APPROACHES OF BIG DATA IN HEALTHCARE: A CRITICAL REVIEW

Faizan Ahmad
Dept. of Computer Science and Engineering
Integral University,
Lucknow, India.
faizanamd@iul.ac.in

Manish Madhav Tripathi
Department of CSE
Integral University
Lucknow, India
mmt@iul.ac.in

Abstract: The term big data refers to the large amount of data that desires new technologies and architectures to seek out valuable knowledge from it by using new and innovative analysis practices. As digitized medical records are currently utilized by most of the healthcare organization and pharmaceutical firms, they need started grouping and storing more and a lot of healthcare data in order to analyze it and obtain insights to solve the challenges of enormous and fast growing data bulks. This paper provides information concerning all the significant developments that have carried out so far within the field of big data analysis in the healthcare sector. This paper also covers key big data implementation challenges and big data solutions that attempt to solve the challenges of enormous and fast growing data bulks whereas reducing worth and notice its potential analytical value.

Keywords: big data; healthcare; big data analytics; medical image processing; signal processing

I. Introduction

The healthcare trade historically has generated large amount of data, driven by record keeping, compliance necessities, and patient care [1]. Whereas most data is hold on in hard copy form, the present trend is toward rapid digitization of those large amounts of data. Driven by obligatory requirements and therefore the potential to enhance the standard of health care delivery meanwhile reducing the prices, these huge quantities of data (known as ‘big data’) hold the promise of supporting a wide range of medical and healthcare functions, together with among others clinical call support, illness police work, and population health management [2-3].

Huge knowledge in healthcare refers to electronic health data sets thus massive and complicated that they’re difficult (or impossible) to manage with traditional software and/or hardware; nor will they be easily managed with traditional or common data management tools and methods [4]. Big data in healthcare is overwhelming not solely due to its volume however also because of the diversity of data varieties and therefore the speed at which it must be managed [4].

What exactly is big data? A report delivered to the U.S. Congress in August 2012 defines big data as “large volumes of high velocity, complex, and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management and analysis of the information” [5]. Big data encompasses such characteristics as variety, velocity and, with respect specifically to healthcare, veracity [6]. Existing analytical techniques can be applied to the vast amount of existing (but presently unanalyzed) patient-related health and medical data to reach a deeper understanding of outcomes that then will be applied at the point of care. Ideally, individual and population knowledge would inform every doc and her patient during the decision-making method and facilitate confirm the foremost appropriate treatment option for that particular patient. Doug Laney expressed the definition of big data as three V’s i.e.

- **Volume**: the amount of massive data or the quantity of generated and stored data.
- **Velocity**: denotes the speed of data at which it is produced and managed to fulfill the demands and challenges for its growth and development. Big data is mostly available in real-time. Activities like regular monitoring, e.g. of daily measurements of glucose of a diabetic patient, blood pressure and ECGs.
- **Variety**: The nature of the data as data comes in all types of formats. This helps people who analyse it to effectively use the resulting insight.

Big data Analytics have a widen role in handling the data which is generated from varied resources to supply quality information. Big data have characteristics like volume, velocity, selection and truthfulness. Since the data is increasing immensely day-by-day, big data not only defines the size but also finds insights from unstructured, complex, noisy, heterogeneous, longitudinal and voluminous data [1].

II. Advantages to health care

By digitizing, combining and effectively using big data, health care organizations ranging from single-physician offices and multi-provider groups to large hospital networks and accountable care organizations stand to realize significant advantages [2]. Potential benefits include detecting diseases at earlier stages when they can be treated more simply and effectively; managing specific individual and population health and detecting health care fraud more quickly and efficiently. Numerous queries may be addressed with big data analytics. certain developments or outcomes is also foreseen and/or calculable based on vast amounts of historical data, such as length of stay (LOS); patients who will choose elective surgery; patients who likely won't like surgery;
Comprehensive effectiveness analysis to see a lot of clinically relevant and cost-efficient ways in which to diagnose and treat patients.

Research & development:
- Predictive modelling to lower attrition and turn out a throw, faster, more targeted R & D pipeline in drugs and devices;
- Statistical tools and algorithms to enhance clinical trial style and patient recruitment to better match treatments to individual patients, so reducing trial failures and speeding new treatments to market;
- Analysing clinical trials and patient records to spot in the context of the problem of providing cohesive storage patient are also utilized during the diagnoses, prognosis, and treatment processes, then the problem of providing cohesive storage capacities if stored for long term. It also demands fast and accurate algorithms if any decision assisting automation were to be performed using the data.

In addition, if other sources of data acquired for each patient are also utilized during the diagnoses, prognosis, and treatment processes, then the problem of providing cohesive storage and developing efficient methods capable of encapsulating the broad range of data becomes a challenge.

Signal Processing: Similar to medical images, medical signals also pose volume and velocity obstacles especially during continuous, high-resolution acquisition and storage from a multitude of monitors connected to each patient. However, in addition to the data size issues, physiological signals also pose complexity of a spatiotemporal nature.

Analysis of physiological signals is often more meaningful when presented along with situational context awareness which needs to be embedded into the development of continuous monitoring and predictive systems to ensure its effectiveness and robustness. Currently healthcare systems use numerous disparate and continuous monitoring devices that utilize singular physiological waveform data or discretized vital information to provide alert mechanisms in case of overt events. However, such uncompounded approaches towards development and implementation of alarm systems tend to be unreliable and their sheer numbers could cause “alarm fatigue” for both care givers and patients [10]. In this setting, the ability to discover new medical knowledge is constrained by prior knowledge that has typically fallen short of maximally utilizing high-dimensional time series data. The reason that these alarm mechanisms tend to fail is primarily because these systems tend to rely on single sources of information while lacking context of the patients’ true physiological conditions from a broader and more comprehensive viewpoint. Therefore, there is a need to develop improved and more comprehensive approaches towards studying interactions and correlations among multimodal clinical time series data. This is important because studies continue to show that humans are poor in reasoning about changes affecting more than two signals [11].

Genomics. The cost to sequence the human genome (encompassing 30,000 to 35,000 genes) is rapidly decreasing with the development of high-throughput sequencing technology [12]. With implications for current public health policies and delivery of care analysing genome-scale data for developing actionable recommendations in a timely manner is a significant challenge to the field of computational biology. Cost and time to deliver recommendations are crucial in a clinical setting. Initiatives tackling this complex problem include tracking of 100,000 subjects over 20 to 30 years using the predictive, preventive, participatory, and personalized health, referred to as P4, medicine paradigm as well as an integrative personal omics profile. The P4 initiative is using a system approach for (i) analysing genome-scale datasets to determine disease states, (ii) moving towards blood based diagnostic tools for continuous monitoring of a subject, (iii) exploring new approaches to drug target discovery, developing tools to deal with big data challenges of capturing, validating, storing, mining, integrating, and finally (iv) modelling data for each individual. The integrative personal omics profile (iPOP) combines physiological monitoring and multiple high-throughput methods for genome sequencing to generate a detailed health and disease states of a subject [11]. Ultimately, realizing actionable recommendations at the clinical level remains a grand challenge for this field.

Utilizing such high density data for exploration, discovery, and clinical translation demands novel big data approaches and analytics.

1. Medical Image Processing from Big Data Point of View

The rapid growth in the number of healthcare organizations as well as the number of patients has resulted in the greater use of computer-aided medical diagnostics and decision support systems in clinical settings. Many areas in health care such as diagnosis, prognosis, and screening can be improved by utilizing computational intelligence. The integration of computer analysis with appropriate care has potential to help clinicians improve diagnostic accuracy [13]. The integration of medical images with other types of electronic health record (EHR) data and genomic data
can also improve the accuracy and reduce the time taken for a diagnosis. In the following, data produced by imaging techniques are reviewed and applications of medical imaging from a big data point of view are discussed.

1.1 Data Produced by Imaging Techniques
Medical imaging encompasses a wide spectrum of different image acquisition methodologies typically utilized for a variety of clinical applications. For example, visualizing blood vessel structure can be performed using magnetic resonance imaging (MRI), computed tomography (CT), ultrasound, and photoacoustic imaging. From a data dimension point of view, medical images might have 2, 3, and four dimensions. Positron emission tomography (PET), CT, 3D ultrasound, and functional MRI (fMRI) are considered as multidimensional medical data. Modern medical image technologies can produce high resolution images such as respiration-correlated or “four dimensional” computed tomography (4D CT) [14]. Higher resolution and dimensions of these images generate large volumes of data requiring high performance computing (HPC) and advanced analytical methods. For instance, microscopic scans of a human brain with high resolution can require 66TB of storage space. Although the volume and variety of medical data make its analysis a big challenge, advances in medical imaging could make individualized care more practical and provide quantitative information in variety of applications such as disease stratification, predictive modelling, and decision making systems. In the following we refer to two medical imaging techniques and one of their associated challenges.

1.2 Methods
The volume of medical images is growing exponentially. For instance, ImageCLEF medical image dataset contained around 66,000 images between 2005 and 2007 while just in the year of 2013 around 300,000 images were stored everyday. In addition to the growing volume of images, they differ in modality, resolution, dimension, and quality which introduce new challenges such as data integration and merging especially if multiple datasets are involved. Compared to the volume of research that exists on single modal medical image analysis, there is considerably lesser number of research initiatives on multimodal image analysis. When utilizing data at a local/institutional level, an important aspect of a research project is on how the developed system is evaluated and validated. Having annotated data or a structured method to annotate new data is a real challenge. This becomes even more challenging when large-scale data integration from multiple institutions are taken into account. As an example, for the same applications (e.g., traumatic brain injury) and the same modality (e.g., CT), different institutions might use different settings in image acquisitions which makes it hard to develop unified annotation or analytical methods for such data. In order to benefit the multimodal images and their integration with other medical data, new analytical methods with real-time feasibility and scalability are required. In the following we look at analytical methods that deal with some aspects of big data.

2. Medical Signal Analytics
Telemetry and physiological signal monitoring devices are ubiquitous. However, continuous data generated from these monitors have not been typically stored for more than a brief period of time, thereby neglecting extensive investigation into generated data. However, in the recent past, there has been an increase in the attempts towards utilizing telemetry and continuous physiological time series monitoring to improve patient care and management [15]. Streaming data analytics in healthcare can be defined as a systematic use of continuous waveform (signal varying against time) and related medical record information developed through applied analytical disciplines (e.g., statistical, quantitative, contextual, cognitive, and predictive) to drive decision making for patient care. The analytics workflow of real-time streaming waveforms in clinical settings can be broadly described using Figure 1. Firstly, a platform for streaming data acquisition and ingestion is required which has the bandwidth to handle multiple waveforms at different fidelities. Integrating these dynamic waveform data with static data from the EHR is a key component to provide situational and contextual awareness for the analytics engine. Enriching the data consumed by analytics not only makes the system more robust, but also helps balance the sensitivity and specificity of the predictive analytics. The specifics of the signal processing will largely depend on the type of disease cohort under investigation. A variety of signal processing mechanisms can be utilized to extract a multitude of target features which are then consumed by a retrained machine learning model to produce an actionable insight. These actionable insights could either be diagnostic, predictive, or prescriptive. These insights could further be designed to trigger other mechanisms such as alarms and notification to physicians. Harmonizing such continuous waveform data with discrete data from other sources for finding necessary patient information and conducting research towards development of next generation diagnoses and treatments can be a daunting task. For bed-side implementation of such systems in clinical environments, there are several technical considerations and requirements that need to be designed and implemented at system, analytic, and clinical levels. The following subsections provide an overview of different challenges and existing approaches in the development of monitoring systems that consume both high fidelity waveform data and discrete data from non-contiguous sources.

2.1 Data Acquisition
Historically streaming data from continuous physiological signal acquisition devices was rarely stored. Even if the option to store this data were available, the length of these data captures was typically short and downloaded only using proprietary software and data formats provided by the device manufacturers. Although most major medical device manufacturers are now taking steps to provide interfaces to access live streaming data from their devices, such data in motion very quickly poses archetypal big data challenges. The fact that there are also governance challenges such as lack of data protocols, lack of data standards, and data privacy issues is adding to this. On the other side there are many challenges within the healthcare systems such as network bandwidth, scalability, and cost that have staled the widespread adoption of such streaming data collection [16]. This has allowed way for system-wide projects which especially cater to medical research communities. Research community has interest in consuming data captured from live monitors for developing continuous monitoring technologies. There have been several indigenous and off-the-shelf efforts in developing and implementing systems that enable such data capture [17]. There are also products being developed in the industry that facilitate device manufacturer agnostic data acquisition from patient monitors across healthcare systems.

2.2 Data Storage and Retrieval
With large volumes of streaming data and other patient information that can be gathered from clinical settings, sophisticated storage mechanisms of such data are imperative. Since storing and retrieving can be computational and time expensive, it is key to have a storage infrastructure that facilitates rapid data pull and commits based on analytic demands. With its capability to store and compute large volumes of data, usage of systems such as Hadoop, MapReduce, and MongoDB is becoming much more common with the healthcare research communities. MongoDB is a free cross-platform document-oriented database which eschews traditional table-based relational database. Typically each health system has its own custom relational database schemas and data models which inhibit interoperability of healthcare data for multi-institutional data sharing or research studies. Furthermore, given the nature of traditional databases
integrating data of different types such as streaming waveforms and static EHR data is not feasible. This is where MongoDB and other document-based databases can provide high performance, high availability, and easy scalability for the healthcare data needs. Apache Hadoop is an open source framework that allows for the distributed processing of large datasets across clusters of computers using simple programming models. It is a highly scalable platform which provides a variety of computing modules such as MapReduce and Spark. For performing analytics on continuous telemetry waveforms, a module like Spark is especially useful since it provides capabilities to ingest and compute on streaming data along with machine learning and graphing tools. Such technologies allow researchers to utilize data for both real time as well as retrospective analysis, with the end goal to translate scientific discovery into applications for clinical settings in an effective manner.

2.3 Data Aggregation Integration of disparate sources of data, developing consistency within the data, standardization of data from similar sources, and improving the confidence in the data especially toward using analytics are some of the challenges facing data aggregation in healthcare systems [18]. Medical data can be complex in nature as well as being interconnected and interdependent; hence simplification of this complexity is important. Medical data is also subject to the highest level of scrutiny for privacy and provenance from governing bodies, therefore developing secure storage, access, and use of the data is very important. Analysis of continuous data heavily utilizes the information in time domain. However, static data does not always provide true time context and, hence, when combining the waveform data with static electronic health record data, the temporal nature of the time context during interation can also add significantly to the challenges. There are considerable efforts in compiling waveforms and other associated electronic medical information into one cohesive database that are made publicly available for researchers worldwide. For example, MIMIC II and some other datasets included in Physionet provide waveforms and other clinical data from a wide variety of actual patient cohorts.

2.4 Signal Analytics Using Big Data Research in signal processing for developing big data based clinical decision support systems (CDSSs) is getting more prevalent [19]. In fact, organizations such as the Institution of Medicine have long advocated use of health information technology including CDSS to improve care quality. CDSSs provide medical practitioners with knowledge and patient-specific information, intelligently filtered and presented at appropriate times, to improve the delivery of care. A vast amount of data in short periods of time is produced in intensive care units (ICU) where a large volume of physiological data is acquired from each patient. Hence, the potential for developing CDSS in an ICU environment has been recognized by many researchers. A scalable infrastructure for developing a patient care management system has been proposed which combines static data and stream data monitored from critically ill patients in the ICU for data mining and alerting medical staff of critical events in real time. Similarly, Bressan et al. developed an architecture specialized for a neonatal ICU which utilized streaming data from infusion pumps, EEG monitors, cerebral oxygenation monitors, and so forth to provide clinical decision support. A clinical trial is currently underway which extracts biomarkers through signal processing from heart and respiratory waveforms in real time to test whether maintaining stable heart rate and respiratory rate variability throughout the spontaneous breathing trials, administered to patients before extubation, may predict subsequent successful extubation. An animal study shows how acquisition of noninvasive continuous data such as tissue oxygenation, fluid content, and blood flow can be used as indicators of soft tissue healing in wound care. Electrocardiograph parameters from telemetry along with demographic information including medical history, ejection fraction, laboratory values, and medications have been used to develop an in-hospital early detection system for cardiac arrest.

A study presented by Lee and Mark uses the MIMIC II database to prompt therapeutic intervention to hypotensive episodes using cardiac and blood pressure time series data. Another study shows the use of physiological waveform data along with clinical data from the MIMIC II database for finding similarities among patients within the selected cohorts. This similarity can potentially help caregivers in the decision making process while utilizing outcomes and treatments knowledge gathered from similar disease cases from the past. A combination of multiple waveform information available in the MIMIC II database is utilized to develop early detection of cardiovascular instability in patients. Many types of physiological data captured in the operative and preoperative care settings and how analytics can consume these data to help continuously monitor the status of the patients during, before and after surgery, are described in. The potential of developing data fusion based machine learning models which utilizes biomarkers from breathomics (metabolomics study of exhaled air) as a diagnostic tool is demonstrated in [1].

3. Big Data Applications in Genomics

The advent of high-throughput sequencing methods has enabled researchers to study genetic markers over a wide range of population, improve efficiency by more than five orders of magnitude since sequencing of the human genome was completed, and associate genetic causes of the phenotype in disease states. Genome-wide analysis utilizing microarrays has been successful in analysing traits across a population and contributed successfully in treatments of complex diseases such as Crohn’s disease and age related muscular degeneration. Analytics of high-throughput sequencing techniques in genomics is an inherently big data problem as the human genome consists of 30,000 to 35,000 genes. Initiatives are currently being pursued over the timescale of years to integrate clinical data from the genomic level to the physiological level of a human being. These initiatives will help in delivering personalized care to each patient. Delivering recommendations in a clinical setting requires fast analysis of genome-scale big data in a reliable manner. This field is still in a nascent stage with applications in specific focus areas, such as cancer, because of cost, time, and labour intensive nature of analysing this big data problem. Big data applications in genomics cover a wide variety of topics. Here we focus on pathway analysis, in which functional effects of genes differentially expressed in an experiment or gene set of particular interest are analysed, and the reconstruction of networks, where the signals measured using high-throughput techniques are analysed to reconstruct underlying regulatory networks. These networks influence numerous cellular processes which affect the physiological state of a human being.

3.1 Pathway Analysis Resources for inferring functional effects for “omics” big data are largely based on statistical associations between observed gene expression changes and predicted functional effects. Experiment and analytical practices lead to error as well as batch effects. Interalization of functional effects has to incorporate continuous increases in available genomic data and corresponding annotation of genes. There are variety of tools, but no “gold standard” for functional pathway analysis of high-throughput genome-scale data. Three generations of methods used for pathway analysis are
described as follows. The first generation encompasses overrepresentation analysis approaches that determine the fraction of genes in a particular pathway found among the genes which are differentially expressed. Examples of the first generation tools are Onto-Express, GoMiner, and ClueGo. The second generation includes functional class scoring approaches which incorporate expression level changes in individual genes as well as functionally similar genes. GSEA is a popular tool that belongs to the second generation of pathway analysis. The third generation includes pathway topology based tools which are publicly available pathway knowledge databases with detailed information of gene products interactions: how specific gene products interact with each other and the location where they interact. Pathway-Express is an example of a third generation tool that combines the knowledge of differentially expressed genes with biologically meaningful changes on a given pathway to perform pathway analysis.

### 3.2 Reconstruction of Regulatory Networks

Pathway analysis approaches do not attempt to make sense of high-throughput big data in biology as arising from the integrated operation of a dynamical system. There are multiple approaches to analyzing genome-scale data using a dynamical system framework. Due to the breadth of the field, in this section we mainly focus on techniques to infer network models from biological big data. Applications developed for network inference in systems biology for big data applications can be split into two broad categories consisting of reconstruction of metabolic networks and gene regulatory networks. Various approaches of network inference vary in performance, and combining different approaches has shown to produce superior predictions. Available reconstructed metabolic networks include Recon 1, Recon 2, SEED, IOMA, and MADE. Recon 2 (an improvement over Recon 1) is a model to represent human metabolism and incorporates 7,440 reactions involving 5,063 metabolites. Recon 2 has been expanded to account for known drugs for drug target prediction studies and to study off-target effects of drugs. Reconstruction of gene regulatory networks from gene expression data is another well-developed field. Network inference methods can be split into five categories based on the underlying model in each case: regression, mutual information, correlation, Boolean regulatory networks, and other techniques. Over 30 inference techniques were assessed after DREAM5 challenge in 2010.

Table 1: Some popular methods and toolkits with their applications.

<table>
<thead>
<tr>
<th>Toolkit name</th>
<th>Category</th>
<th>Selected applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>GoMiner</td>
<td>Pathway analysis</td>
<td>Pancreatic cancer</td>
</tr>
<tr>
<td>Onto-Express</td>
<td>Pathway analysis</td>
<td>Breast cancer</td>
</tr>
<tr>
<td>MapReduce</td>
<td>Pathway analysis</td>
<td>MapReduce provides the interface for</td>
</tr>
<tr>
<td></td>
<td></td>
<td>the distribution of sub-tasks and the</td>
</tr>
<tr>
<td></td>
<td></td>
<td>gathering of outputs. When tasks are</td>
</tr>
<tr>
<td></td>
<td></td>
<td>executed, MapReduce tracks the processing</td>
</tr>
<tr>
<td></td>
<td></td>
<td>of each server/nodes.</td>
</tr>
<tr>
<td>ClueGo</td>
<td>Pathway analysis</td>
<td>Colorectal tumors</td>
</tr>
<tr>
<td>GSEA</td>
<td>Pathway analysis</td>
<td>GSEA Pathway analysis</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Diabetes</td>
</tr>
<tr>
<td>The Hadoop Distributed</td>
<td></td>
<td>HDFS enables the storage for the</td>
</tr>
<tr>
<td>File System (HDFS)</td>
<td></td>
<td>Hadoop cluster. It divides the data</td>
</tr>
<tr>
<td></td>
<td></td>
<td>into smaller parts and</td>
</tr>
<tr>
<td></td>
<td></td>
<td>distributes it across the various servers/</td>
</tr>
<tr>
<td></td>
<td></td>
<td>nodes.</td>
</tr>
<tr>
<td>Recon 2</td>
<td>Reconstruction</td>
<td>Drug target prediction studies</td>
</tr>
<tr>
<td></td>
<td>of metabolic</td>
<td></td>
</tr>
<tr>
<td></td>
<td>networks</td>
<td></td>
</tr>
<tr>
<td>HBase</td>
<td>Reconstruction</td>
<td>HBase is a column-oriented database</td>
</tr>
<tr>
<td></td>
<td>of gene regulatory networks</td>
<td>management system that sits on top of HDFS. It uses a non-SQL approach.</td>
</tr>
<tr>
<td>Boolean methods</td>
<td>Reconstruction</td>
<td>Cardiac Differentiation</td>
</tr>
<tr>
<td></td>
<td>of gene regulatory networks</td>
<td></td>
</tr>
<tr>
<td>ODE models</td>
<td>Reconstruction</td>
<td>Reconstruction of gene regulatory</td>
</tr>
<tr>
<td></td>
<td>of gene regulatory networks</td>
<td></td>
</tr>
<tr>
<td>Oozie</td>
<td></td>
<td>Oozie, an open source project, streamlinesthe workflow and coordination among the tasks.</td>
</tr>
</tbody>
</table>

### IV. Conclusion:

Big data analytics which leverages legions of disparate, structured, and unstructured data sources is going to play a vital role in how healthcare is practiced in the future. One can already see a spectrum of analytics being utilized, aiding in the decision making and performance of healthcare personnel and patients. Here we focused on three areas of interest: medical image analysis, physiological signal processing, and genomic data processing. Data face similar challenges and opportunities in dealing with disparate structured and unstructured big data sources. Although there are some very real challenges for signal processing of physiological data to deal with. Apart from the obvious need for further research in the area of data wrangling, aggregating, and harmonizing continuous and discrete medical data formats, there is also an equal need for developing novel signal processing techniques specialized towards physiological signals. Research pertaining to mining for biomarkers and clandestine patterns within bio signals to understand and predict disease cases has shown potential in providing actionable information. However, there are opportunities for developing algorithms to address data filtering, interpolation, transformation, feature extraction, feature selection, and so forth. Furthermore, with the notoriety and improvement of machine learning algorithms, there are opportunities in improving and developing robust CDSS for clinical prediction, prescription, and diagnostics [20].

### V. References:

4. L. A. Celi, R.G. Mark, D. J. Stone, and R.A. Montgomery, Big data in the intensive care unit:


