Volume 3, Issue , 2024
PP 131-141

International Journal of Futuristic Innovation in Arts, Humanities and Management (IJFIAHM)

# CHALLENGES AND APPROACHES IN MACHINE LEARNING WITH BIG DATA

**1st Miss.Manisha Ghode**
manishaghode555@gmail.com
Student, Department of Computer Science & Engineering,
Shri Sai College of Engineering & Technology, Chandrapur, India.

**2nd Mr. Lowlesh Yadav**
lowlesh.yadav@gmail.com
Assistant Professor, Department of Computer Science & Engineering,
Shri Sai College of Engineering & Technology, Chandrapur, India.

**3rd Mr. Neehal Jiwane**
neehaljiwane@gmail.com
Assistant Professor, Department of Computer Science & Engineering,
Shri Sai College of Engineering & Technology, Chandrapur, India.

## ABSTRACT

The Big Data revolution promises to transfigure how we live, work, and suppose by enabling process optimization, empowering sapience discovery and perfecting decision timber. The consummation of this grand eventuality relies on the capability to prize value from similar massive data through data analytics; machine literacy is at its core because of its capability to learn from data and give data driven perceptivity, opinions, and prognostications. still, traditional machine learning approaches were developed in a different period, and therefore are grounded upon multiple hypotheticals, similar as the data set befitting entirely into memory, what unfortunately no longer holds true in this new environment. These broken hypotheticals, together with the Big Data characteristics, are creating obstacles for the traditional ways. Accordingly, this paper compiles, summarizes, and organizes machine literacy challenges with Big Data. In discrepancy to other exploration that discusses challenges, this work highlights the cause – effect relationship by organizing challenges according to Big Data Vs or confines that instigated the issue volume, haste, variety, or veracity. also, arising machine learning approaches and ways are bandied in terms of how they're able of handling the colorful challenges with the ultimate ideal of helping interpreters elect applicable results for their use cases. Eventually, a matrix relating the challenges and approaches is presented. Through this process, this paper provides a perspective on the sphere, identifies exploration gaps and openings, and provides a strong foundation and stimulant for farther exploration in the field of machine literacy with Big Data.

**Keyword:**Machine Learning,Big Data,Literacy, Challenges Approaches.

## 1.INTRODUCTION:

Today, the quantum of data is exploding at an unknown rate as a result of developments in Web technologies, social media, and mobile and seeing bias. For illustration, Twitter processes over 70M tweets per day, thereby generating over TB diurnal (1). ABI Research estimates that by 2020, there will be further than 30 billion connected bias (2). These Big Data retain tremendous eventuality in terms of business value in a variety of fields similar as health care, biology, transportation, online advertising, energy operation, and fiscal services (3), (4). still, traditional approaches are floundering when faced with these massive data. The conception of Big Data is defined by Gartner (5) as high volume, high haste, and/ or high variety data that bear new processing paradigms to enable sapience discovery, bettered decision timber, and process optimization. According to this description, Big Data aren't characterized by specific size criteria, but rather by the fact that traditional approaches are floundering to reuse them due to their size, haste or variety. The eventuality of Big Data is stressed by their description; still, consummation of this implicit depends on perfecting traditional approaches or developing new bonesable of handling similar data. Because of their eventuality, Big Data have been appertained to as a revolution that will transfigure how we live, work, and suppose (6). The main purpose of this revolution is to make use of large quantities of data to enable knowledge discovery and better decision timber (6). The capability to prize value from Big Data depends on data analytics; Jagadish etal. (7) consider analytics to be the core of the Big Data revolution. Data analytics involves colorful approaches, technologies, and tools similar as those from textbook analytics, business intelligence, data visualization, and statistical analysis. This paper focusses on machine literacy (ML) as an abecedarian com ponent of data analytics. The McKinsey Global Institute has stated that ML will be one of the main motorists of the Big Data revolution (8). The reason for this is its capability to learn from data and give data driven perceptivity, opinions, and prognostications (9). It's grounded on statistics and, also to statis tical analysis, can prize trends from data; still, it does not bear the unequivocal use of statistical attestations. According to the nature of the available data, the two main orders of learning tasks are supervised literacy when both inputs and their asked labors(markers) are known and the system learns to collude inputs to labors and unsupervised literacy when asked labors aren't known and the system itself discovers the structure within the data. Bracket and retrogression are exemplifications of supervised literacy in bracket the labors take separate values (class markers) while in regression the labors are nonstop. exemplifications of bracket algorithms

are k- nearest neighbor, logistic retrogression, and Support Vector Machine (SVM) while retrogression exemplifications include Support Vector Regression (SVR), direct retrogression, and polynomial retrogression. Some algorithms similar as neural networks can be used for both, bracket and regression. Unsupervised literacy includes clustering which groups objects grounded on established similarity criteria; k- means is an illustration of similar algorithm. Prophetic analytics relies on machine literacy to develop models erected using once data in an attempt to prognosticate the future (10); multitudinous algorithms including SVR, neural networks, and Naïve Bayes can be used for this purpose.

## 2.METHODOLOGY:

This paper highlights the challenges specific or largely relevant to machine literacy in the environment of Big Data, associates them with the V confines, and also provides an overview of how arising approaches are responding to them. In the being literature, some experimenters have described general machine learning challenges with Big Data (4), (14), (16), (17) whereas others have bandied them in the environment of specific methodologies (14), (18). Najafabadi etal. (14) concentrated on deep literacy, but noted the following general obstacles for machine literacy with Big Data unshaped data formats, presto moving(streaming) data,multi-source data input. also, Sukumar (16) linked three main bearmentsdesigning flexible and largely scalable infrastructures, understanding statistical data characteristics before applying algorithms; and eventually, developing capability to work with larger datasets. Both studies, Najafabadi etal. (14) and Sukumar (16) reviewed aspects of machine literacy with Big Data; still, they didn't essay to associate each identified challenge with its cause. also, their conversations are on a veritably high position without presenting affiliated results. In discrepancy, our work includes a thorough discussion of challenges, establishes their relations with Big Data confines, and presents an overview of results that Miti gate them. Qiu etal. (17) presented a check of machine literacy for Big Data, but they concentrated on the field of signal processing. Their study linked five critical issues (large scale, different data types, high speed of data, uncertain and deficient data, and data with low value viscosity) and related them to Big Data confines. Our study includes a more comprehensive view of challenges, but also relates them to the V confines. likewise, Qiu etal. (17) also linked colorful literacy ways and bandied representative work in signal processing for Big Data. Although they do a great work of identifying being problems and possible results, the lack of categorization and direct relationship between each approach and its

challenges makes it delicate to make an informed decision in terms of which literacy paradigm or result would be stylish for a specific use case or script. Accordingly, in our work emphasis is on establishing correlation between results and challenges. Al- Jarrah etal. (4) reviewed machine literacy for Big Data riveting on the effectiveness of large- scale systems and new algorithmic approaches with reduced memory footmark. Although they mentioned colorful Big Data hurdles, they did not present a methodical view as is done in this work. Al-Jarrah etal. were interested in the logical aspect, and styles for reducing computational complexity in dis tributed surroundings weren't considered. This work, on the other hand, considers both the logical aspect and computational complexity in distributed surroundings. Being studies have effectively bandied the obstacles encountered by specific ways similar as deep learning (14), (18). still, these studies riveted on a narrow aspect of machine literacy; a more comprehensive view of challenges and approaches in the Big Data environment is demanded. analogous to our work, Gandomi and Haider distributed challenges in agreement with the Big Data Vs (19). still, their characterization is general and not in terms of machine literacy. To understand the origin of machine literacy challenges, the present work categorizes them using the Big Data definition. In addition, colorful machine learning approaches are reviewed, and how each approach is able of addressing known challenges is bandied. This enables experimenters to make better informed decision regarding which learning paradigm or result to use grounded on the specific Big Data script. It also makes it possible to identify exploration gaps and openings in the sphere of machine literacy with Big Data. Accordingly, this work serves as a comprehensive foundation and facilitator for unborn exploration.

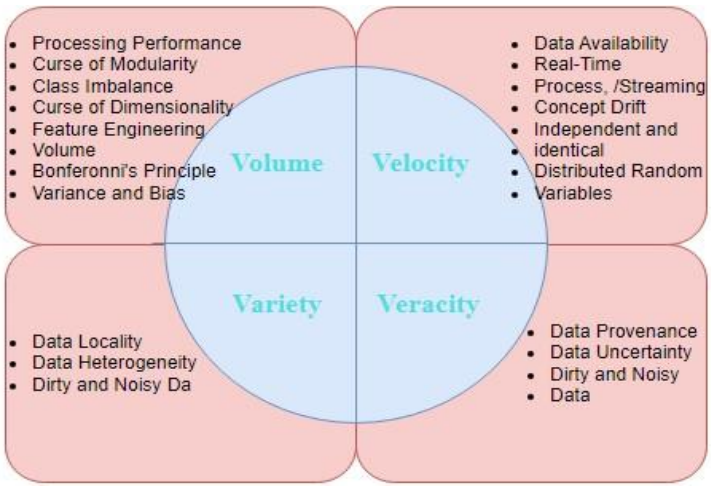


**Figure 1. Big Data Characteristics with Associated Challenges.**

## 3.MACHINE LITERACY CHALLENGES FORMING FROM BIG DATA DEFINITION:

Big Data are frequently described by its confines, which are appertained to as itsVs. before delineations of Big Data riveted on three Vs (24) (volume, haste, and variety); still, a more generally accepted description now relies upon the following four Vs (25) volume, haste, variety, and veracity. It's important to note that other Vs can also be set up in the literature. For illustration, value is frequently added as a 5th V (22), (26). still, value is defined as the asked outgrowth of Big Data processing (27) and not as defining characteristics of Big Data itself. For this reason, this paper considers only the four confines that characterize Big Data (28). This provides an occasion to relate challenges directly to the defining characteristics of Big Data, rendering the origin and cause of each explicitly. This section identifies machine literacy challenges and associates each challenge with a specific dimension of Big Data.Fig.1 illustrates the confines of Big Data along with their associated challenges as farther bandied in the followingsub-sections.

### 3.1. Volume:

The first and the most talked about specific of Big Data is volume it's the quantum, size, and scale of the data. In the machine literacy environment, size can be defined either vertically by the number of records or samples in a dataset or horizontally by the number of features or attributes it contains. likewise, volume is relative to the type of data a lower number of veritably complex data points may be considered original to a larger volume of simple data (19). This is maybe the easiest dimension of Big Data to define, but at the same time, it's the cause of multitudinous challenges. The followingsub-sections bandy machine literacy challenges caused by volume.

### 3.2. Variety:

The variety of Big Data describes not only the structural variation of a dataset and of the data types that it contains, but also the variety in what it represents, its semantic interpretation (7)

and its sources. Although not as numerous as for other V dimensions, the challenges associated with this dimension have substantial impact.

### 3.3. Velocity:

The haste dimension of Big Data refers not only to the speed at which data are generated, but also the rate at which they must be anatomized. With the omnipresence of smart phones and real- time detectors and the impending need to inter act snappily with our terrain through the development of technologies similar as smart homes, the haste of Big Data has come an important factor to consider.

### 3.4. Veracity:

The veracity of Big Data refers not only to the trust ability of the data forming a dataset, but also, as IBM has described, to the essential unreliability of data sources (19). The provenance and quality of Big Data together define the veracity component (62), but also pose a number of challenges as bandied in the followingsub-sections.

| APPROACHES | | | CHALLENGES | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | VOLUME | | | | | | | | VARIETY | | | VELOCITY | | | | VERACITY | | |
| | | | Processing Performance | Curse of Modularity | Class Imbalance | Curse of Dimensionality | Feature Engineering | Non-linearity | Bonferonni"s Principle | Variance and Bias | Data locality | Data Heterogeneity | Dirty and noisy Data | Data availability | Real-time Processing/Streaming | Concept drift | I.i.d | Data Provevance | Data Uncertainty | Dirty and Noisy Data |
| MANIPULATIONS | Data Manipulations | Dimensionality Reduction | ✓ | | | ✓ | | | | | | | | | | | | | | |
| | | Instance Selection | ✓ | ✓ | | | | | | | | | | | | | | | | |
| | | Data Cleaning | | | | | | | | | | | ✓ | | | | | | | ✓ |
| | Processing Manipulations | Vertical Scaling | ✓ | | | | | | | | | | | | | | | * | | |
| | | Hori-zontal Scaling: Batch-oriented | ✓ | ✓ | | * | | | | | ✓ | | | | | | | * | | |
| | | Hori-zontal Scaling: Stream-oriented | ✓ | ✓ | | | | | | | | | | ✓ | ✓ | | | * | | |
| | Algorithm Manipulations | Algorithm Modifications | ✓ | * | | * | | | | | ✓ | | | | ✓ | | | | | |
| | | Algorithm Mod. with new Paradigm | ✓ | * | | * | | | | | ✓ | | | | ✓ | | | | | |
| LEARNING PARADIGMS | Deep Learning | | | | | | ✓ | ✓ | | | | ✓ | * | | | | | | * | * |
| | Online Learning | | ✓ | ✓ | * | | | | | | ✓ | | * | ✓ | ✓ | * | ✓ | | | * |
| | Local Learning | | ✓ | ✓ | ✓ | | | | | ✓ | ✓ | | | | | | | | | |
| | Transfer Learning | | | | ✓ | | | | | | | ✓ | * | | | | | | * | * |
| | Lifelong Learning | | ✓ | | ✓ | | | | | | | ✓ | * | ✓ | ✓ | * | | | * | * |
| | Ensemble Learning | | ✓ | ✓ | | | | | | | | | | | | ✓ | | | | |

**Table 1. Machine Learning Approaches and The Challenges They Address.**

## 4. APPROACHES:

In response to the presented challenges, colorful approaches have been developed. Although designing entirely new algorithms would appear to be a possible result (68), experimenters have substantially preferred other styles. numerous approaches have been suggested and checks have been cantinalashed on specific orders of results; exemplifications include checks on platforms for Big Data analytics (20), (21) and review of data mining with Big Data (23). This paper reviews and organizes colorful proposed machine literacy approaches and discusses how they address the linked challenges.The big picture of approach-challenge correlations is presented in Table 1; it includes a list of approaches along with the challenges that each stylish address.

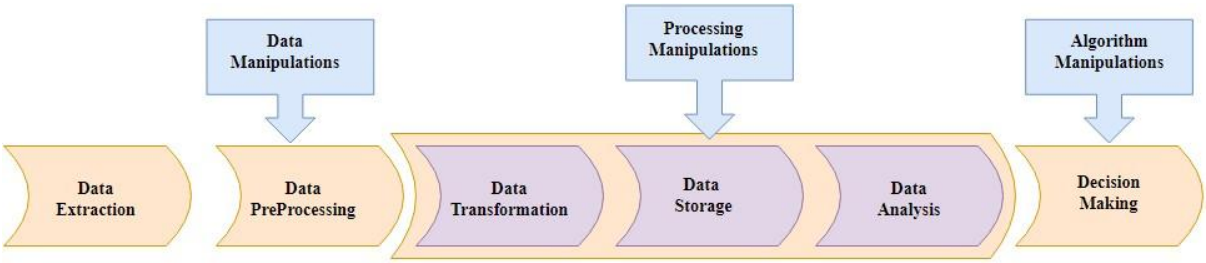


**Figure 2. Data Analytics Pipeline.**

Symbol ‘’ indicate high degree of remedy while ‘∗’ represents partial resolution. As it can be seen from the table, there are two main orders of results. The first order relies on data, processing, and algorithm manipulations to handle Big Data. The alternate order involves the creation and adaption of different machine learning paradigms and the revision of being algorithms for these paradigms. In addition to these two orders, it's important to note several machine literacies as a service immolations Microsoft Azure Machine Learning, now part of Cortana Intelligence Suite (69); Google Cloud Machine Learning Platform (70); Amazon Machine Learning (71); and IBM Watson Analytics (72). Because these services are backed up by important pall providers, they offer not only scalability but also integration with other pall platform services. still, at the moment, they support a limited number of algorithms compared to the R language (73), MATLAB (74), or Weka (75). also, calculation happens on pall coffers, which requires data transfer to remote bumps. With Big Data, this results in high network business and may indeed come infeasible due to time or bandwidth conditions. Because these ML services are personal, information about their beginning technologies is veritably limited; thus, this paper doesn't bandy them further. The followingsub-sections

introduce ways and methodologies being developed and used to handle the challenges associated with machine literacy with Big Data. First, manipulation ways used in confluence with being algorithms are presented. Second, colorful machine literacy paradigms that are especially well suited to handle Big Data challenges are bandied.

## 5. CONCLUSION:

This paper has handed a methodical review of the challenges associated with machine literacy in the environment of Big Data and distributed them according to the V dimensions of Big Data. also, it has presented an overview of ML approaches and bandied how these ways overcome the colorful challenges linked. The use of the Big Data description to classify the challenges of machine literacy enables the creation of cause effect connections for each of the issues. likewise, the creation of unequivocal relations between approaches and challenges enables a more thorough understanding of ML with Big Data. This fulfills the first ideal of this work; to produce a foundation for a deeper understanding of machine literacy with Big Data. Another ideal of this study was to give experimenters with a strong foundation for making easier and better-informed choices with regard to machine literacy with Big Data. This ideal was achieved by developing a comprehensive matrix that lays out the connections between the colorful challenges and machine literacy approaches, thereby pressing the stylish choices given a set of conditions. This paper enables the creation of connections among the colorful issues and results in this field of study, which wasn't fluently possible on the base of the being literature. From the development or adaption of new machine learning paradigms to attack undetermined challenges, to the combi nation of being results to achieve farther performance advancements, this paper has linked exploration opportunities. This work has thus fulfilled its last ideal by furnishing the academic community with implicit directions for unborn work and will hopefully serve as root for great advancements in the field of machine literacy with Big Data.

## REFERENCES

[1] R. Krikorian. (2010). Twitter by the Numbers, Twitter. [Online].Available: http://www.slideshare.net/raffikrikorian/twitter-by-the-numbers?

ref=http://techcrunch.com/2010/09/17/twitter-seeing-6-billion-api-callsper-day-70k-per-second/

[2] ABI. (2013). Billion Devices Will Wirelessly Connect to the Internet of Everything in 2020, ABI Research. [Online]. Available: https://www.abiresearch.com/press/more-than-30-billion-devices-willwirelessly-conne/

[3] W. Raghupathi and V. Raghupathi, ‘‘Big data analytics in healthcare: Promise and potential,’’ Health Inf. Sci. Syst., vol. 2, no. 1, pp. 1–10, 2014.

[4] O. Y. Al-Jarrah, P. D. Yoo, S. Muhaidat, G. K. Karagiannidis, and K. Taha, ‘‘Efficient machine learning for big data: A review,’’ Big Data Res., vol. 2, no. 3, pp. 87–93, Sep. 2015.

[5] M. A. Beyer and D. Laney, ‘‘The importance of ‘big data’: A definition,’’ Gartner Research, Stamford, CT, USA, Tech. Rep. G00235055, 2012.

[6] V. Mayer-Schönberger and K. Cukier, Big Data: A Revolution That Will Transform How We Live, Work, and Think. Boston, MA: Houghton Mifflin Harcourt, 2013.

[7] H. V. Jagadish et al., ‘‘Big data and its technical challenges,’’ Commun. ACM, vol. 57, no. 7, pp. 86–94, 2014.

[8] M. James, C. Michael, B. Brad, and B. Jacques, Big Data: The Next Frontier for Innovation, Competition, and Productivity. New York, NY: McKinsey Global Institute, 2011.

[9] M. Rouse. (2011). Machine Learning Definition. [Online].Available: http://whatis.techtarget.com/definition/machine-learning

[10] M. Rouse. (2009). Predictive Analytics Definition. [Online].Available: http://searchcrm.techtarget.com/definition/predictive-analytics

[11] K. Grolinger, M. Hayes, W. A. Higashino, A. L’Heureux, D. S. Allison, and M. A. M. Capretz, ‘‘Challenges for MapReduce in big data,’’ in Proc. IEEE World Congr. Services (SERVICES), Jun. 2014, pp. 182–189.

[12] J. Dean and S. Ghemawat, ‘‘MapReduce: Simplified data processing on large clusters,’’ in Proc. 6th Symp. Oper. Syst. Design Implement., 2004, pp. 137–149.

[13] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, ‘‘The hadoop distributed file system,’’ in Proc. IEEE 26th Symp. Mass Storage Syst. Technol. (MSST), May 2010, pp. 1–10.

[14] M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald, and E. Muharemagic, ‘‘Deep Learning Applications and Challenges in Big Data Analytics,’’ J. Big Data, vol. 2, no. 1, p. 1, Feb. 2015.

[15] C. Parker, ‘‘Unexpected challenges in large scale machine learning,’’ in Proc. 1st Int. Workshop Big Data, Streams Heterogeneous Source Mining Algorithms, Syst., Programm. Models Appl. (BigMine), 2012, pp. 1–6.

[16] S. R. Sukumar, ‘‘Machine learning in the big data era: Are we there yet?’’ in Proc. 20th ACM SIGKDD Conf. Knowl. Discovery Data Mining, Workshop Data Sci. Social Good (KDD), 2014, pp. 1–5.

[17] J. Qiu, Q. Wu, G. Ding, Y. Xu, and S. Feng, ‘‘A survey of machine learning for big data processing,’’ EURASIP J. Adv. Signal Process., vol. 67, pp. 1–16, Dec. 2016.

[18] X.-W. Chen and X. Lin, ‘‘Big data deep learning: Challenges and perspectives,’’ IEEE Access, vol. 2, pp. 514–525, 2014.

[19] A. Gandomi and M. Haider, ‘‘Beyond the hype: Big data concepts, methods, and analytics,’’ Int. J. Inf. Manage., vol. 35, no. 2, pp. 137–144, Apr. 2015.

[20] D. Singh and C. K. Reddy, ‘‘A survey on platforms for big data analytics,’’ J. Big Data, vol. 2, no. 1, pp. 1–20, 2015.

[21] P. D. C. de Almeida and J. Bernardino, ‘‘Big data open source platforms,’’ in Proc. IEEE Int. Congr. Big Data, Jun. 2015, pp. 268–275.

[22] W. Fan and A. Bifet, ‘‘Mining big data: Current status, and forecast to the future,’’ SIGKDD Explorations Newslett., vol. 14, no. 2, pp. 1–5, Dec. 2012.

[23] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, ‘‘Data mining with big data,’’ IEEE Trans. Knowl. Data Eng., vol. 26, no. 1, pp. 97–107, Jan. 2014.

[24] R. Narasimhan and T. Bhuvaneshwari, ‘‘Big data—A brief study,’’ Int. J. Sci. Eng. Res., vol. 5, no. 9, pp. 350–353, 2014.

[25] Lowlesh Nandkishor Yadav, “Predictive Acknowledgement using TRE System to reduce cost and Bandwidth”IJRECE VOL. 7 ISSUE 1 (JANUARY- MARCH 2019) pg no 275-278.