KNOWLEDGE – International Journal
Vol. 28.5
December, 2018

AUTOMATED MEDICAL DATA ANALYSES OF DISEASES USING BIG DATA

Lindita Loku
University Goce Delcev Stip, Republic of Macedonia lindita.loku@ugd.edu.mk
Bekim Fetaji
University Mother Teresa – Skopje, Republic of Macedonia bekim.fetaji@unt.edu.mk;
Aleksandar Krsteski
University Goce Delcev Stip, Republic of Macedonia aleksandar.krsteski@ugd.edu.mk;

Abstract: Diagnosis of different diseases is a growing concern and one of the most difficult challenges for modern medicine. Current diagnosis technologies (e.g. magnetic resonance imaging, electroencephalogram) produce huge quantity data (in size and dimension) for detection, monitoring and treatment of neurological diseases. In general, analysis of those medical big data is performed manually by experts to identify and understand the abnormalities. It is really difficult task for a person to accumulate, manage, analyse and assimilate such large volumes of data by visual inspection. As a result, the experts have been demanding computerised diagnosis systems, called "computer-aided diagnosis (CAD)" that can automatically detect the neurological abnormalities using the medical big data. This system improves consistency of diagnosis and increases the success of treatment, save lives and reduce cost and time. Recently, there are some research works performed in the development of the CAD systems for management of medical big data for diagnosis assessment. Such data analyses to realize diagnosis is very interesting for diabetes and autism. Many companies and research groups are working to treat diabetes, but preventing the disease will have a greater impact on health in at-risk groups. A team of US researchers are using data analytics to create a precision medicine approach to prevention of diabetes that steers efforts towards those who are at highest risk of developing the disease and who would benefit most from drug treatment or preventive lifestyle strategies. The analyses yielded most important 17 factors that were assessed that could predict an individual's risk of diabetes. Autism Spectrum Disorder (ASD) is characterized by difficulties in social communication, social interactions, and repetitive behaviors. It is diagnosed during the first three years of life. Early and intensive interventions have been shown to improve the developmental trajectory of the affected children. The earlier the diagnosis, the sooner the intervention therapy can begin, thus, making early diagnosis is set as our important research goal. Because ASD is not a neurodegenerative disorder, many of the core symptoms can improve as the individuals learn to cope with their environments under the right conditions. The earlier the age at which intervention can be started, the better their learning and daily function can be facilitated. Recent Big Data software packages and innovations in Artificial Intelligence have tremendous potential to assist with early diagnosis and improve intervention programs. The research study will focus on methodological evaluation of emerging technologies and will investigate by comparing different data sets and find a pattern that can be established as prognosis system. The research study investigated peer-reviewed studies in order to understand the current status of empirically-based evidence on the clinical applications in the diagnosis and treatment of Autism Spectrum Disorders (ASD). Also a survey and investigation on different sensing technologies for ASD like: eye trackers, movement trackers, electrodermal activity monitors, tactile sensors, vocal prosody and speech detectors. We assess their effectiveness and study their limitations. We also examine the challenges faced by this growing field that need to be addressed before these technologies can perform up to their theoretical potential In some cases, a technology is unable to deliver up to its potential, not due to the hardware but due to the inefficiency of the accompanied algorithms, as in the case of classifiers for repetitive behavior detection. Therefore, equal emphasis needs to be placed on the improvement of all aspects of a tracking technology. The nature of the sensors makes the tracked data very sensitive to experimental and systematic errors, often causing the collected data to be discarded due to unreliability. Efforts to reduce such inaccuracies can significantly improve the performance and potential of the overall technology. By collecting specific data, these sensors may be able to acquire objective measures that can be used to identify symptoms specific to ASD. The contribution of the analyses will assist not only the therapists and clinicians in their selection of suitable tools, but to also guide the developers of the technologies and devise new algorithm in prediction of autism.
Keywords: big data, analytical tools in medicine, big data tools and methods, effectiveness of data analytics

1. INTRODUCTION

Current modern technology like Big Data and Analytics is helping in the control and treatment of diseases such as diabetes, thalassemia, autism, etc. Worldwide, diabetes is a leading public health concern which affects over 422 million people and results in 1.5 million mortalities every year, says the World Health Organisation (WHO). In

India,it is growing at an alarming rate, with the country home to over 65.1 million people living with the disease. Due to the growing unstructured nature of Big Data form health industry, it is necessary to structure and emphasis its size into nominal value with possible solution.

Healthcare industry faces many challenges that make us to know the importance to develop the data analytics. Medical Apps powered by Big Data and Internet of Things can help doctors share best practices as well as treat patients internationally. The Apps generally combine live streaming and instant messaging and networking in a secure environment. Direct and double encrypted conversations among doctors can be achieved through these state-of-the-art Applications. Encryption ensures the privacy of interactions, even as a patient receives the best medical advice and care. Healthcare professionals using the Apps are put through a credentials check before being admitted into the system.

2. RESEARCH METHODOLOGY

The research methodology used was qualitative research and involved as method of collecting the data, literature review. In general, the literature search was done in two main stages. First, we have searched in the top 10 conferences and journals in data mining. Second, we have identified the toodp 10 cited papers in Google Scholar. A detailed description of the process is provided below.

To review the relevant big data analytics literature, we have decided to search using the term "Big Data Analytics", in the last 10 years (2008-2018), in peer-reviewed articles. Specifically, in the top 10 leading conferences and journals in data mining. As for the conference and journal ranking, we have adopted academic.research.micosoft.com ranking.. Among the top conferences in data mining are: ACM Conference on Knowledge Discovery and Data Mining (KDD), IEEE International Conference on Data Engineering (ICDE), International Conference on Information and Knowledge Management (CIKM), IEEE International Conference on Data Mining (ICDM), and SIAM International Conference on Data Mining (SDM-SIAM). Additionally, the top five journals in data mining are: IEEE Transactions on Knowledge and Data Engineering (TKDE), Information Processing Letters (IPL), The International Journal on Very Large Data Bases (VLDB), Data Mining & Knowledge Discovery (DATAMINE), and ACM's Special Interest Group on Knowledge Discovery and Data Mining Explorations (SIGKDD Explorations). Consequently, the search keyword together with the remaining search conditions have been used in two search databases; IEEE Xplore andACM DL. Furthermore, and to give the business dimension to big data analytics as well as avoid focusing only on the technical aspect of analytics, the same search term and conditions were applied to EBSCO research database. In EBSCO, two prime sources were selected: Business and Information.

3. LITERATURE REVIEW OF BIG DATA

Big data refers to datasets that are not only big, but also high in variety and velocity, which makes them difficult to handle using traditional tools and techniques. Due to the rapid growth of such data, solutions need to be studied and provided in order to handle and extract value and knowledge from these datasets [3].

Big data sizes are constantly increasing, currently ranging from a few dozen terabytes (TB) to many petabytes (PB) of data in a single data set. Consequently, some of the difficulties related to big data include capture, storage, search, sharing, analytics, and visualizing. Today, enterprises are exploring large volumes of highly detailed data so as to discover facts they didn't know before [4].

Big Data becomes very important, making possible to turn into this amount of data in information, knowledge, decision making and, ultimately, insights. A view of what are the Big Data has been exposed to Gartner that defines Big Data as high volume, velocity and variety information requires innovative way of information processing and to derive enhanced insights through data analytical tools, automation of process and effective decision making. Also it demands for the cost effective solution." [5],[6]. In fact the Big Data doesn't mean only the volume .Big Data characterized as 3Vs – Volume (data size and number), Velocity (the rate at which data are generated or need to the processed) and Variety (different types / different forms of content).

The volume of Big Data is expanding beyond terabytes into peta bytes and even exabytes (1 million TB). Varity refers to the data from different types of sources like sensors, devices, machines and unknown things. It means different data types, data formats, structured, semi structured and unstructured data. Speed of the production of data and to process to the data to generate valuable insights referred as Velocity. In fact the life of data can be very short and this may become obsolete after some time. So efficient usage of Big Data analytics results will bring good useful insights from high volume, variety of data. [7]. Sometimes quality of the data is a concern area because it fetches data from different applications for making decisions. Not all the data captured from various devices are

useful for making decisions. These are just information and the information has to be converted into knowledge for decision making..

Big Data are very large and complex that it is really tough and only with traditional approaches it is very difficult to process and analyze the data. Effective data management for Big Data sets is not possible with traditional RDBMS (Relational database management systems). Due to the size of Big Data it is very difficult to extract the information in a proper and required manner 10].

Bringing insights from the large amount of data is very much useful. In fact the raw input is the data that is processed into information. The data has no meaning when it is individual. But volume of data will provide some meaningful output and it provide trends and patterns. The converted knowledge will be the used for further analysis. The combination of knowledge and experience is Wisdom [9]. Conversion from raw data to valuable information is a challenge. To handle Big Data, Innovation in latest technologies and application of different techniques will help the individuals and organizations to collect the data, various analysis and visualize various formats of data in different industries and various domains. The aim of this chapter is to provide technical aspects of Big Data and Big Data challenges in Internet of things (IoT).

Hence, big data analytics is where advanced analytic techniques are applied on big data sets. Analytics based on large data samples reveals and leverages business change. However, the larger the set of data, the more difficult it becomes to manage [17]. Big data is data whose scale, distribution, diversity, and/or timeliness require the use of new technical architectures, analytics, and tools in order to enable insights that unlock new sources of business value. Three main features characterize big data: volume, variety, and velocity, or the three V's. The volume of the data is its size, and how enormous it is. Velocity refers to the rate with which data is changing, or how often it is created. Finally, variety includes the different formats and types of data, as well as the different kinds of uses and ways of analyzing the data [13].

Data volume is the primary attribute of big data. Big data can be quantified by size in TBs or PBs, as well as even the number of records, transactions, tables, or files. Additionally, one of the things that make big data really big is that it's coming from a greater variety of sources than ever before, including logs, clickstreams, and social media. Using these sources for analytics means that common structured data is now joined by unstructured data, such as text and human language, and semi-structured data, such as eXtensible Markup Language (XML) or Rich Site Summary (RSS) feeds. There's also data, which is hard to categorize since it comes from audio, video, and other devices. Furthermore, multi-dimensional data can be drawn from a data warehouse to add historic context to big data. Thus, with big data, variety is just as big as volume [14].

Moreover, big data can be described by its velocity or speed. This is basically the frequency of data generation or the frequency of data delivery. The leading edge of big data is streaming data, which is collected in real-time from the websites [16].

Some researchers and organizations have discussed the addition of a fourth V, or veracity. Veracity focuses on the quality of the data. This characterizes big data quality as good, bad, or undefined due to data inconsistency, incompleteness, ambiguity, latency, deception, and approximations [15].

4. LITERATURE REVIEW OF BIG DATA IN ANALYSES OF DISEASES IN MEDICINE

In recent years the healthcare industry has generated large amounts of data. The value based treatment in hospitals and digitization of world likes to have the computerized data rather than hard copy form. The health care data includes Electronic Health Reports (EHR) of patients data, clinical reports, doctor's prescription, diagnostic reports, medical images, pharmacy information, health insurance related data, data from social medias and medicinal journals [13]. All these information collectively forms Big Data in health care. By employing the analysis of big data will produce the predicted results for understanding the trends to improve the health care and life time expectancy, proper treatment at early stages at low cost. The analytics associated with big data is described by four characteristics: volume, velocity, variety and veracity [8]. The accumulation of health-related data continuously, diabetic Mellitus (DM) is one of the Non Communicable Diseases (NCD), is a major health hazard in developing countries such as India. The acute nature of DM is associated with long term complications and numerous of health disorders. There are three main types of this disease. Type1 DM results from the body's failure to produce insulin, and presently requires the person to inject insulin. This form is referred as Insulin - Dependent Diabetes Mellitus (IDDM). Type 2 DM results from insulin resistance, a condition in which cells fail to use insulin properly, sometimes combined with an absolute insulin deficiency. This form was previously referred to as Non-Insulin - Dependent Diabetes Mellitus (NIDDM). The third main form, gestational diabetes occurs when pregnant women without a previous diagnosis of diabetes develop a high blood glucose level. It may precede development of type 2

DM. It was estimated that 61.3 million people aged 20-79 years live with diabetes at 2011 in India. This number was expected to increase to 101.2 million by 2030.

There is a set of products that are designed specifically for individuals with autism and/or intellectual disabilities or can be adopted for this population. In this section, we provide a review of existing solutions listed for each category. The criteria for selection were: (1) devices that are designed specifically for individuals with ASD or ID, then (2) devices designed for general population but can be adopted for ASD or ID population in terms of offering the functionality that delivers the service. Selected devices are consumer electronic device or medically approved ones. Another more advanced solution called TouchPoints wristband and produced by TouchPoints Inc. [32]. This solution provides not only emotional monitoring but also claims to relieve stress using stimulating electrical pulses [44]. Other devices that can be worn include: BioHarness, Equivital Sensor Belt, Zepher belt and Hexoskin which is clinically validated to provide reliable ECG data. The last product listed is called AIO Sleeve, developed by Komodo Technologies but it is only a consumer device which does not provide clinical grade data.

Recently, smart clothing is becoming the new trend for wearable devices especially for physiological monitoring and emotional assessment as it provides a seamless experience for the users compared to wristbands which can be obtrusive to some users. An example of such solutions is Hexoskin Smart Shirt, developed by Hexoskin Inc. [20], which incorporates fabric sensors that collect: ECG, heart rate, heart rate variability, respiration rate, and body movement data.

5. LITERATURE REVIEW OF BIG DATA ANALYTICS TOOLS AND METHODS

With the evolution of technology and the increased multitudes of data flowing in and out of organizations daily, there has become a need for faster and more efficient ways of analyzing such data. Having piles of data on hand is no longer enough to make efficient decisions at the right time [19].

Such data sets can no longer be easily analyzed with traditional data management and analysis techniques and infrastructures. Therefore, there arises a need for new tools and methods specialized for big data analytics, as well as the required architectures for storing and managing such data. Accordingly, the emergence of big data has an effect on everything from the data itself and its collection, to the processing, to the final extracted decisions.

Consequently, [8] proposed the Big – Data, Analytics, and Decisions (B-DAD) framework which incorporates the big data analytics tools and methods into the decision making process [8]. The framework maps the different big data storage, management, and processing tools, analytics tools and methods, and visualization and evaluation tools to the different phases of the decision making process. Hence, the changes associated with big data analytics are reflected in three main areas: big data storage and architecture, data and analytics processing, and, finally, the big data analyses which can be applied for knowledge discovery and informed decision making. Each area will be further discussed in this section. However, since big data is still evolving as an important field of research, and new findings and tools are constantly developing, this section is not exhaustive of all the possibilities, and focuses on providing

a general idea, rather than a list of all potential opportunities and technologies.

Big Data Analytics (BDA) is the use of advanced techniques, mostly data mining and statistical, to find (hidden) patterns in (big) data. BDA is where advanced techniques operate on big data sets [7]. The term “Big Data” has recently been applied to datasets that grow so large that they become awkward to work with using traditional database management systems [8]. A significant amount of these techniques rely on commercial tools such as relational DBMS, data warehousing, ETL, OLAP, and business analytics tools. During the IEEE 2006 International Conference on Data Mining (ICDM), the top-ten data mining algorithms were defined based on expert nominations, citation counts, and a community survey. In order, those algorithms are: C4.5, k-means, SVM (support vector machine), Apriori, EM (expectation maximization), PageRank, AdaBoost, kNN (k-nearest neighbours), Naïve Bayes, and CART. They cover classification, clustering, regression, association analysis, and network analysis. Indeed, mostly, they have been incorporated in commercial as well as open source tools. Furthermore, multivariate analysis techniques such as regression, factor analysis, clustering, and discriminant analysis have been associated with many business applications [12].

The architecture of predictive analysis system includes various phases like data collection, data warehousing, predictive analysis, processing analyzed reports. Big Data Analysis Platforms And Tools are as follows:

1) Hadoop And MapReduce : This is one of the popularly used Big Data tool. Hadoop MapReduce is a Big Data programming model used for writing applications to process very huge amount of data in parallel on various clusters of commodity hardware in a reliable and fault tolerant manner. The scheduling, monitoring and re-execution of the failed tasks taken care by the master and the slave execute the tasks as per the direction of the master.

2) Gridgain: This is an alternative of Mapreduce and this also supports HDFS. This is used for fast analysis of real time data using in –memory processing.
3) Hpcc: It's expansion is High performance computing cluster. Both paid version and open source is available.
4) Storm: It works in many programming languages and owned by Twitter. It works under Linux operating system.

B. Data Bases / Warhouses

1) Apache Cassandra: This is another open source distributed database management system developed by Facebook. This is a high performance, scalability and high availability software. It has a good built in Cache.
2) Apache HBase: designed to run on the top of HDFS(Hadoop Distributed File system). It provides real time access to Hadoop and it provides distributed and scalable data set. It is modeled after Google's BigTable and it used Java for programming.
3) MongoDB [18]: MapReduce uses this for batch processing. It provides Query by field, Range and regular expression searches. It follows master slave model and the duplicate data is useful during hardware failure.
4) Neo4j [11]: It is a graph database model. Its speed is thousand times higher than Traditional DBMS. It works under REST interface or Java API.
5) Apache CouchDB [17]: It performs MapReduce queries thourgh JavaScript. It provides synchronization even in Smart Objects.
6) Terrastore [16]: This works in all the operating system. It is highly scalable and consistent.
7) FlockDB[16]: It is a graph oriented database and works in all Operting system
8) RIAK [16]: is another open source distributed key-value data store. It works with map/reduce, HTTP, REST and JSON.
9) Hypertable [19]: This is designed after Bigtable. It runs on the top of HDFS, GlusterFS, or the Kosmos File System (KFS). Its own querying language is HQL (Hypertable querying language)
10) Hive: Like Hypertable it uses its own querying language called HiveQL . This runs in all operating system. Hive is the Hadoop based data warehouse.

6. CONCLUSION

The research study was focused on investigating: medical data analyses of diseases using big data and we conducted a review on commercially off the-shelf wearable devices suitable for monitoring and tracking individuals with autism spectrum disorder and/or intellectual disability. Although the Big Data boom started few years ago, the opportunities are growing as the speed of data keeps growing. A global survey conducted by McKinsey on Big Data to understand the innovation, competition, and productivity. The survey covered Healthcare, Public sector, Retail, Manufacturing, Telecommunications. In depth study carried out with the help of existing literature reviews and various interview with industry executives. The research conducted in Economics and management. The research focused on Productivity, Competitiveness and growth .The evolution of global financial market and the economic impact of technology. Following key domains will have the great opportunities.

Large amount of data required for better analysis. Critical insights derived from clinical data will provide good care to the patients. Clinics can play a better role due to the availability of transparent and largely available information. Best practices to be deployed to meet the challenges and complete rethinking and change in IT structure required at the time of deployment. A big revolution is happening in the genetic field. New research direction arrived by Genome project. Big Data plays significant role in data storage, retrieval, sequence analysis and visualization.

A literature review reveals many results on diabetes carried out by different methods and materials of diabetes problem in India. Many people have developed various prediction models using data mining to predict diabetes.

Combination of classification-regression-genetic-neural network, handles the missing and outlier values in the diabetic data set, and also they replaced the missing values with domain of the corresponding attribute [1]. The classical neural network model is used for prediction, on the pre-processed dataset. In predictive analysis of diabetic treatment using regression based data mining techniques to diabetes data, they discover patterns using SVM algorithm that identify the best mode of treatment for diabetes across different age [2]. They concluded that drug treatment for patients in the young age group can be delayed whereas; patients in the old age group should be prescribed drug treatment immediately. Prediction and classification of various type of diabetes

The diabetes may associate with severe diseases such as heart attacks, strokes, eye diseases and kidney diseases, etc. Analyzing the risk value by the level of patient health condition using above results of can be used by the physicians at remote locations to serve the people. Detecting diseases at earlier stages can help to be treated more easily and effectively. In developing countries such as India, it is mandatory to manage specific individual and population health and detecting health care fraud more quickly. The middle-income families can be with the high availability of medical facility at minimum cost. This system leads to the improved focus on every individual patient health.

Thereby we can reduce and save our next generation from diabetic mellitus. Various Big Data Analysis Platforms and Tools Data bases / warhouses , Business Intelligence, Data Mining, File systems and Programming languages were discussed under Big Data technologies. Opportunities available in various domains discussed under Big Data Potentials. The biggest challenges in front of all the enterprises are the requirement of cultural and technological change to adopt the new technology. Valuable insights will be derived from available traditional data also. Organisational leaders should take the initiative to understand and move torwards the Big Data. Because it involves changes in all levels. Future research problems will promise the benefits of Big Data

REFERENCES

[1] V. H. Bhat, P. G. Rao, and P. D. Shenoy, "An Efficient Prediction Model for Diabetic Database Using Soft Computing Techniques," Architecture, Springer-Verlag Berlin Heidelberg, pp. 328-335, 2009.
[2] Abdullah A. Aljumah, Mohammed Gulam Ahamad, Mohammad Khubeb Siddiqui, "Application of data mining: Diabetes health care in young and old patients", Journal of King Saud University – Computer and Information Sciences, vol. 25, pp. 127–136, 2012.
[3] K. Rajesh, V. Sangeetha, "Application of Data Mining Methods and Techniques for Diabetes Diagnosis" in International Journal of Engineering and Innovative Technology (IJEIT) Vol 2(3), 2012.
[4] Sadhana, Savitha Shetty, "Analysis of Diabetic Data Set Using Hive and R", International Journal of Emerging Technology and Advanced Engineering, vol 4(7), 2014.
[5] Sabibullah M, Shanmugasundaram V, Raja Priya K, "Diabetes Patient's Risk through Soft Computing Model", International Journal of Emerging Trends & Technology in Computer Science, vol 2(6), 2013.
[6] 6. V. H. Bhat, P. G. Rao, S. Krishna, and P. D. Shenoy, "An Efficient Framework for Prediction in Healthcare," Most, Springer-Verlag Berlin Heidelberg , pp. 522-532, 2011.
[7] Nishchol Mishra, Dr.Sanjay Silakari, "Predictive Analytics: A Survey, Trends, Applications, Oppurtunities & Challenges", International Journal of Computer Science and Information Technologies, vol. 3(3), 4434-4438 4434, 2012.
[8] Wullianallur Raghupathi, and Viju Raghupathi, "Big data analytics in healthcare: promise and potential", Health Information Science and Systems, vol. 2(3) pp. 2-10, 2014.
[9] Mansoor Khan," Managing Vulnerable Populations: How Predictive Analytics is a Key Component for Understanding User Behavior and Improving Care Quality", Managed Care Outlook, vol 27(4), February 15, 2014.
[10] Andrew Pearson, Qualex Asia, "Predictive Analytics for the Healthcare Industry", Andrew Pearson, Qualex Asia Limited, 2012. http://www.intel.com/content/www/us/en/healthcare-it/bigger-data-better-healthcare-idc-insights-whitepaper.html
[11] D. Peter Augustine, "Leveraging Big Data analytics and Hadoop in Developing India's Health Care Services", International Journal of Computer Applications, vol 89(16), pp 44-50, 2014.
[12] Muni kumar N, Manjula R,"Role of Big Data Analytics in Rural Health Care – A Step Towards Svasth
[13] Bharath", International Journal of Computer Science and Information Technologies, vol 5(6), pp 7172-7178,
2014.
[14] Andre W. Kushniruk, "Predictive Analytics and Forecasting in Health Care: Integrating Analytics with Electronic Health Records", SAS Institute Inc, 2008.
[15] Radu-Ioan,Ciobanu,Valentin Cristea, Ciprian Dobre and Florin Pop, Big Data Platforms for the Internet of Things,2014,Springer
[16] Flavio Bonomi, Rodolfo Milito, Preethi Natarajan and Jiang Zhu,Fog Computing: A Platform for Internet of Things and Analytics, springer (2014)
[17] Shintaro Yamamoto, Shinsuke Matsumoto,Sachio Saiki, and Masahide Nakamura Kobe University, 1-1 Rokkodai-cho, Nada-ku, Kobe, Hyogo 657-8501, Japan, Using Materialized View as a Service of Scallop4SC for Smart City Application Services (2014)
[18] Mukherjee, A.; Datta, J.; Jorapur, R.; Singhvi, R.; Haloi, S.; Akram, W. "Shared disk big data analytics with Apache Hadoop" (18-22 Dec. 2012)
[19] Kudakwashe Zvarevashe1, Dr. A Vinaya Babu, Towards MapReduce Performance Optimization: A Look into the Optimization Techniques in Apache Hadoop for BigData Analytics (2014)
[20] K. C Dominick, N. O. Davis, J. Lainhart, H. Tager-Flusberg, and S Folstein, "Atypical behaviours in children with autism and children with a history of language impairment. Research in Developmental Disabilities,," Research in Developmental Disabilities, vol. 28, no. 2, pp. 145–162, 2007.