definition and large sizes. The advanced analysis of biomedical image datasets has many beneficial applications. It enables to personalize remotely radiological services (e.g., doctors can monitor online image of patients in order to provide a prescription). However, specialized doctors are very few and cannot diagnose all these millions of images generated. With this rise of biomedical image data, new demands to Artificial Intelligent (AI) for machine learning (ML) systems to learn complex models are made. ML is used as the primary mechanism for distilling structured information and knowledge from raw data, turning them into automatic predictions and actionable hypotheses for diverse applications [13].

In this paper, we will focus specifically on biomedical imaging with Big Data technologies, along with Artificial Intelligent (AI) for machine learning. An architectural workflow describes the optimal algorithm and method reported in the literature. We will present a workflow performing the steps of acquisition of biomedical image data, analysis, storage, processing, querying, classification, and automatic diagnosis of biomedical images. We describe the importance of applying compressed biomedical images in a big data architecture. Two main big data architectures are proposed. The one is based on MapReduce in Hadoop and the other is based on Spark. The two proposed architectures will be compared. The paper is organized as follows: section 2 reviews published methods in the field. In section 3, these methods are exploited theoretically throughout our work. Section 4 presents the design and construction of the architectures. Results are analyzed and discussed in section 5. A conclusion and future work are provided in section 6.

2. State of the art

In medicine, the data encountered are mainly obtained from patients. These data consist of physiological signals, images, and videos. They can be stored or transmitted using appropriate hardware and techniques. One of the services used in medicine for the storage and transmission of image data is the Picture Archiving and Communication System (PACS). PACS are popular for delivering images to local display workstations, which is accomplished primarily through existing protocols like digital image communication in medicine (DICOM). However, data exchange with a PACS is highly standardized [14], and this system relies on using structured data solely to retrieve medical images rather than leveraging the unstructured content of the biomedical images [6]. Many works have been performed in managing and analyzing structured and unstructured data images using the concept of big data and artificial intelligence (AI).

AI is now used more intensively in medicine. Indeed, AI is required to automate the decision and diagnosis of diseases [15, 16]. In medicine, AI can be used to develop a classification algorithm [17, 18], make decisions [13, 19, 20] and for predictive analysis [13]. Therefore we need to develop a solution that can analyze and assist with diagnosis using these images. Hence, it is necessary to implement ML algorithms in order to automate decision-making in the diagnostic system of medical images. Human physicians may not be replaced by machines in the future, but AI can definitely assist physicians to make better clinical decisions or even replace human judgment in certain functional areas of healthcare (e.g., radiology) [16]. Concerning the radiology domain, we propose within this work to develop an architecture that implements AI to diagnose or make decisions concerning a biomedical image. It is worth mentioning that several papers have been published concerning big data and AI in biomedical imaging. In that way, Istephan *et al.* in [6] implemented and examined the feasibility of having a framework to provide efficient querying of unstructured data in unlimited ways. Their proposed framework is used to evaluate a query in two phases. In phase one, structured data are used to filter the clinical data warehouse, while in phase two, feature extraction modules are executed on the unstructured data in a distributed manner via Hadoop, to complete the query. However, their work was only limited to Hadoop, which does not include many libraries such as Machine Learning, SQL, etc. In 2017, Yang *et al.* examine two important aspects that are central to modern big data bioinformatics analysis – software scalability and validity [5]. They discussed how state-of-the-art software testing techniques that are based on the idea of multiple executions, such as metamorphic testing, can be used to

implement an effective bioinformatics quality assurance strategy. Dilsizian and Siegel in [15] showed the importance of AI, big data, and massively parallel computing in medicine and cardiac imaging to personalize diagnosis and treatment. Although their works forecast future application of AI systems in medicine, they did not provide an explicit big data architecture for these implementations.

To the best of our knowledge, none of the existing researches present a complete workflow to manage biomedical images. This drawback is the main interest of this paper. Indeed, we have designed a workflow implementing optimal algorithms combining AI and ML to efficiently manage (acquire, analyze, process, share …) biomedical images. Therefore, we propose a complete and optimal workflow based on big data technology and optimal algorithms (AI and ML) drawn from the literature, to manage biomedical images. The classification step within the proposed optimal flow will be considered as a study case implementing big data analysis technology (Hadoop and Spark) and can be customized to all remaining steps.

3. Methods

Medical imaging supplies important information on organ function and anatomy in order to detect the state of diseases. We propose a workflow to handle the steps of image processing. The main goal of the workflow is to give in each step the optimal method that we have to implement so as to have an optimal big data architecture solution.

In this section, a conceptual framework was developed to provide a systematic method necessary for analyzing big data in biomedical imaging from patient data. The conceptual framework proposed is summarized in Fig.1. This figure shows the parts of big data processes for biomedical image processing. We rely on results of recent publications to design optimal algorithms or methods for each big data processing step.

Data management is the organization, administration, and governance of large volumes of both structured and unstructured data. The goal of big data management is to ensure a high level of data quality and accessibility for business intelligence and big data analytics applications [21].

- *Clinical images acquisition.* In biomedical imaging, the techniques that are well established within clinical settings to acquire an image are [3]: computed tomography, magnetic resonance imaging, x-ray, molecular imaging, ultra sound, photo-acoustic imaging, fluoroscopy, positron emission tomography-computed tomography (PET-CT). These techniques take the medical images in a higher definition and large sizes. These methods generally give an internal image of human body parts. However, if a patient

suffers from a disease that affects the skin, the technician can then use a camera or smartphone to take the picture of the skin and port it into the system. This enables smartphone cameras to act as acquisition devices. The captured data image is then transferred into the database of the big data platform for processing.

- *Extraction, Cleaning, and Annotation*. Extraction refers to a technique that enables to obtain useful biomedical images from the raw data and, refines them so that they can be used in the following analytic steps. Cleaning is the process that eliminates noise on acquired images. At this stage, we just need a filter. Annotations rely on a technique, which allows adding some information concerning the patient on images.
- *Integration and representation*. This is the step which involves the automatic clustering of images in the databases. Preview of images is also possible at this level before analyses.

Concerning big data analysis and share, it is an entire program that bears the development of theoretical, mathematical, artificial intelligence, statistical methods for analysis of biomedical images, clinical diagnosis and patient monitoring.

- *Modeling.* The modeling step is based on mathematical models and computational algorithms. This can be used to format images in a way that is easier to understand. This step is not compulsory, and depends upon the nature of the image. For example, a 3D image can be modeled in 2D to facilitate its manipulation.
- *Classification.* Classification is one of the classical concerns in image processing [22-24]. Classification is an example of pattern recognition [25]. Classification in machine learning concerns a problem of identifying to which set of categories a new population belongs. When category membership is known, the classification is done on the basis of a training set of data containing observations. An example would be to assign a given biomedical image into “anatomic body part” or “biological systems” classes. It is worth noticing that ML algorithms can be classified into three major categories which are: supervised learning, semi-supervised learning, and unsupervised learning. Supervised learning is suitable for predictive modeling via building some relationships between the patient traits (as input) and the outcome of interest (as output) [16]. Unsupervised learning is known as clustering for feature extraction [16, 26]. Semi-supervised learning is a hybrid between supervised and unsupervised learning, which is suitable for scenarios where the outcome is missing for certain subjects [16]. Thus, supervised learning is used

to classify, regress or estimate data processing tasks. And unsupervised learning is utilized to do data processing tasks such as clustering or prediction.

In our workflow, the classification steps are processed under a supervised learning algorithm via a support vector machine (SVM). SVM is chosen from several other supervised learning algorithms because SVM and neural network are two well-known techniques used to classify biomedical image data. Indeed, in medical imaging, SVM and neural networks take up to 42% and 31% respectively of the most used algorithms [16]. This statistic shows the efficiency of the SVM algorithm. SVM is mainly used for classifying the subjects into two groups, where the outcome *Yi* is a classifier. $Yi = -1$ or 1 and represents whether the *i*th considered patient belongs to group 1 or 2, respectively [16, 22]. SVM uses the learned features and patterns for application on labeled data from a given source domain, resulting in a linear classification model that outperforms other methods [27, 28]. SVM is successfully applied to biomedical images datasets as shown in [16, 28]. The classification step could be assistive to organize image databases into image categories prior to retrieval or diagnostics. Henceforth, each specialist will see only the biomedical images of his competence field.

- *Prediction and decision.* Many computer-aided diagnoses have experience that is more intensive in the medical imaging field. These methods are based on the ML algorithm. Deep Convolutional Neural Network (CNN) is one of the most used to automate the process of diagnosing symptoms from patient information. This is because the CNN yields over 88% accuracy for diagnosis and treatment suggestion [16, 29]. For example, in 2017, Esteva *et al.* trained clinical images taken by smartphones using CNN and identified skin cancer [30]. Esteva *et al.* obtained a specificity and sensitivity over 91%, which indicates the performance of CNN. Geert *et al.* applied CNN on medical images dataset to detect automatically, symptoms like cancer prostate or sentinel lymph node [31]. The CNN, which consists of multiple layers of neuron-like computational connections with step-by-step minimal processing, achieves significant improvements [32]. Given that our architecture has to work on very large image data volumes, the CNN will be appropriate for the automatic diagnostic step. Further, a CNN requires a huge number of training images (e.g., 1,000,000) to determine a large number of parameters in the CNN [33].
- *Validation*. Validation is performed by calculating sensitivity and specificity [30]:

$$Sensitivity = \frac{True\ positive}{positive};\ Specificity = \frac{True\ negative}{negative}$$

Where *true positive* is the number of symptoms correctly predicted on the images, *positive* is the total number of symptoms shown, *true negative* is the number of correctly predicted benign symptoms, and *negative* is the number of benign symptoms shown.

- *Transformation (Compression).* This step provides transformation of the images. Here transformation refers to compression. Compressing data in big data architecture is important as we see in references [34-36]. Indeed, big data compression techniques allow the taming of the complexity of big data management tasks within such frameworks. This beneficially influences all the other activities that are delivered as services in a reference Cloud architecture [37]. The compression method is representative of data reduction for big data analytics. In fact, reducing the size of data makes them analytically computational, less expensive and thus faster, especially for the data through putting to the system rapidly [34]. Basically in this step, the idea behind big data compression consists of reducing the size of data (images) to gain storage capacity, transmission time, management efficiency and querying. In image compression, there are always new approaches that are being tried and tested to improve the quality of the reconstructed image. There are two types of compression: lossless compression, where reconstruction data is identical to the original; and lossy compression, where there is a loss of data. However, lossless compression is limited because compression rates are very low [38]. The compression ratio of lossy compression scheme is very high. Some biomedical images cannot tolerate distortions of the reconstructed image because the slightest information on the image is important. Thus we focused on lossless compression in our architecture as in [36]. Lossless compression algorithms are achieved generally by *entropy encoding*, such as the Shannon-Fano algorithm, Huffman coding, arithmetic coding, Lempel-Ziv-Welch algorithm [38, 39]. Huffman coding is chosen to be implemented in our architecture because it is a compression algorithm based on the frequency of appearance of characters in an original document. Developed by David Huffman in 1952, the method relies on the principle of shorter codes allocation for frequent values and longer codes for less frequent values [38-42]. This coding uses a table containing the apparent frequencies of each character to establish an optimal binary string representation. The procedure is broken up into three parts:
  - First, the creation of the frequency appearance of character table in the original data.

- Afterward, the creation of a binary tree according to the previously calculated table.
- Finally, encoding symbols in an optimal binary representation.

- *Share or storage.* Big data applications commonly use Not Only SQL (NoSQL) technologies as a database [43-46]. NoSQL refers to a database category that appeared in 2009 which differs from the relational databases [43]. Indeed, one of the recurring problems of relational databases is the loss of performance when one should process a very large volume of data. Moreover, distributed architectures provide the need to adapt solutions natively to replication mechanisms of data and load management [43, 47]. Cloud computing technologies can also be used to facilitate sharing of data. Cloud computing is an on-demand computing model composed of autonomous, networked IT (hardware and/or software) resources [48]. Cloud computing is suited for big data bioinformatics applications as it allows for on-demand provisioning of resources with a pay-as-you-go model, thus eliminating the need of purchasing and maintaining costly local computing infrastructure for performing analyses [5]. Cloud computing platforms use hypervisor technology to provide dynamic access to virtualize computing resources. Virtualisation is a technique that enables a single hardware resource to host a number of the independent virtual machines, where each virtual machine shares some of the hardware resources.

4. Results

In this section, we present the big data architectures to handle the step of workflow described in Section 3. The main goal of these architectures is to see how the data image is processed since implementation. However, we base on Hadoop framework, and Spark framework, and propose two architectures for classification step as shown in Fig.1. Indeed, the classification stage represents one of the main parts of the proposed workflow. In fact, the classification step groups each category of biomedical images (lunch cancer, pelvis, skin image…) with each order. Finally, diagnostic and analysis time will be minimized both for specialist or CNN algorithms. Henceforth, the classification step has to be well-designed.

4.1 Hadoop architecture

Hadoop is an Apache open source framework based on parallel programming. The Hadoop File System was developed using a distributed file system design and is called HDFS (Hadoop Distributed File System) [7]. HDFS holds a very large amount of data, provides

easier access, and makes applications available for parallel processing. The distributed file system is designed to process large amounts of data with sequential read and write operation. Each file is broken into chunks, and stored across multiple data nodes.

Hadoop implements MapReduce programming. MapReduce is a processing technique and programming model done in a lateral and scattered manner [7, 49-51].

MapReduce programming is a special form of a directed acyclic graph (DAG) which is applicable to a wide range of used cases. MapReduce is organized in two functions [51, 52]. The first one is a Map function, which transforms an element of data into some number of key/value pairs. The second is the Reduce function, which is used to merge the values (of the same key) into a single result. The proposed architecture is shown in Fig. 2. In this architecture, we can observe the simplicity of the implementation of MapReduce programming. All of the images resulting from the modeling step will be automatically classified in each defined category. That will optimize the prediction and decision methods to be applied to the images. Thus, we can use Hadoop and apply a deep learning algorithm in each category resulting from the classification step, in order to predict and make decisions automatically on each image. The architecture of Fig. 2 can be customized and applied in all process of Fig. 1.

Hadoop is suited for processing large amounts of data. However, others frameworks such as Spark, allows the achievement of real-time processing, and is already implemented in several libraries which facilitate its usage and programming.

4.2 Spark architecture

Spark has a programming model similar to MapReduce but extends it with a data-sharing abstraction called Resilient Distributed Datasets (RDDs) [7, 53]. RDDs are fault-tolerant collections of objects partitioned across a cluster that can be manipulated in parallel [54]. Spark offers a unified and complete framework to manage the different requirements for big data processing with a variety of datasets (graph data, image/video, text data, etc.) from different sources (batch, real-tome streaming). In addition to Map and Reduce operations, Spark also supports SQL queries, streaming data, machine learning and graph processing data. With capabilities like in-memory data storage and near real-time processing, the performance can be several times faster than other big data technologies. Spark runs over the existing Hadoop Distributed File System (HDFS) infrastructure to provide enhanced and additional functionalities. Users create RDD’s by applying operations called “transformations” (such as map,filter and groupBy) to their data. We use these properties to

develop an architecture enabled to make the classification using the Map and groupBy methods. Fig. 3 presents our Spark architecture model for the classification of image data. In order to calculate the number of images in each class, we used the method ReduceByKey proposed in the Spark framework. In Fig. 3, we used only one ReduceByKey. However, depending on the processing, we can find several ReduceByKey in a Spark architecture. Thus, in Fig. 3, the images will be counted and formed into a matrix. In addition, we will be able to locate an image in its original sample, thanks to this matrix.

5. Discussion

New imaging technologies give rise to new challenges in image management. Imaging techniques produce huge amounts of structured/unstructured data that require reliable and efficient algorithms and methods for interpretation and analysis. This work proposed a workflow based on big data technology in biomedical image analysis. The peculiarity of our workflow is that it gives us the optimal methods and algorithms to use in each design step. By using big data technology and AI techniques, we can automate the processes of acquisition, management, processing/analysis, and sharing/storage of biomedical image data. Thus, our proposed workflow does not only allow the exchange of image data as in the case of conventional systems [14, 54]. It is interesting to mention that we have designed two architectures based on Hadoop for the first one and Spark for the second one. Both proposed architectures allow performing of the classification step. We specify that these architectures proposed can be customized on all steps of Fig. 1. With regard to our two architectures (Fig. 2 and Fig. 3), we notice that Hadoop applications are easier to implement than Spark applications. However, Spark includes all libraries used to automate our image workflow process. Therefore, for complete automation, we need to work with the Spark framework. The proposed architecture can be compared to the architecture proposed in [5]. Our Hadoop architecture for classification is almost the same type as in [5]. Spark architecture proposed in [5] makes it possible to count the number of genes in the processing system; however the proposed Spark architecture within this work groups together the counting of images and the classification of images by categories. Thus, this architecture can be considered as a valuable contribution compared with the results encountered in the literature. The implementation of these proposed architectures are beyond the scope of this work, and will be addressed in our future works.

The ability to easily adapt our architectures can be used to improve or modify the end user's systems, such as electronic medical records or PACS, with the evolution of imaging technologies, processing, and storage. Our workflow based on the NoSQL database can support biomedical images available in other commonly used formats, and also DICOM data. Our workflow provides another aspect to how this will be structured for management systems and data analysis biomedical images. Thus, it will enable the facilitation of tasks for remote diagnosis and telemedicine.

6. Conclusion

Big data biomedical image was considered herein, including the methods to generate, manage, represent, and analyze imaging information for biomedical application. In this paper, we proposed a workflow for the management and analysis of biomedical image data based on the tools of big data technology. To design our workflow, we conducted a literature review to identify the best algorithms and methods most suitable for the management and analysis of biomedical images. Thus, we were able to give for each step of our workflow, a method/algorithm to finally obtain an optimal architecture. Our proposed workflow does not only allow the exchange of image data as in the case of conventional systems, but it manages also from acquisition, to the storage and sharing of images. In order to show the use of big data framework in our workflow, we proposed and designed two architectures to perform the classification step. The first architecture proposed is based on the Hadoop framework and the second on the Spark. We noted that the Spark architecture was the most complete because it facilitates the implementation of algorithms with its embedded libraries. Our proposed architectures are more complete, easier, and are adaptable in all the steps of conception than those proposed in literature. As future work, we should develop/implement a real application of our workflow proposed in Fig. 1 with the Spark framework.

Acknowledgments including declarations

Acknowledgments: The authors are very grateful to Mrs. Talla Isabelle from linguistic center of Yaounde for improving the overall English level and mistakes within our manuscript. The authors would like to acknowledge and thank Pr. Fotsin Hilaire, M. Fozin Theophile, M. Mezatio Anicet, M. Kuetche Christian from LAMACETS and M. Bogne Baron, M. Tamko Clarence, M. Mache Kevin and M. Kamguia Herve from Inchtech's Team for their support and assistance during the conception of that work.

Funding and competing interests: We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

Ethical approval: This article does not contain any studies with human participants and/ or animals performed by any of the authors.

Figure caption

Figure.1 Big Data workflow for biomedical image processing. Only classification step will be designed with Hadoop/ Spark framework.

Figure.2: Hadoop MapReduce pipeline for biomedical image classification.

Figure.3: Spark Map Reduce pipeline for biomedical image classification and counting.

References

[1] J. Andreu-Perez, C. C. Y. Poon, R. D. Merrifield, S. T. C. Wong, and G-Z Yang, Big Data for Health, Journal of Biomedical and Health Informatics. 19 (4) (2015) 1193-1208.

[2] J. Luo, M. Wu, D. Gopukumar and Y. Zhao, Big Data Application in Biomedical Research and Health Care: A Literature Review, Biomedical Informatics Insights 8 (2016) 1-10.

[3] A. Belle, R. Thiagarajan, S. M. R. Soroushmehr, F. Navidi, D. A. Beard and K. Najarian1, Big Data Analytics in Healthcare, BioMed Research International. (2015) 16p.

[4] M. Viceconti, P. Hunter, and R. Hose, Big Data, Big Knowledge: Big Data for Personalized Healthcare, Journal of Biomedical and Health Informatics. 19 (4) (2015) 1209-1215.

[5] A. Yang, M. Troup, J. W.K. Ho, Scalability and Validation of Big Data Bioinformatics Software, Computational and Structural Biotechnology Journal. (2017) 8p. Article in press.

[6] S. Istephan, M-R. Siadat, Unstructured medical image query using big data – An epilepsy case study, Journal of Biomedical Informatics. 59 (2016) 218–226.

[7] A. Oussous, F-Z. Benjelloun, A. A. Lahcen, S. Belfkih, Big Data technologies: A survey, Journal of King Saud University – Computer and Information Sciences (2017) 18 p. Article in press.

[8] L. Wang, Y. Wang and Q. Chang, Feature Selection Methods for Big Data Bioinformatics: A Survey from the Search Perspective, Methods. 111 (2016) 21-31. doi.org/10.1016/j.ymeth.2016.08.014

[9] S. Wolfert, L. Ge, C. Verdouw, M-J. Bogaardt, Big Data in Smart Farming – A review, Agricultural Systems. 153 (2017) 69–80.

[10] H. Zhang, X. Wei, T. Zou, Z. Li, and G. Yang, Agriculture Big Data: Research Status, Challenges and Countermeasures, International Federation for Information Processing, CCTA. (2015) 137-143.

[11] P. Pääkkönen, D. Pakkala, Reference Architecture and Classification of Technologies, Products and Services for Big Data Systems, Big Data Research. 2 (2015) 166–186.

[12] J. Wang, M. Qiu, B. Guo, Enabling Real-time Information Service on Telehealth System over Cloud-based Big Data Platform, Journal of Systems Architecture (2016) doi: 10.1016/j.sysarc.2016.05.003

[13] E. P. Xing, Q. Ho, P. Xie, D. Wei, Strategies and Principles of Distributed Machine Learning on Big Data, Engineering. 2 (2016) 179–195.

[14] T. Doel, D. I. Shakir, R. Pratt, and al, GIFT-Cloud: A data sharing and collaboration platform for medical imaging research, Elsevier computer methods and programs in biomedicine. 139 (2017) 181-190. doi.org/10.1016/j.cmpb.2016.11.004.

[15] S. E. Dilsizian and E. L. Siegel, Artificial Intelligence in Medicine and Cardiac Imaging: Harnessing Big Data and Advanced Computing to Provide Personalized Medical Diagnosis and Treatment, Curr Cardiol Rep. 16:441 (2014) 1-8.

[16] F. Jiang, Y. Jiang, H. Zhi and al, Artificial intelligence in healthcare: past, present and future, BMJ Stroke and Vascular Neurology. (2017) 1-14. doi:10.1136/svn-2017-000101.

[17] J. L. Warner, P. Zhang, J. Liu, G. Alterovitz, Classification of hospital acquired complications using temporal clinical information from a large electronic health record, Journal of Biomedical Informatics. 59 (2016) 209–217.

[18] K. Mei, J. Peng, L. Gao, N. Zheng, J. Fan, Hierarchical Classification of Large-Scale Patient Records for Automatic Treatment Stratification, Journal of Biomedical and Health Informatics. 19 (4) (2015) 2168-2194.

[19] C. C. Bennett and K. Hauser, Artificial intelligence framework for simulating clinical decision-making: A Markov decision process approach, Artificial Intelligence in Medicine. 57 (2013) 9–19.

[20] E. S. Berner, Clinical Decision Support Systems: Theory and Practice, Third Edition Springer, Switzerland 2016. 319p.

[21] http://searchdatamanagement.techtarget.com/definition/big-data-management (Accessed on 14-09-17)

[22] X. Wang, N. Thome, M. Cord, Gaze latent support vector machine for image classification improved by weakly supervised region selection, Pattern Recognition. 72 (2017) 59-71. doi.org/10.1016/j.patcog.2017.07.001

[23] Q-N. Yuan, J. Lu, H. Huang, W. Pan, Research of batik image classification based on support vector machine, Computer Modelling & New Technologies. 18(12B) (2014)193-196.

[24] L. H. Thai, T. S. Hai, N. T. Thuy, Image Classification using Support Vector Machine and Artificial Neural Network, I.J. Information Technology and Computer Science. 5 (2012) 32-38. DOI: 10.5815/ijitcs.2012.05.05

[25] S. M. Weiss and L. Kapouleas, An Empirical Comparison of Pattern Recognition Neural Nets and Machine Learning Classification Methods, In Proceedings of the Eleventh International Joint Conference on Artificial Intelligence. (1989) 781-787.

[26] J. Qiu, Q. Wu, G. Ding, Y. Xu and S. Feng, A survey of machine learning for big data processing, EURASIP Journal on Advances in Signal Processing. 67 (2016) 1-16. DOI 10.1186/s13634-016-0355-x.

[27] M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald, E. Muharemagic, Deep learning applications and challenges in big data analytics, Journal of Big Data. 2 (1) (2015) 1-21. doi 10.1186/s40537-014-0007-7.

[28] E. Rezk, Z. Awan, F. Islam, A. Jaoua, S. Al Maadeed, N. Zhang, G. Das, N. Rajpoot, Conceptual data sampling for breast cancer histology image classification, Elsevier Computers in Biology and Medicine (2017), doi: 10.1016/j.compbiomed.2017.07.018.

[29] U.R. Acharya, S.L. Oh, Y. Hagiwara, J.H. Tan, H. Adeli, Deep convolutional neural network for the automated detection and diagnosis of seizure using EEG signals, Computers in Biology and Medicine (2017), doi: 10.1016/j.compbiomed.2017.09.017.

[30] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, S. Thrun, Dermatologist-level classification of skin cancer with deep neural networks, Nature. (2017) 1-11. Doi:10.1038/nature21056

[31] G. Litjens, C. I. Sánchez, N. Timofeeva, and al, Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis, Nature. (2016) 1-11. doi: 10.1038/srep26286.

[32] J-G. Lee, S. Jun, Y-W. Cho, and al, Deep Learning in Medical Imaging: General Overview, Korean J Radiol. 18 (4) (2017) 570-584. doi.org/10.3348/kjr.2017.18.4.570.

[33] K. Suzuki, Overview of deep learning in medical imaging, Radiol Phys Technol. (2017) 17p. DOI 10.1007/s12194-017-0406-5

[34] C-W. Tsai, C-F. Lai, H-C. Chao, A. V. Vasilakos, Big data Analytics, chapter 2, journal of Big data, vol 2 No 21 2015, Switzerland. B. Furht and F. Villanustre, Big Data Technologies and Applications.

[35] C. Yang, X. Zhang, C. Zhong and al, A spatiotemporal compression based approach for efficient big data processing on Cloud, Journal of Computer and System Sciences 80 (2014) 1563–1583.

[36] Z. QU, G. CHEN, Big data compression processing and verification based on Hive for smart substation, J. Mod. Power Syst. Clean Energy (2015) 3(3):440–446
DOI 10.1007/s40565-015-0144-9.

[37] A. Cuzzocrea, Big Data Compression Paradigms for Supporting Efficient and Scalable Data-Intensive IoT Frameworks, Proceedings of the Sixth International Conference on Emerging Databases: Technologies, Applications, and Theory, October 17-19, 2016, Jeju Island, Republic of Korea. 67-71.

[38] K. Sayood, Introduction to Data Compression, third ed., Morgan Kaufmann, San Francisco, 2006.

[39] D. Salomon, A Concise Introduction to data compression, Springer-Verlag, London, 2008.

[40] D. A. Huffman, A Method for the Construction of Minimum-Redundancy Codes, Proceeding of the I.R.E. 40 (1952) 1098-1101.

[41] U. Nandi and J. Kumar Mandal, Windowed Huffman Coding with Limited Distinct Symbols, Procedia Technology. 4 (2012) 589-594.

[42] M. Tomasz Biskup and W. Plandowski, Shortest Synchronizing Strings for Huffman Codes, Theoretical Computer Science. 410 (30-40) (2009) 3925-3941

[43] R. Bruchez, Les bases de données NoSQL et le Big Data: Comprendre et mettre en œuvre, 2e Edition Eyrolles, 2015 paris, 315p.

[44] S. Sakr and A. Elgammal, Towards a Comprehensive Data Analytics Framework for Smart Healthcare Services, Elsevier Big Data Res. (2016), http://dx.doi.org/10.1016/j.bdr.2016.05.002.

[45] T. Huang, L. Lan, X. Fang, P. An, J. Min, F. Wang, Promises and Challenges of Big Data Computing in Health Sciences, Elsevier Big Data Res. 2 (1) (2015) 2-11. doi.org/10.1016/j.bdr.2015.02.002.

[46] J. Bhogal, I. Choksi, Handling Big Data using NoSQL, IEEE 29th International Conference on Advanced Information Networking and Applications Workshops, 2015. DOI 10.1109/WAINA.2015.19.

[47] K. K-Y. Lee, W-C. Tang, K-S. Choi, Alternatives to relational database: Comparison of NoSQL and XML approaches for clinical data storage, Elsevier Computer Methods and Programs in Biomedicine. 110 (2013) 99-109. doi.org/10.1016/j.cmpb.2012.10.018.

[48] Q. F. Hassan, Demystifying Cloud Computing, The Journal of Defense Software Engineering. (2011) 16–21.

[49] Z. Lu, X. Wang, J. Wu, P. C.K. Hu, InSTechAH: Cost-Effectively Autoscaling Smart Computing Hadoop Cluster in Private Cloud, Journal of Systems Architecture (2017), doi: 10.1016/j.sysarc.2017.07.002

[50] R. Manjunath, Tejus, R. K. Channabasava, S. Balaji, A BigData MapReduce Hadoop Distribution Architecture for Processing Input Splits to solve the Small Data Problem, 2nd International Conference on Applied and Theoretical Computing and Communication Technology (ICATCCT) (2016), 484-487.

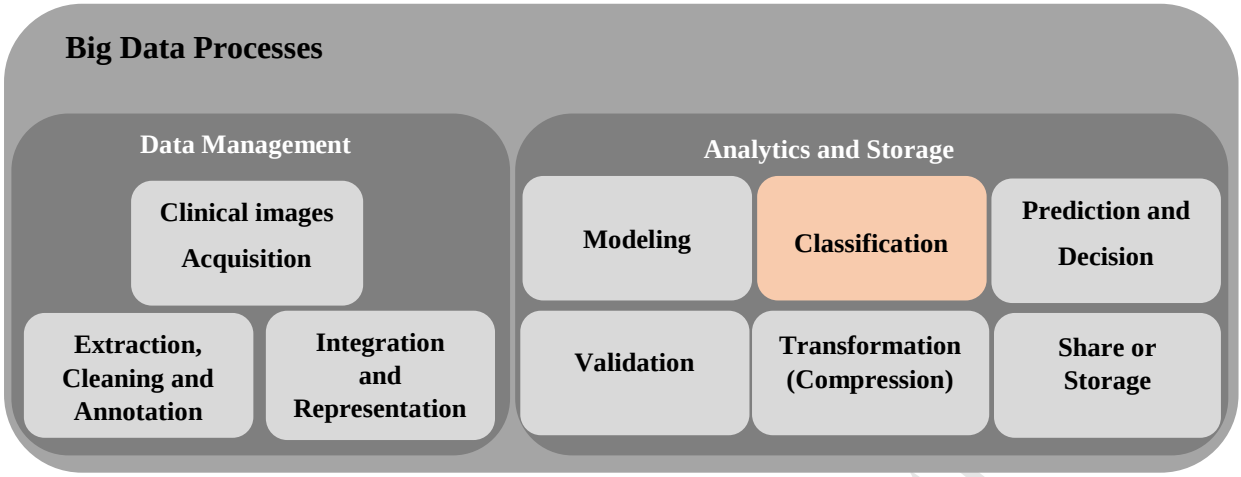
[51] T. Estrada, B. Zhang, P. Cicotti, R.S. Armen, M. Taufer A scalable and accurate method for classifying protein–ligand binding geometries using a MapReduce approach, Elsevier Computers in Biology and Medicine. 42 (2012) 758–771.

[52] S. Kamal, S. H. Ripon, N. Dey, A. S. Ashour V. Santhi, A MapReduce approach to diminish imbalance parameters for big deoxyribonucleic acid dataset, Computer Methods and Programs in Biomedicine. 131 (2016) 191-206. doi.org/10.1016/j.cmpb.2016.04.005.

[53] M. Zaharia, R. S. Xin, P. Wendell and al, Apache Spark: A Unified Engine for Big Data Processing, Communications of the ACM. 59 (11) (2016) 56-65. doi:10.1145/2934664.

[54] I. Moodley, S. Moodley, A comparative cost analysis of picture archiving and communications systems (PACS) versus conventional radiology in the private sector, S Afr J Rad. 19 (1) (2015) 1-7. http://dx.doi.org/10.4102/sajr.v19i1.634

Figure 1: Tchagna et *al.*



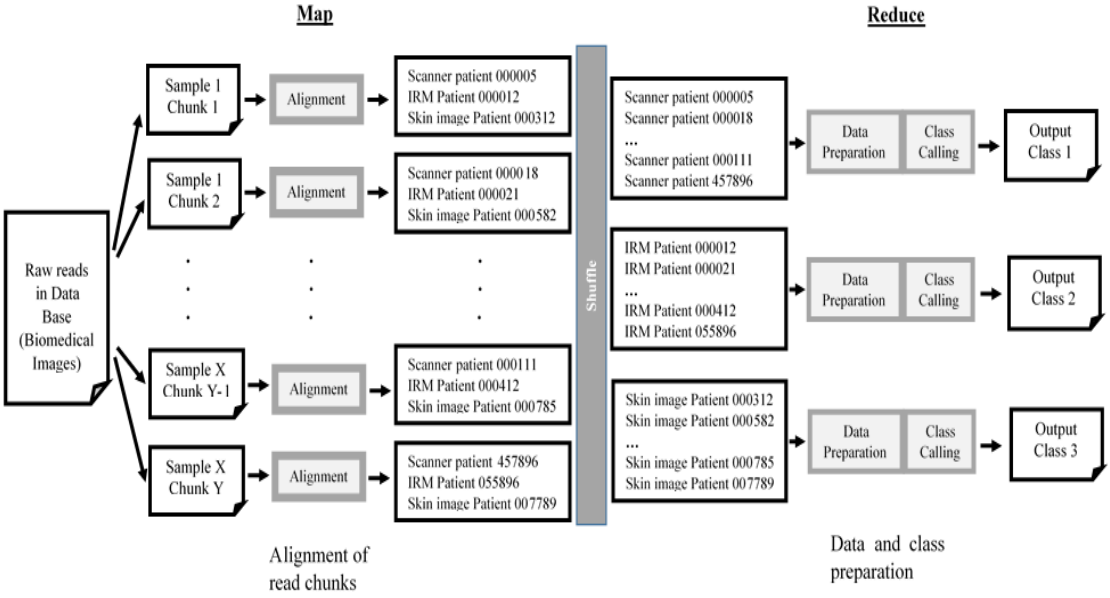


Figure 2: Tchagna et *al.*

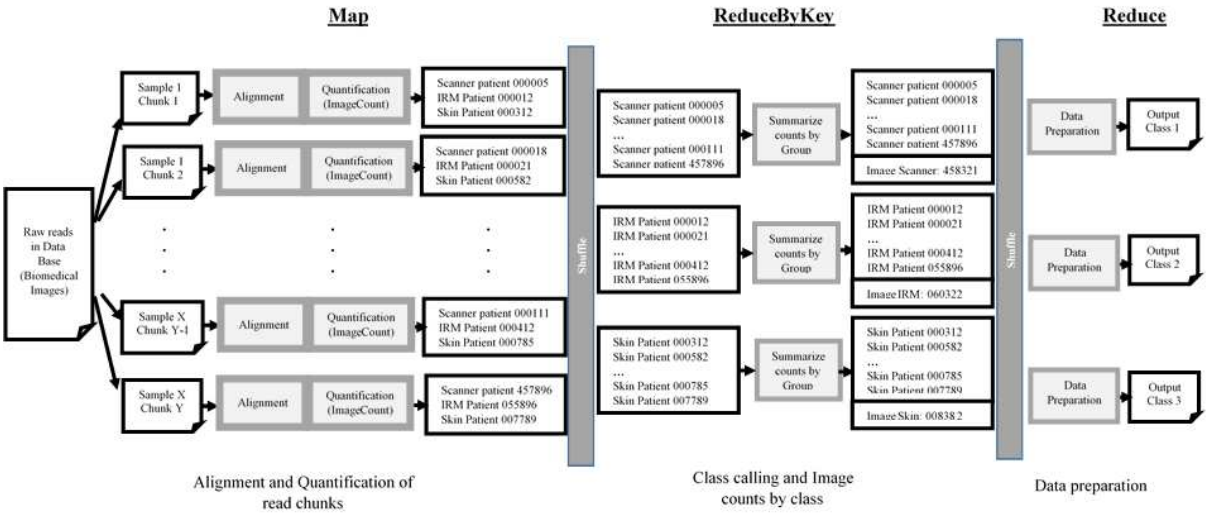


Figure 3: Tchagna et *al.*

Our reference: IMU 109

Article reference: IMU_2018_82

Article title: An Optimal Big Data Workflow for Biomedical Image Analysis

Caption of Fig4.a, Fig4.b is below

Figure 4: (a) General big data architecture to automate biomedical image analyses (b) Classification architecture to group images by category using Hadoop or Spark.