Chapter 17
Big Data-Based System: A Supportive Tool in Autism Spectrum Disorder Analysis

Tanu Wadhera
https://orcid.org/0000-0002-7646-2303
Dr. B. R. Ambedkar National Institute of Technology, Jalandhar, India

Deepti Kakkar
https://orcid.org/0000-0002-9681-1291
Dr. B. R. Ambedkar National Institute of Technology, Jalandhar, India

ABSTRACT

In the health domain, the move of generating big data is opening new methodologies in detection as well as prediction of various diseases and disorders. The first phase of the present chapter has provided insights into the role of big data analytics in the detection of one such neuro-disorder, that is, autism spectrum disorder (ASD). The data lake concept has provided a direction to resolve the issue by providing a common platform for storing tremendous amount of data in all formats (structured, unstructured, or raw). However, if the entire data have potential value, the data lakes need to be strategically designed as otherwise it can lead to data swamps. Therefore, in the second phase, data lake based on Hadoop architecture and Apache Spark engine has been provided for the analysis of the health data. The proposed system has resolved the data storage issue, management, and analytics on a single platform. Hence, the novelty of the chapter is that it is pointing towards the faster exploration as well as management of data so that the timely generation of hypothesis can help in analyzing ASD.

INTRODUCTION

Big data is the information hub with data collection from different sources and allows working on huge samples of the data that can determine the results even due to minute variations in the factors. This hub provides various options for pooling and exploring the different aspects of data in order to manage as well as predict the therapeutic outcomes. Hence, the data have unprecedented generation volume, veloc-

DOI: 10.4018/978-1-7998-3069-6.ch017

ity, and variety (Sarkar, 2017). The term 'volume' indicates the huge amount of data originating from sensors, recording machines, internet and healthcare data. 'Velocity' refers to generation, retrieval and speed of the data and 'variety' refers to the different types and formats of data that are generated such as structured, unstructured or real-time data and used for various purposes, such as big data analytics. The term veracity was added to big data for the healthcare domain referring to the trustworthiness, quality, accuracy and precision of the data (Feldman, Martin & Skotnes, 2012). The storage, processing, analysis and retrieval of such type of data need a more sophisticated environment and infrastructure and big data technology is providing an efficient platform for these data sets. The big data analytics approach takes the information collection concept to a completely different and new level and is being effectively used in a number of sectors such as information technology, customer care services, and risk management in fields such as business, astronomy and forecasting and in the healthcare sector (Andreu-Perez et al., 2015). In healthcare system, the patient data is heterogeneous, involves incomplete and imprecise observations derived from various sources such as diagnosis, treatment, injury, and mental impairments. It can be structured data such as signals, phenotypes, omics, International Classification of Diseases (ICD) codes and unstructured data such as clinical prescriptions, imaging, electronic health records and environmental factors (Cyganek et al., 2016, Dinov, 2016, Tanu &Kakkar, 2018a, Wang, Kung, & Byrd, 2018). In 2020, the worldwide digital healthcare data is expected to cross 25000 petabytes from 500 petabytes (2012 record) (Sun, J., & Reddy, 2013). In case of neurological disorders, especially ASD, the heterogeneity of the disorder demands that the data needs to be big not only in volume (sample size) but also in availability (open access) and depth (multiple amounts of data from the same individual). This database can be utilized in identification, prognosis, and diagnosis as well as in the treatment of the disorder. One of the studies provided an oracle-based computational model deploying big data for treating the ASD (Mani, Berkovich, & Liao, 2014). This model inserts new information into the brain of autistic individuals to understand the information-processing in their brain and improve their behavior.

Autism Spectrum Disorder (ASD) is a developmental disorder majorly targeting social communication and interaction domain and leading to restricted & repetitive behavior in individuals. In the world, almost 1-2% of the population suffers from ASD (Lombardo, Lai & Baron-Cohen, 2018). The seriousness of the disorder sounds from the percentage affecting the children worldwide. It is no more a western disease, as in India the number of ASD affected individuals is roughly 23 after 10,000 (Tanu and Kakkar, 2018b). The diagnostic age for the disorder is 2.5–3 years and at that time the disorder gets fully developed, leaving very minimum possibility of retracing the symptoms. These figures indicate that the experts/clinicians need to be vigilant in detecting and monitoring ASD suspected or affected individuals in order to reduce the diagnostic age. But, it is not possible, for the clinicians, to consider the entire amount of data, especially unstructured, to diagnose the individual properly at the proper time. There is a need to develop new strategies or systems that can diagnose the disorder and monitor health care effectively and cheaply (Weil, 2014). Moreover, to analyze the structured data such as brain signals or images, Machine Learning (ML) algorithms also demand a huge variety of data for training and testing. As if the training will be done on large datasets, only then the disorder can be tested efficiently. Since the disorder symptoms and deficits prevail unevenly, big data can make it possible to identify assistive personalized diagnostic and treatment methodologies (Raghupathi and Raghupathi, 2014). For big data based healthcare applications, the integration and exchange of the various heterogeneous datasets is an important requirement.