Chapter 8
Machine Learning and Deep Learning for Big Data Analysis

Tarun Pradhan
Manipal University Jaipur, India

Prathamesh Nimkar
Manipal University Jaipur, India

Kavita Jhajharia
https://orcid.org/0000-0002-6424-2127
Manipal University Jaipur, India

ABSTRACT

As data plays a role in machine learning and provides insights across various sectors, organizations are placing more emphasis on collecting, organizing, and managing information. However, traditional methods of analysing data struggle to keep up with the increasing complexity and volume of big data. To extract insights from datasets, advanced techniques like machine learning and deep learning have emerged. In the field of self-driving cars, analysing sensor data relies on methodologies developed from data analytics. These trends extend beyond cases; big data and deep learning are driving forces supported by enhanced processing capabilities and the expansion of networks. Managing the complexities involved in processing amounts of data requires scalable architectures that leverage distributed systems, parallel processing techniques and technologies such as GPUs. This development is particularly relevant for industries like banking, healthcare, and public safety, which have pressing demands, for transparency and interpretability in models.

1. INTRODUCTION

Big Data is a kind of data that is so vast and complex that it cannot be handled by standard systems or data-warehousing technology. Big data is unable to stored using a relational database management system or processed using standard SQL-style queries (Ishwarappa & Anuradha, 2015). Due to the develop-

DOI: 10.4018/979-8-3693-0413-6.ch008

ment of technology and services, massive amounts of structured, unstructured and semi-structured data have been generated from a variety of sources (Ishwarappa & Anuradha, 2015). Big data is originated from various sectors like agriculture, medical, IOT. In healthcare sector big data is produced by medical equipment, automated medical tools, such as sensor devices and high-throughput instruments, etc (Mohammed Alqahtani, 2023). By attaching IOT devices to fields, soil, and plants to collect data in real time directly from the ground, the agriculture industry creates big data.

Big Data can be identified using 7 Vs includes Velocity, Volume, Variety, Veracity, Variability, Visualization, and Value. Keep in mind that these Vs are provisional and could go up or down in the near future (Tyagi & G, 2019).

- " Volume" is a term used to refer to the quantity of Big Data. The IOT, medical devices, cloud computing traffic, and other variables all play a part in the rapid increase in data volume. Many big companies like Google, Apple etc. have a large amount of information or data in the form of logs (Ishwarappa & Anuradha, 2015).
- "Velocity" describes the rate with which data is gathered and the rate with which it is processed, stored, and evaluated by databases.
- "Variety" is a term that describes the range of data types. Different form of data can be gathered from numerous sources. Data might be structured, unstructured or semi-structured.
- "Variability" refers to the inconsistencies in the data that gets generated. This is mostly due to varied data sources, distinct data layouts, or errors in data filling.
- "Veracity" reflects on the accuracy as well as data quality. when coping with massive amount of data. It might not be entirely accurate or it might. Missing information may be present in the accumulating data. (Ishwarappa & Anuradha, 2015)
- "Visualisation" refers to the ability to convert data into pictorial representations that are easy to understand and analyse.
- "Value" is discussing the potential value of big data, which directly affects how organizations might use the information obtained.

Different sensor technologies used in various industries are producing a tremendous amount of data. But traditional data mining methods, however, were not powerful enough to manage huge data or uncover hidden patterns. As a result, Arthur Samuel invented the name "Machine Learning" (Weiss, 1992), and according to Samuel, " The discipline of machine learning equips computers with the ability to learn without being explicitly programmed" (Weiss, 1992). Classification and prediction are the core focuses of machine learning (ML) (Tyagi & G, 2019). Deep learning algorithms are very effective for learning from massive amounts of unsupervised data. Usually, they adopt a greedy layer-wise learning strategy to pick up data representations (Najafabadi et al., 2016). A Google-Brain Project member said, " The deep learning models might be thought of as a rocket engine, and the massive amounts of data we must feed these (exiting) algorithms as fuel (Tyagi & Rekha, 2020)." The Graphical Processing Unit (GPU) has emerged as a crucial component in the Deep Learning algorithm's execution. GPU is utilized to handle enormous amounts of data where parallel processing is used (Tyagi & Rekha, 2020). Deep learning has been used to improve the performance of classification models in some Big Data fields, like speech recognition and computer vision (Najafabadi et al., 2016). In the classification of biomedical images, Machine Learning and Deep Learning algorithms have achieved exceptional results. A wide range of procedures, including the structuring of databases of biomedical image to picture classifications before