# Big Data Analytics and Implementation Challenges of Machine Learning in Big data

**Rajeswari P[1,6], Sathishkumar V E[2] , Chunduru Anilkumar[3] , P Thilakaveni[4] and Usha Moorthy[5]**

[1] Department of Computer Engineering, PSG Polytechnic College, Peelamedu, Coimbatore, Tamilnadu- 641004
[2] Department of Industrial Engineering, Hanyang University, 222 Wangsimini-ro, Seongdong-gu, Seoul, Republic of Korea
[3] Department of Information Technology, GMR Institute of Technology, Srikakulam, AP-532127
[4] Department of Computer Engineering, PSG Polytechnic College, Peelamedu,, Coimbatore, Tamilnadu-641004.
[5] School of Computing and Information Technology, REVA University, Bengaluru, Karnataka, India

[6]srisathishkumarve@gmail.com

**Abstract.** The term "big data" refers to an information processing system that combines different conventional data techniques. Big data also includes a large amount of personally identifiable and authenticated data, making privacy a major concern. Various techniques have been developed to provide security and efficient data processing. Machine learning is a form of data technology that deals with one of the most important and least understood aspects of the data collected. Deep learning algorithms, similar to machine learning algorithms, learn programmers automatically from data and are thought to improve the efficiency and security of large data sets. The efficiency of machine learning and deep learning in a sensitive environment was evaluated in this paper, which reviewed security problems in big data. This paper begins by providing an overview of machine learning and deep learning algorithms. The research then moves on to machine learning problems and challenges, as well as potential solutions. The investigation into deep learning principles of big data continues after that. Finally, the report examines approaches used in recent research developments and concludes with recommendations for the future.

**Keywords:** Machine learning, Deep learning, big data, data mining, artificial intelligence, security issues.

## 1. Introduction

Bigdata creates a vital role in many applications for huge data storage and it was applicable for data related functionalities such as data prediction, data retrieval, and etc., In general, many existing models can be utilized for data analysis with traditional manner. But these have taken so much complexities of both time and space. To avoid such complexities with the help of advanced technologies like Artificial Intelligence (AI), Machine Learning (ML) as well as Deep learning (DL). The large amount of complex data could be solved by these advanced technologies. This article could help to get a clear idea of how those technologies are utilized. The roadmap of an entire analysis paper on analyzing the success of machine learning and deep learning algorithms is shown in Figure 1.

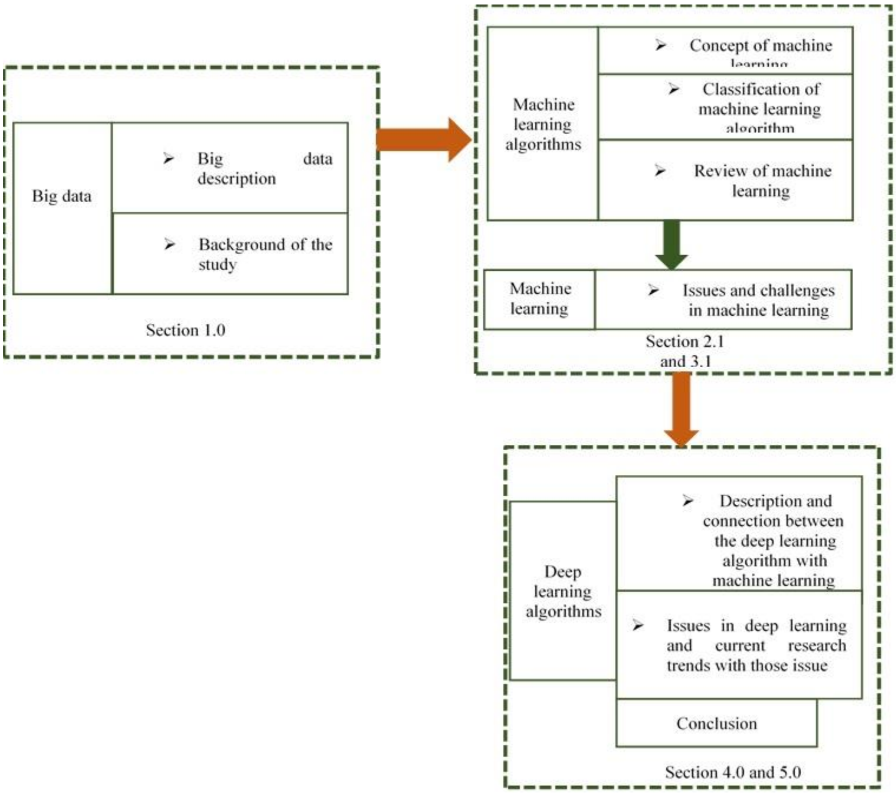
Big data
➢ Big data description
➢ Background of the study
Section 1.0
Machine learning algorithms
➢ Concept of machine learning
➢ Classification of machine learning algorithm
➢ Review of machine learning
Machine learning
➢ Issues and challenges in machine learning
Section 2.1 and 3.1
Deep learning algorithms
➢ Description and connection between the deep learning algorithm with machine learning
➢ Issues in deep learning and current research trends with those issue
Conclusion
Section 4.0 and 5.0

**Figure 1.** Roadmap of the present survey.

The following is the format of this survey paper: The Section I included a summary of big data and existing methods. The following is the remainder of the section: Section 2 summarizes key principles and classifications in machine learning algorithms, as well as studies performed by various scholars. The machine learning algorithm faces critical issues and challenges in Section 3. Section 4 contains a summary of deep learning algorithms as well as the relationship between machine learning and deep learning algorithms. Section 5 discusses the challenges in deep learning that are unique to big data, as well as the latest developments in these areas. Section 6 contains the conclusion.

## 2. Big data analytics

Despite the fact that using big data analytics to security issues is a vital guarantee, we must overcome a few challenges to fully comprehend its potential. New demands for sharing data among industry segments and to law enforcement agencies clash with the security guideline of avoiding data reuse that is, using data only for the purposes for which it was collected. Protection used to be heavily reliant on 75 mechanical limits on the ability to segregate, analyze, and link potentially sensitive datasets [1]. However, advances in big data analysis have provided us with instruments to concentrate and relate this data, making security breaches less demanding.

As a result, we should develop massive data apps while understanding security requirements and recommendations. Despite the fact that security control occurs in a few places on occasion, the Federal Communications Commission works with media communications organizations, the Health Insurance Administration, and other government agencies in the United States. The Human Services Data Portability and Accountability Act addresses human services data, Public Utility Commissions in a few states limit the use of sophisticated lattice data, and the Federal Trade Commission is developing rules for Web action. This movement has been broad in scope and open to interpretation for much of the time.

Even with security measures in place, we must recognize that the large scale accumulation and capacity of data makes these data stores appealing to a variety of groups, including industry (which will use our data for marketing and public relations), government (which will argue that this data is necessary for national security or legal requirements), and hoodlums (who might want to take our personalities). As huge data application draughtsman and planners, our role is to be proactive in putting up barriers to prevent misuse of these massive data warehouses. Another test is the question of data provenance.

Because large amounts of data allow us to expand the data sources we use for preparation, it's tough to ensure that each data source satisfies the level of dependability that our examination calculations want in order to produce precise results. To differentiate and alleviate the effects of vindictively embedded data, the study analyses thoughts from antagonistic machine learning and hearty insights.

**Table 1.** Data models and machine learning

| Machine Learning Technique | Data Type |
|---|---|
| Method of Changing Directions | Large Scale Data |
| Hypothesis for data combination for two-dimensional range heterogeneous data | Different Data Type |
| Keep forward neural networks alive. | High-speed streaming data |
| Knowledge discovery in databases (KDD) | Low-density data with a wide range of significance |

This research article is about to analyze the machine learning and deep learning approach for cloud computing environment. Table 1 shows the effectiveness of machine learning techniques for specific data types based on a study of current machine learning techniques. Through the examination of existing machine learning approaches various techniques like Sustain forward neural system, KDD, ADMM and data combination technique have been observed. The examination of articles demonstrates that KDD performs effectively for data with low-value density and density techniques, ADMM is suitable for large scale dataset and Keep forward neural networks alive is suitable for high-speed data streaming.

## 3. Implementation challenge of machine learning in big data

When applied to a large number of data sets, machine learning faces numerous challenges. Machine learning is challenging for large datasets with different data files, according to a study of existing studies [2]. This section offers a comprehensive overview of the machine learning implementation challenge for various data.

### *3.1. Large-scale data learning*

The amount of data is discovered to be the most essential element of big data, which poses a great machine learning challenge. Using **computerized** data as an example, Google alone must prepare approximately 24 petabytes of data on a regular basis. Allowing individuals to examine various data sources would also considerably enhance the size of the data. Data saved and dissected by large businesses will definitely reach the petabyte to exabyte range soon, based on existing growth patterns.

### *3.2. Different types of data may be learned*

The tremendous diversity of data is the second feature that makes large data both fascinating and overwhelming. This is because data comes from a multitude of sources and in many different formats.

Data sources that are structured, semi-structured, or entirely unstructured enable the age of diverse, high-dimensional, and nonlinear data with multiple representation frames. The vast test is obvious when studying with such a dataset, and the degree of complexity is not believed until we drill down into it.

### *3.3. Data streaming at high speeds necessitates learning*

When working with large volumes of data, rate or speed is critical, and this is becoming an increasingly  important test for learning. We need to complete an errand within a specific timeframe, for example, seismic tremor prediction, securities exchange forecast, and operator based self-ruling trade (purchasing/offering) structures, and so on; typically, the preparing results turn out to be less profitable or even useless. In these time-sensitive situations, the ability to estimate data is contingent on data freshness, which should be maintained indefinitely.

### *3.4. Uncertain and incomplete data learning*

Previously, machine learning algorithms were frequently rewarded with typically precise data from well-understood and constrained sources, resulting in consistently accurate learning outcomes; as a result, veracity was never an issue. However, with today's massive amounts of data, consistency and reliability of the source data soon become a problem, as data sources are frequently of varying origins and data quality is not always assured. As a result, we've added veracity as the fourth critical issue for big data learning, highlighting the need of dealing with and managing with data quality insecurity and inadequacy.

### *3.5. Data with a low density of values and a wide variety of meanings are good candidates for machine learning*

In fact, the final design uses a variety of learning strategies to examine massive datasets, focusing significant data from monstrous measures of data as profound understanding or business advantages. As a consequence, esteem is often cited as a distinguishing characteristic of big data. In any case, it's unclear how vital value can be extracted from vast volumes of data with a low esteem thickness. When dealing with criminal cases, for example, the police often need to review some reconnaissance recordings. Unfortunately, in some video streams, a few profitable data casings are occasionally blurred.

## 4. Other type of data processing using machine learning

### *4.1. Incremental learning for non-stationary data*

When an owner uses deep learning algorithms, he would know that there is one layer to extract features to decode the corrupted data. This layer is hidden. This helps in mapping and learning new features thus improving the generative and discriminative objective function [3]. You can also add features like over-fitting the data. The time when data is changed over time the massive data streams are also changed. This problem can quickly be solved by incremental feature extraction. This would help in avoiding expensive analysis involved in cross-validation while selection of large scales data sets.

### *4.2. High dimensional data*

For non-stationary data, incremental learning when a business owner employs deep learning algorithms, he is aware that one layer extracts features in order to decipher corrupted data. This layer has been turned off [4]. This aids in the mapping and learning of new features, thus enhancing the generative and discriminative objective functions. For non-stationary data, incremental learning when a business owner employs deep learning algorithms, he is aware that one layer extracts features in order to decipher corrupted data. This layer has been turned off. This aids in the mapping and learning of new features, thus enhancing the generative and discriminative objective functions.

### *4.3. Models at a large scale*

Deep Learning models are highly suited to managing the massive amounts of data associated with Big Data on a big scale, and they are frequently better at extracting complicated data designs from large datasets, as indicated in prior publications [5]. With Deep Learning for Big Data Analytics, identifying the ideal number of model parameters in such large-scale models and improving their computational reasonableness In addition to the difficulties of working with massive volumes of data, large-scale Deep Learning models for Big Data Analytics must cope with other Big Data issues such as region adjustment and gushing data. These advances underscore the need for more large-scale Deep Learning algorithms and architectures study.

**Table 2.** Deep Learning Techniques and data types.

| Deep Learning Technique | Data Type |
|---|---|
| Deep dynamic model (DDM) | High-dimensional data |
| Big scale Deep Learning models | Large scale data |

## 5. Implementation

Various problems related to big data analytics and issues with applying machine learning and deep learning algorithms have been examined in this report. We tried to introduce a solution for one of the problems in this analysis, which was an uncertain and incomplete dataset. To solve this problem, we used a machine learning algorithm to implement solution learning for unknown and incomplete data sets in Hadoop. Parkinson telemonitoring data set is the data set that we can adapt. The estimation of expression has demonstrated a significant increase in Parkinson's disease development. Around 90% of PWPs (people with Parkinson's disease) suffer from vocal deterioration. As a result, this dataset, which specifically focuses on discourse signals, was selected. The Parkinson Disease dataset for arrangement design was developed by Max Little of the University of Oxford in collaboration with the National Center for Voice and Speech in Denver, Colorado. This partnership picked up on the debate signals.

## 6. Conclusion

As of now, many traditional mechanisms implemented for providing a better solution in bigdata environment. But it faces so many challenges for a good performance. To overcome these challenges with ML and DL algorithms and it delivers a positive solution. Also, these algorithms reduce the crucial experimental computations without human presence. This analysis is focused to address the various challenges of big data and deliver a few points of research aspects in terms of big data with artificial technologies. With this learning, it could easy for analyze the huge data with less time complexity as well as less space complexity.

## References

[1] Qiu, J., Wu, Q., Ding, G., Xu, Y., & Feng, S. (2016). A survey of machine learning for big data processing. *EURASIP Journal on Advances in Signal Processing*, 2016(1), 1-16

[2] Al-Jarrah, O. Y., Yoo, P. D., Muhaidat, S., Karagiannidis, G. K., & Taha, K. (2015). Efficient machine learning for big data: A review. *Big Data Research*, **2(3),** 87-93.

[3] Sun, A. Y., & Scanlon, B. R. (2019). How can Big Data and machine learning benefit environment and water management: a survey of methods, applications, and future directions. *Environmental Research Letters*, **14(7),** 073001.

[4] López Belmonte, J., Segura-Robles, A., Moreno-Guerrero, A. J., & Parra-González, M. E. (2020). Machine learning and big data in the impact literature. A bibliometric review with scientific mapping in Web of science. *Symmetry*, **12(4),** 495.

[5] Zincir-Heywood, N., Casale, G., Carrera, D., Chen, L. Y., Dhamdhere, A., Inoue, T., ... & Samak, T. (2020). Guest Editorial: Special Section on Data Analytics and Machine Learning for

Network and Service Management–Part I. *IEEE Transactions on Network and Service Management*, **17(4),** 1971-1974.