Big Data Challenges and Learning Paradigms: A Review

S. L SWAPNA[1*]and V. SARAVANAN[2]

[1]Research Scholar, Hindusthan College of Arts and Science (Autonomous), Coimbatore, India
[2]Professor, Hindusthan College of Arts and Science (Autonomous), Coimbatore, India
[1]swapnamartin2003@gmail.com
[2]vsreesaran@gmail.com

Abstract: *Big data is one of the impacts of information revolution due to technological advancements such as communication, mobile and cloud services. The uncontrolled accumulation of structured and unstructured enormous volumes of data creates challenges in storing and manipulating data and obtaining valuable insights from these data. Big Data Analytics is progressively becoming popular and the organizations are in forefront to devise and adopt diversified approaches including machine learning for Big Data Analytics. Business organizations are using data learning as a scientific method for dealing with big data. The use of appropriate data analytics tools is crucial for the organizations to withstand in their business, to face the challenges in the market and gain out of competitive advantage. By considering the overwhelming demand on the data analytics tools, this review paper presents the comprehensive view on various Big Data Analytics methods in place and the state-of-the-art approaches towards Big Data Analytics. This paper also presents upcoming challenges towards big data and suggests certain mechanisms to thwart those challenges.*

Keywords: Big Data, Big Data analytics, Machine Learning

1. INTRODUCTION

A sharp increase in the rate of exploding data is unprecedented at par with the advancements in social interfaces, mobile and sensing technologies. It is estimated that twitter possesses around 100M tweets per day and generating more than 10TB per day [1]. According to ABI Research, it is projected that there will be 30 billion connected devices by 2020 [2]. Big Data have power to hold enormous business potential in various areas such as finance and insurance, retail, logistics, health care and bioinformatics, travel, advertising, energy services and so on [3], [4]. On the other hand, firms rely on conventional approaches are facing difficulties to process these big data and owing to their size, velocity or variety. Data Analytics has various approaches and methods to extract insights from Big Data [5] and considered to be the centre of data rebellion. Data analytics tools at large in place today such business intelligence, text analytics, visualization and image analytics and statistical analysis.

The centre of focus of this paper is data learning as it is in forefront for handling Big Data. It is evident that machine learning (ML) is one of the major driving factors of the Big Data Analytics (BDA) today [6] as it has the ability to learn from data without explicit programming and do predictions [7]. The two classes of learning based on the nature of the data namely supervised and unsupervised. Classification and regression under the class of supervised learning. In classification, the algorithm takes discrete values as class labels whereas in regression, the outputs are continuous. Wide range of classification type algorithms are in place like k-nearest neighbour (kNN), logistic

**Corresponding Author*

regression, Decision Tree (DT) and Support Vector Machine (SVM) whilst regression type comprise linear regression, polynomial regression and Support Vector Regression. Few neural networks algorithms are also been used for the problems of classification and regression. Clustering algorithms such as k-means are unsupervised which creates clusters of entities anchored in similarity index. Predictive analytics is a concept relying on machine learning which creates functional models built using historical data in an effort to predict future scenarios [8]. In ML, algorithms are used to train the models with considerably more data and give better results [9]. However, extensive use of these huge datasets creates number of problems as the conventional algorithms not capable of handling data at large. Many ML algorithms are designed to function with an assumption that the entire dataset is made available for processing at the time of training the model. The increasingly evolving big data break these assumptions and the conventional ML algorithms are made unusable and their performance become uncertain. Various techniques were developed to adapt machine learning algorithms to work with large datasets such as MapReduce [10] and distributed processing frameworks such as Hadoop [11]. Sub domains of machine learning including deep and online learning are also adapted in an effort to overcome the challenges of data learning with Big Data.

This paper summarizes data learning challenges with Big Data and the focus is on linking the identified challenges with the Big Data V dimensions: volume, velocity, variety, and veracity. Secondly, latest machine learning approaches are analysed with the emphasis on how they address the identified challenges and provides a perspective on the domain and identifies pit falls and future prospects in the area of machine learning with Big Data.

2. RELATED WORK

This review overlooks the challenges relevant to machine learning in the context of Big Data, and the V dimensions, and then provides an overview of how emerging approaches are responding to them. On looking at the existing research insights, few researchers have described general machine learning challenges with Big Data [4], [12], [13], [14] whereas some others have discussed them in the context of specific methodologies [12], [15]. Najafabadi et al. [12] noted the following issues for machine learning with Big Data: unstructured data formats, fast moving (streaming) data, multi-source data input, noisy and poor-quality data, high dimensionality, scalability of algorithms, imbalanced distribution of input data, unlabelled data, and limited labeled data. Also, Sukumar [13] acknowledged three requirements: designing flexible and highly scalable architectures, understanding statistical data characteristics before applying algorithms; and developing ability to work with larger datasets. Qiu et al. [14] given a survey of machine learning for Big Data Analytics and focused on the field of signal processing. Both studies identified five critical issues such as large scale, different data types, high speed of data, uncertain and incomplete data, and data with low value density; and related them to Big Data dimensions. Likewise, John Martin et al.[16] also brought another work and presented challenges in feature analysis of signal processing. Al-Jarrah et al. [4] approached machine learning for Big Data focussing on the efficiency of large-scale systems and new algorithmic approaches with reduced memory footprint. Even though they mentioned various Big Data challenges, they fail to present a systematic view of this work. Many reviews interested in the analytical aspect, and nothing is found in their researches for reducing computational complexity were not considered. Our work, on the other hand, presents both the analytical and computational aspects in distributed environments. Existing literatures effectively discussed the challenges faced by specific techniques such as machine learning and deep learning [12], [15]. However, those reviews focussed a very narrow approach of machine learning and a more comprehensive view of the challenges in the Big Data context is needed. Various big data Analytics platforms are presented in [17] and [18]. Also, another review [21] revealed the use of open source platforms including Apache Mahout, massive online analysis (MOA), the R Project, Vowpal, Pegasos, and GraphLab. All these studies examined and compared existing platforms, but the present study relates these platforms to the issues they address. The limitations of data mining with Big Data have been exposed in the articles [19] and [20]. Fan and Bifet [19] concentrates on challenges and not classify those challenges nor provide possible solutions. According to Wu et al. [20], the challenges are categorized into three tiers namely tier I -big data mining platforms, tier II -Semantics and application knowledge, and Tier III -Big Data mining algorithms.

In order to understand the causes of machine learning issues, this study categorizes them using the Big Data definition. Furthermore, different machine learning approaches and models are studied, and how each one of them is able to address these challenges. This will be a valuable source of knowledge enables researchers to make better informed decision regarding which machine learning paradigm or solution to adopt for the Big Data scenario. In addition, it makes possible to identify and reduce the research gaps and facilitate for further research prospects.

3. BIG DATA DIMENSIONS

Big Data are generally described by its dimensions: volume, velocity, variety, and veracity. This part of this paper briefs machine learning challenges and relate each challenge with a specific dimension of Big Data as illustrated in fig.1. This will open discussion among the researchers on big data challenges by bringing under these four key dimensions.

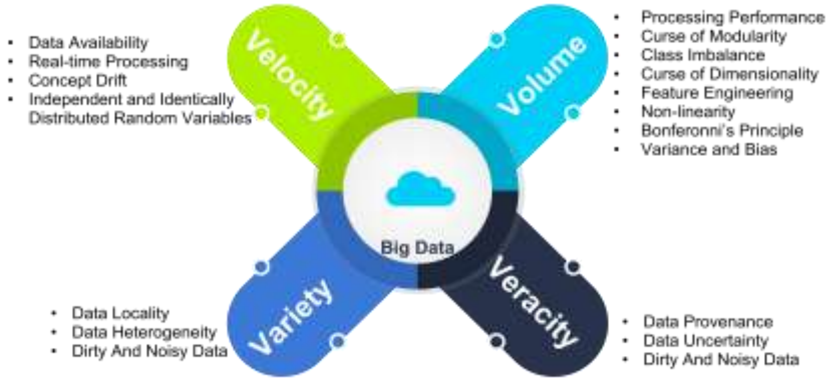

Figure 1. Characteristics of Big Data and challenges

Volume: Volume is the most significant characteristic of Big Data and is characterized by amount, size, and scale of the data. In machine learning, the size will be defined either vertically by the number of samples in a dataset or horizontally by the number of features or attributes. Also, volume is relative to the type of data: a smaller number of very complex data points may be considered equivalent to a larger quantity of simple data [21]. This is the simplest dimension of Big Data to represent; however, data volume is the major cause of several challenges.

Variety: It shows the semantic representation [5] of Big Data. Variety describes not only the structural variation of a dataset, but also the variety in what it represents. The challenges related to this dimension have significant impact on Big Data.

Velocity: This dimension refers the speed at which data are generated and the rate at which they must be analyzed. The various properties associated with velocity include data availability, real-time processing, concept drift and independent and identically distributed random variables.

Veracity: The veracity of Big Data is characterized by incompleteness, uncertainty, and the reliability of the datasets, and also the unreliability of data sources [21]. These characteristics also cause a number of challenges in handling Big Data.

4. HANDLING OF BIG DATA

In response to the challenges, various approaches have been proposed by the researchers. While proposing entirely new algorithms would appear to be a possible solution [22], researchers have mostly preferred other methods. Many solutions are put forth and review studies have been published on specific categories of solutions; examples include surveys on Big Data analytics platforms [17], [18] and review study of data mining with Big Data [20].

Table 1. Learning Paradigms and Challenges

APPROACHES		CHALLENGES																
		VOLUME								VERIETY			VELOCITY			VERACITY		
		Processing Performance	Curse of Modularity	Class Imbalance	Curse of Dimensionality	Feature Engineering	Non-linearity	Bonferomi's Principle	Variance and Bias	Data Locality	Data Heterogeneity	Dirty and Noisy Data	Data Availability	Real-time Processing	Concept drift	Data Provenance	Data Uncertainty	Dirty and Noisy Data
LEARNING PARADIGMS	Deep Learning					H	H				H	P					P	P
	Online Learning	H	H	P						H		P	H	H	P			P
	Local Learning	H	H	H					H	H								
	Transfer Learning			H							H	P					P	P
	Lifelong learning	H		H							H	P	H	H	P		P	P
	Ensemble Learning	H	H												H			

The wide image of the approaches towards the challenges is presented in Table 1; which contains approaches and challenges that how best they address. "H" indicates high degree of solution while 'P' represents partial remedy. As shown in the table, there are two main classes of solutions: the first one relies on data, processing, and algorithm manipulations to handle Big Data and the latter one show the creation and adaptation of different machine learning paradigms and the modification of existing paradigms. Besides, several machine learning solutions are in place as service offerings which include: Microsoft Azure Machine Learning, now part of Cortana Intelligence Suite [23]; Google Cloud Machine Learning Platform [24]; Amazon Machine Learning [25]; and IBM Watson Analytics [26]. These ML tools and services are backed up by powerful cloud offerings. But currently, they support a less number of algorithms compared to R

[27] and MATLAB [28]. Also, computation happens on cloud resources, which requires data transfer to remote nodes and this lead to high network traffic and shall become infeasible due to time or bandwidth requirements. As these ML services are proprietary, information about their underlying technologies is very restricted; therefore. The upcoming sections propose techniques and methodologies being developed and used to handle the challenges associated with data learning with Big Data.

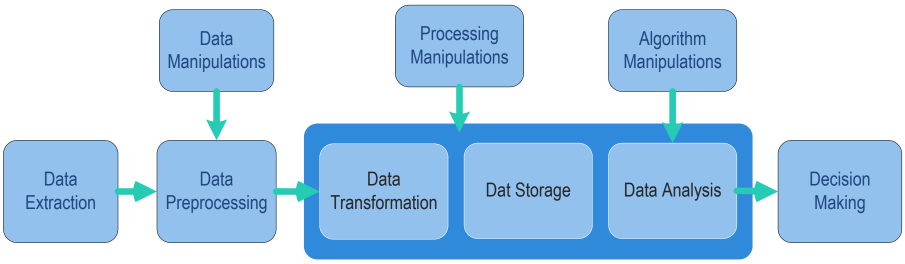

Figure 2: Data analysis process flow

Big Data analytics using machine intelligence relies on an established process which is referred to as the data analytics pipeline. Data manipulations respond to the Big Data challenges in machine learning. Fig.2 shows a representation of the data analysis pipeline. The three manipulations, along with their corresponding sub-categories are depicted in Fig. 3.

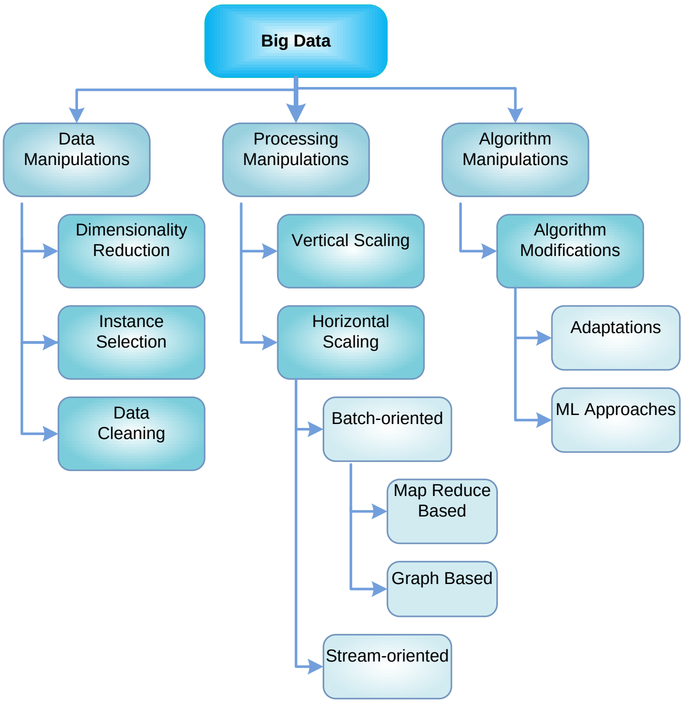

Figure 3. Big Data Manipulations in ML

5. LEARNING PARADIGMS:

Number of learning paradigms emerged under machine learning domain; yet, few are irrelevant to certain areas of research like signal processing [29]. The following sub-sections briefs the machine learning paradigms relevant in the Big Data context, together with the way they address the identified big data challenges.

Deep Learning: Deep learning is branch of a broader family of machine learning methods based on artificial neural networks with representation learning. Representation learning is generally called as feature learning [30]. Deep learning algorithm gets its name from the fact that it uses data representations rather than explicit data features to get insights. The learning process transforms data into abstract representations that facilitate the features to be learnt. These representations are subsequently used to accomplish the machine learning tasks. As the features learned directly from the data, there is no need for feature engineering. The ability to avoid feature engineering is regarded as a great advantage in the context of Big Data due to the challenges associated with this process.

Online Learning: Online Learning paradigm is a state-of-art approach to deal with online live Big Data as it responds in a predictable manner to large-scale processing by nature. It is machine learning paradigm that has been explored to bridge efficiency gaps created by Big Data. It can be seen as an alternative to batch learning or conventional machine learning. As its name suggests, batch learning processes data in batches and requires the entire dataset to be available when the model is created [31]. The limitation of batch learning is once the model is generated no longer be modified. It makes difficult to deal with the dimensions of Big Data for the following said reasons. i) Having to process a very large amount of data at one time is not computationally efficient or always feasible. ii) It is necessary to have the entire dataset available at the beginning of the processing limits the use of data from various sources. iii) Need to have access to the entire dataset at the time of processing does not enable real-time analysis. iv) Since the model cannot be altered, it is highly prone to performance impediments caused by poor data veracity. On the other hand, online learning employ data streams for training, and ML models can learn one instance at a time [15]. This can lessen the computational load and enhance performance as the data need not be entirely held in memory. As a result of this online learning, it enables processing of huge volumes of data, solves the curse of modularity, smoothes the process of real-time processing, and provides the ability to learn from data. Furthermore, as it does not require all data to be present at once or located at the same place, this paradigm remedies data availability and locality.

Local Learning: Local learning is a ML technique that offers an alternative to typical global learning and is first proposed by Bottou and Vapnik in 1992 [32]. Typically, Machine Learning algorithms make use of global learning through strategies such as generative learning [33]. The idea of ML is that based on the data’s underlying distribution, a data learning model can be used to re-generate the input data. It basically attempts to summarize the entire dataset, but local learning is concerned only with subsets of interest. Thus, local learning can be viewed as a semi-parametric approximation of a global model. The stronger but less restrictive assumptions of this hybrid parametric model notably yield low variance and bias [4]. The abstract view of the local learning process is given in Fig 4. The concept of local learning is to separate the input space into clusters and then build a separate model for each cluster. By this way the overall cost and complexity can be reduced. In fact, it is more competent to solve k problems of size m/k than for a single problem of size m. As a result, this method can enable processing of datasets that are considered very large for global paradigms.

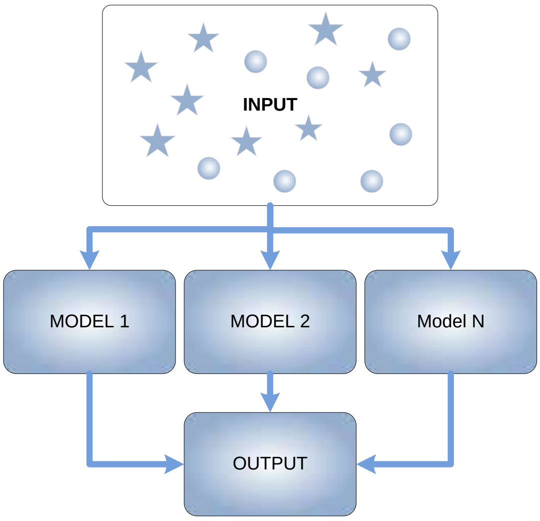

Figure 4. Local Learning

Transfer Learning: It is an approach for enhancing learning for a particular application domain, termed as the target domain. It trains the model with other datasets from multiple domains, referred to as source domains, with similar attributes or features of the data. Transfer learning is preferred when the data size within the target domain is insufficient or the learning task is different [34]. The Fig. 5 illustrates an abstract view of transfer learning.

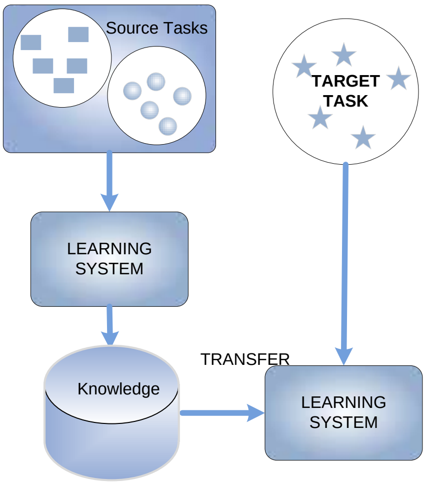

Figure 5. Transfer Learning

Lifelong Learning: Lifelong learning imitates the human learning; learning is continuous process in human; the acquired knowledge is retained to solve different problems. It is aimed at maximize overall learning, in such a way to solve a new problem by training either on one single domain or on varied domains collectively [35]. The outcome of the learning from the training process are collected and combined together to make a model called knowledge model. On training heterogeneous domains, transfer learning might be used in the combining step to create such a topical model. The knowledge rely on this topical model is used to perform a new task or solve new problems regardless of the source of knowledge.

Ensemble Learning: Ensemble learning is achieved by joining multiple learning machines to get enhanced learning outcomes than those obtained from any constituent learner [36]. Fig. 6 depicts a classical view of the ensemble learning process. In general, the overall outcome is obtained by a voting process among the weighted outcomes of individual learners [36]. The individual learners can be similar or from completely different categories, including those belonging to supervised and unsupervised paradigms. The process of weighting mechanism assigns a value to each learning output point and combines them to show up the aggregate outcome. The voting process could be simply aggregating the values of the learning points or by means of the statistical techniques to get a combined value of the learning outputs that may lead to better learning performance [37]. For instance, Waske et al. [38] have adopted SVM for individual learners and also used an SVM in the voting process in their work.

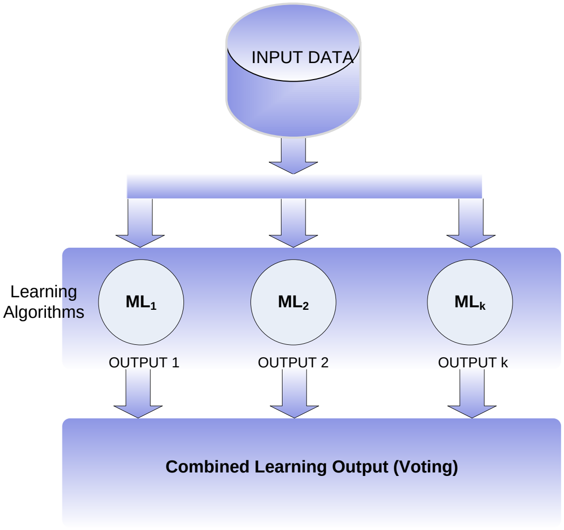

Figure 6. Ensemble Learning

6. CONCLUSION

This review has brought various research findings on big data and the challenges faced by the researchers in dealing with these data for various application domains. It has presented major challenges in machine learning with Big Data, reviewed state-of-art machine learning approaches, and presented how each approach is capable of addressing the identified challenges. This paper has given an well systematic review of the challenges associated with machine learning in the context of Big Data and categorized them according to the V dimensions of Big Data. Besides, this review has reported an overview of ML paradigms and discussed how these techniques beat the various challenges identified. In addition, the explicit relation between ML approaches and Big Data challenges are well established in this paper and fulfils the prime objective of this work. This is to provide deeper understanding of machine learning with Big Data for the future researchers. This work also provide better foundation for making easier and better informed choices with regard to machine learning with Big Data so as to achieve the second objective. It has been achieved by developing a comprehensive matrix that lays out the relationships between the various challenges and machine learning approaches, thus guiding for the best choices given a set of conditions. This paper also opens research opportunities for the researchers like adaptation of new machine learning paradigms for the unsolved problems, combination of existing solutions to achieve further performance improvements. Thereby, the comprehensive review study accomplishes all its musters and provides the research community with potential guidance for their future work to achieve great improvements in the fields of big data and machine learning.

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

[6] John Martin R, Swapna S.L, Sujatha S "Adopting Machine Learning Models for Data Analytics-A Technical Note." International Journal of Computer Sciences and Engineering 6.10 (2018): 360-365.

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

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

[9] 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.

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

[11] 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.

[12] 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.

[13] 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.

[14] 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.

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

[16] John Martin R., Sujatha S. and Swapna S L.. Multiresolution Analysis in EEG Signal Feature Engineering for Epileptic Seizure Detection. International Journal of Computer Applications 180(17):14-20, February 2018.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, Data Mining: Practical Machine Learning Tools and Techniques. San Mateo, CA: Morgan Kaufmann, 2016.

[23] R. Barga, V. Fontama, and W. H. Tok, ‘‘Cortana analytics,’’ in Predictive Analytics With Microsoft Azure Machine Learning. Berkeley, CA, USA: Apress, 2015, pp. 279–283.

[24] Google. (2016). Google Cloud Machine Learning. [Online]. Available: https://cloud.google.com/products/machine-learning/

[25] Amazon Web Services. (2016). Amazon Machine Learning. [Online]. Available: https://aws.amazon.com/machine-learning/

[26] IBM. (2014). IBM Watson Ecosystem Program. [Online]. Available: http://www-03.ibm.com/innovation/us/watson/

[27] R Core Team, R: A Language and Environment for Statistical Computing, Vienna, Austria, 2015, vol.1.

[28] MATLAB, The MathWorks Inc., Natick, MA, USA, 2016.

[29] John Martin R, Swapna S.L, “A Machine Learning Framework for Epileptic Seizure Detection by Analyzing EEG Signals,” International Journal of Computing and Digital Systems, 2021.

[30] Y. Bengio, A. Courville, and P. Vincent, ‘‘Representation learning: A review and new perspectives,’’ IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 8, pp. 1798–1828, Aug. 2013.

[31] J. Leskovec, A. Rajaraman, and J. D. Ullman, Mining of Massive Datasets, vol. 13. Cambridge, U.K.: Cambridge Univ. Press, 2014.

[32] L. Bottou and V. Vapnik, ‘‘Local Learning Algorithms,’’ Neural Comput., vol. 4, no. 6, pp. 888–900, 1992.

[33] K. Huang, H. Yang, I. King, and M. R. Lyu, ‘‘Local learning vs. global learning: An introduction to maxi-min margin machine,’’ in Support Vector Machines: Theory and Applications. Berlin, Germany: Springer, 2005, pp. 113–131.

[34] L. Torrey and J. Shavlik, Handbook of Research on Machine Learning Applications and Trends. Hershey, PA: IGI Global, 2010.

[35] D. L. Silver, Q. Yang, and L. Li, ‘‘Lifelong machine learning systems: Beyond learning algorithms,’’ in Proc. AAAI Spring Symp., 2013, pp. 49–55.

[36] T. Dietterich, ‘‘Ensemble methods in machine learning,’’ in Multiple Classifier Systems, vol. 1857. London, U.K.: Springer-Verlag, 2000, pp. 1–15.

[37] M. Sewell, ‘‘Ensemble learning,’’ Dept. Comput. Sci., UCL, London, U.K., Tech. Rep. RN/11/02, 2011, p. 12.

[38] B. Waske and J. A. Benediktsson, ‘‘Fusion of support vector machines for classification of multisensor data,’’ IEEE Trans. Geosci. Remote Sens., vol. 45, no. 12, pp. 3858–3866, Dec. 2007.