Journal of Soft Computing Paradigm (JSCP) (2019)
Vol.01/ No. 02
Pages: 103- 112
http://irojournals.com/jscp/
DOI: https://doi.org/10.36548/jscp.2019.2.005

REVIEW OF MACHINE LEARNING TECHNIQUES FOR VOLUMINOUS INFORMATION MANAGEMENT

Dr. A. Pasumpon Pandian,
Professor,
Computer Science Engineering,
KGiSL Institute of Technology,
Coimbatore, India.
Email id: pasumponpandian32@gmail.com

Abstract: The recent technological growth at a rapid pace has paved way for the big data that denotes to the exponential growth of the information's. The big data analytics are the trending concepts that have emerged as the promising technology that offers more enhanced perceptions from the huge set of the data that have been produced from the diverse areas. The review in the paper proceeds with the methods of the big-data-analytics and the machine-learning in handling, the huge set of data flow. The overview of the utilization of the machine-learning algorithms in the analytics of high voluminous data would provide with the deeper and the richer analysis of the huge set of information gathered to extract the valuable and turn it into actionable information's. The paper is to review the part of machine-learning algorithms in the analytics of high voluminous data

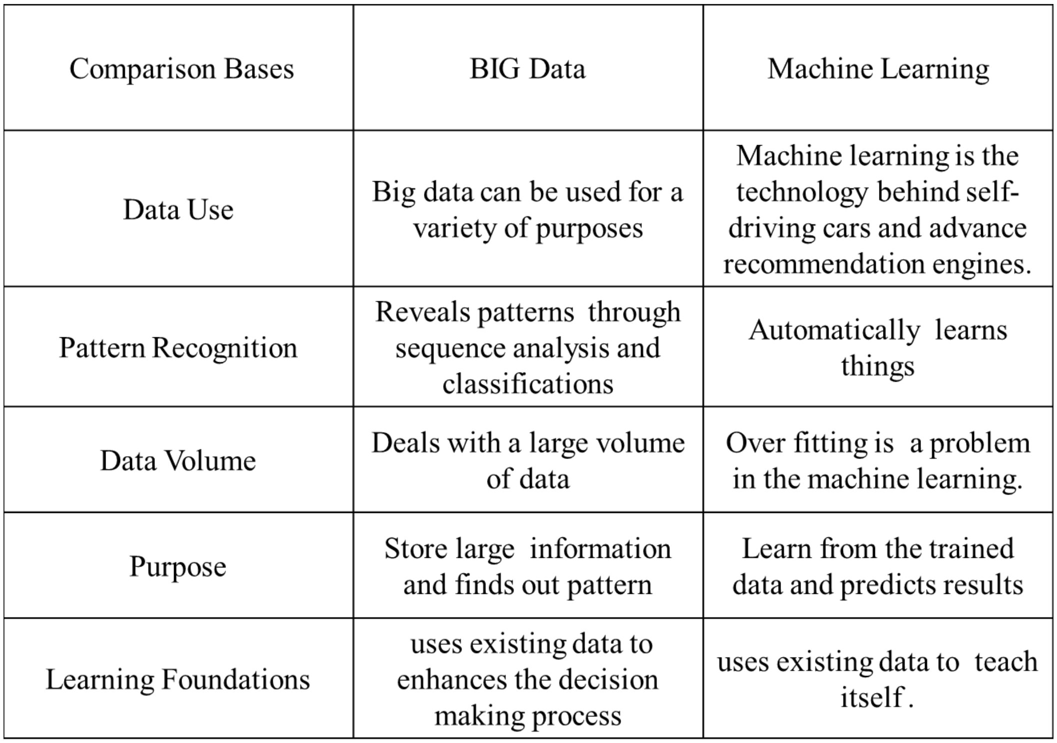
Keywords: Big Data Analytics, Big Data Management, Supervised, Reinforcement Learning, Machine- learning and Deep Learning

1. INTRODUCTION

Big data (high voluminous data) [1] is a term coined to describe the heavy capacious data that exponentially grows at a rapid pace and the analytics on the big-data (B-D) [2] helps in extracting the valuable information from the data flow by framing the probable relations between the diverse set data accumulated. This massive amount of data gathered from variety of sources either online or offline at a very high rate in various formats is challenging to be handled by the human analytics to extract the values from it and convert them into actionable solutions. So the big data analytics are engaged in extracting the values in information's gathered.

The voluminous data [4] is defined as the method to gather and examine the huge set of data i.e. the big data. The analytics of the voluminous data remains useful for discovering hidden patterns and the other useful evidence's / statistics like the new trends in the market, the customer choices etc. so the big data analytics [5] help in the enhancing the decision making capabilities to improve the future steps in business, in health care, in developing smart applications, in security measures, in governance and social networks.

The machine learning (M-L) [12] is the specific type of artificial intelligence. That helps in predicting the actions that are to be taken in the future without any human intervention utilizing only the computers and the machines. The machine learning could help in improving the accuracy of the outputs predicted and remains as the predominant technology in the recent days. The table.1 below provides the comparison between the B-D and the M-L



Comparison Bases	BIG Data	Machine Learning
Data Use	Big data can be used for a variety of purposes	Machine learning is the technology behind self-driving cars and advance recommendation engines.
Pattern Recognition	Reveals patterns through sequence analysis and classifications	Automatically learns things
Data Volume	Deals with a large volume of data	Over fitting is a problem in the machine learning.
Purpose	Store large information and finds out pattern	Learn from the trained data and predicts results
Learning Foundations	uses existing data to enhances the decision making process	uses existing data to teach itself .

Table .1 M-L and B-D Comparison

The paper is to review the involvement of machine learning in the big data analytics [15] to bring out more useful information' s and accuracy in prediction of the of the values. The fig .1 below shows the involvement of the Machine Learning in the Big Data

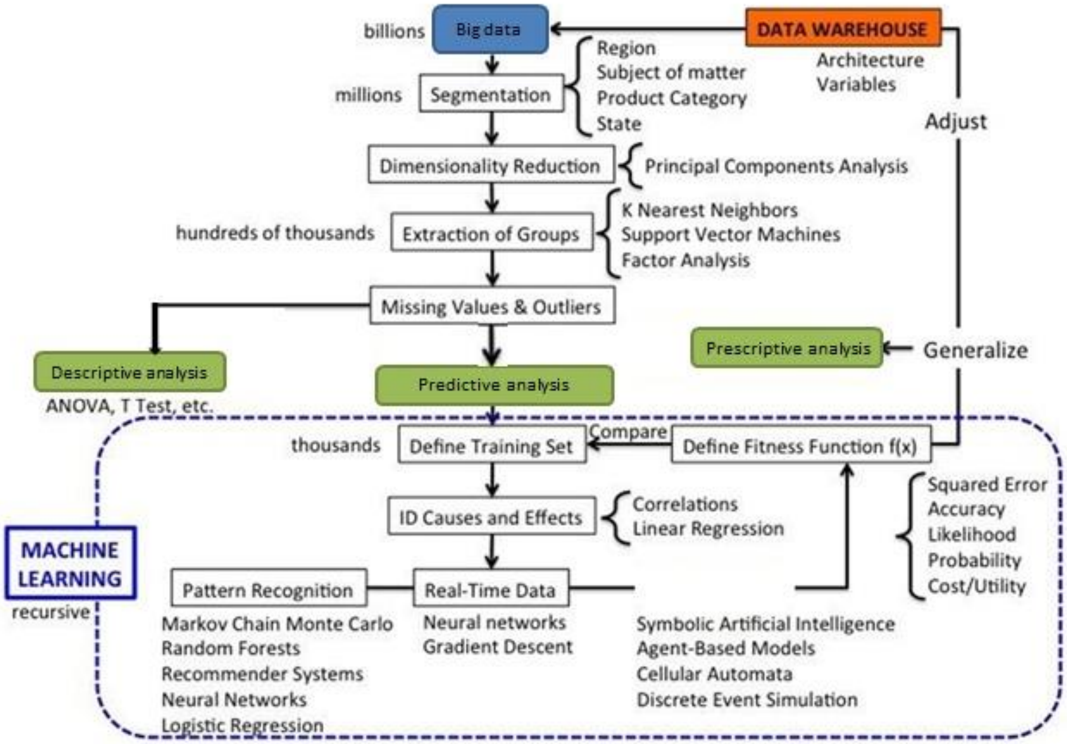
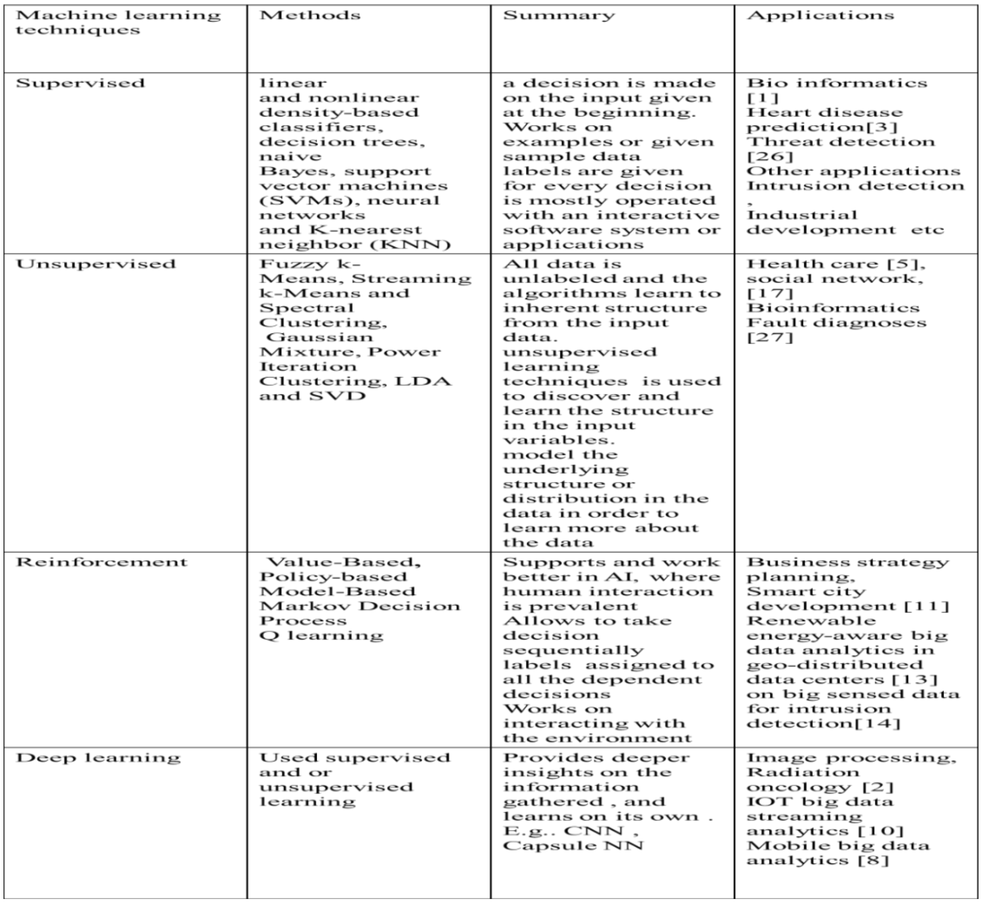


Fig.1. Involvement of M-L in Big-DA [16]

The paper is organized with the 2. Detailing the M-L involvement in B-D, 3.providing the overview M-L tools for B-D. 4. Presenting the challenges and issue faced by M-L on B-D and 5. The conclusion

2. M-L INVOLVEMENT IN B-D

The machine learning techniques utilizes the learning that are reinforcement learning [11, 14] ,unsupervised [24] , deep learning [8, 9, and 10], and supervised [17], efficiently analysis the huge volume of data and measure the correlation between the data that are gathered to segregate the values from them and present the deeper insights that are present in the data. The tabulation below in table.2 shows the deeper insights provided by the different techniques of the machine-learning with the B-D analytics in diverse applications.

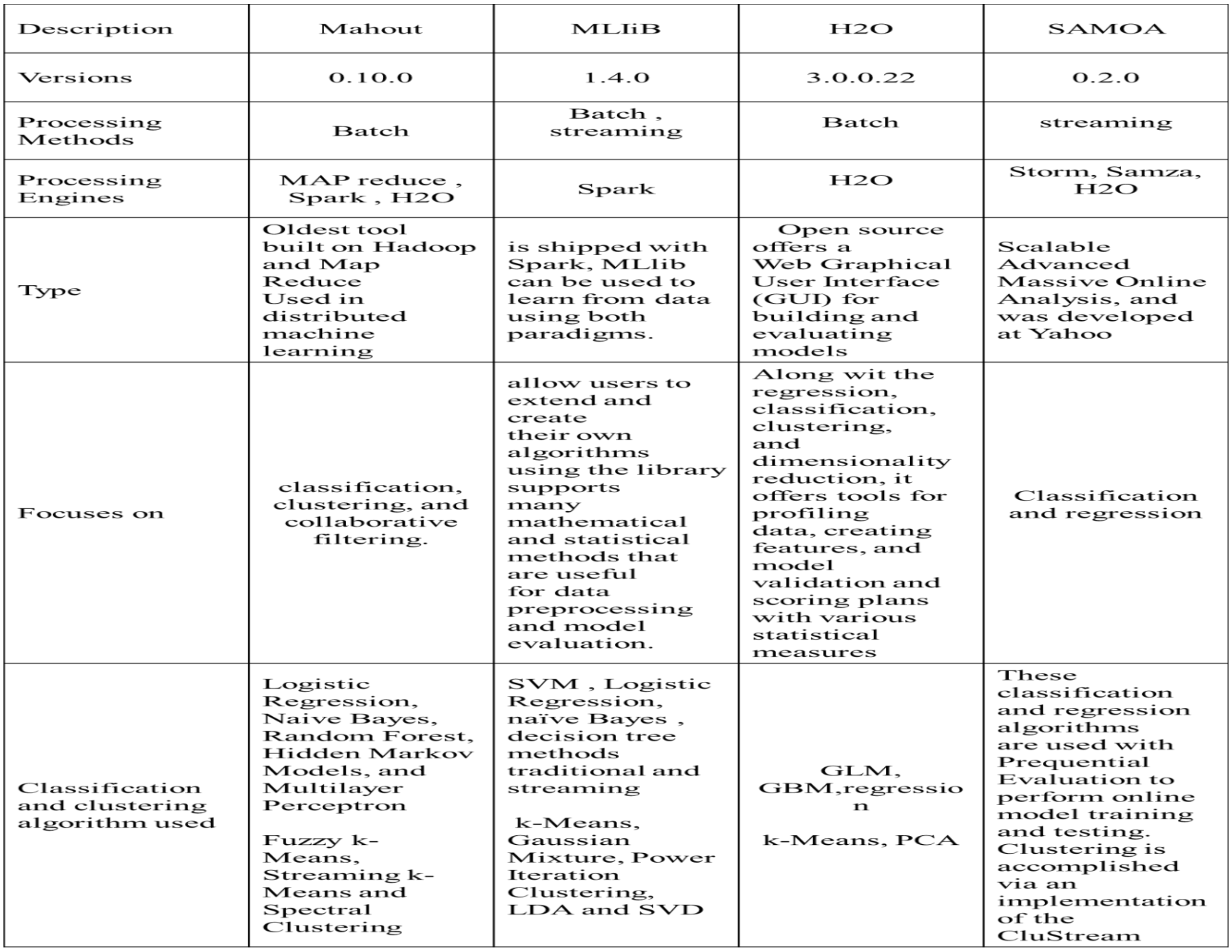


Machine learning techniques	Methods	Summary	Applications
Supervised	linear and nonlinear density-based classifiers, decision trees, naive Bayes, support vector machines (SVMs), neural networks and K-nearest neighbor (KNN)	a decision is made on the input given at the beginning. Works on examples or given sample data labels are given for every decision is mostly operated with an interactive software system or applications	Bio informatics [1] Heart disease prediction[3] Threat detection [26] Other applications Intrusion detection , Industrial development etc
Unsupervised	Fuzzy k-Means, Streaming k-Means and Spectral Clustering, Gaussian Mixture, Power Iteration Clustering, LDA and SVD	All data is unlabeled and the algorithms learn to inherent structure from the input data. unsupervised learning techniques is used to discover and learn the structure in the input variables. model the underlying structure or distribution in the data in order to learn more about the data	Health care [5], social network, [17] Bioinformatics Fault diagnoses [27]
Reinforcement	Value-Based, Policy-based Model-Based Markov Decision Process Q learning	Supports and work better in AI, where human interaction is prevalent Allows to take decision sequentially labels assigned to all the dependent decisions Works on interacting with the environment	Business strategy planning, Smart city development [11] Renewable energy-aware big data analytics in geo-distributed data centers [13] on big sensed data for intrusion detection[14]
Deep learning	Used supervised and or unsupervised learning	Provides deeper insights on the information gathered , and learns on its own . E.g.. CNN , Capsule NN	Image processing, Radiation oncology [2] IOT big data streaming analytics [10] Mobile big data analytics [8]

Table.2 Summary of M-L in B-D

3. THE M-L TOOLS FOR B-D [20] [32]

The M- in B-D is achieved using various tool kits in machine learning. Some of the prominent tool kits that are used in developing the own machine algorithms for the big data analytics are presented in the table.3 below. The selection criteria for the perfect toolkits for the big data depends on the scalability, speed, coverage and the extensibility.



Description	Mahout	MLIiB	H2O	SAMOA
Versions	0.10.0	1.4.0	3.0.0.22	0.2.0
Processing Methods	Batch	Batch , streaming	Batch	streaming
Processing Engines	MAP reduce , Spark , H2O	Spark	H2O	Storm, Samza, H2O
Type	Oldest tool built on Hadoop and Map Reduce Used in distributed machine learning	is shipped with Spark, MLlib can be used to learn from data using both paradigms.	Open source offers a Web Graphical User Interface (GUI) for building and evaluating models	Scalable Advanced Massive Online Analysis, and was developed at Yahoo
Focuses on	classification, clustering, and collaborative filtering.	allow users to extend and create their own algorithms using the library supports many mathematical and statistical methods that are useful for data preprocessing and model evaluation.	Along wit the regression, classification, clustering, and dimensionality reduction, it offers tools for profiling data, creating features, and model validation and scoring plans with various statistical measures	Classification and regression
Classification and clustering algorithm used	Logistic Regression, Naive Bayes, Random Forest, Hidden Markov Models, and Multilayer Perceptron Fuzzy k-Means, Streaming k-Means and Spectral Clustering	SVM , Logistic Regression, naïve Bayes , decision tree methods traditional and streaming k-Means, Gaussian Mixture, Power Iteration Clustering, LDA and SVD	GLM, GBM,regression k-Means, PCA	These classification and regression algorithms are used with Prequential Evaluation to perform online model training and testing. Clustering is accomplished via an implementation of the CluStream

Table.3 Machine Tool Kits for Big Data

4. CHANLLENGES FOR MACHINE LEARNING IN BIG DATA [12], [34], [35]

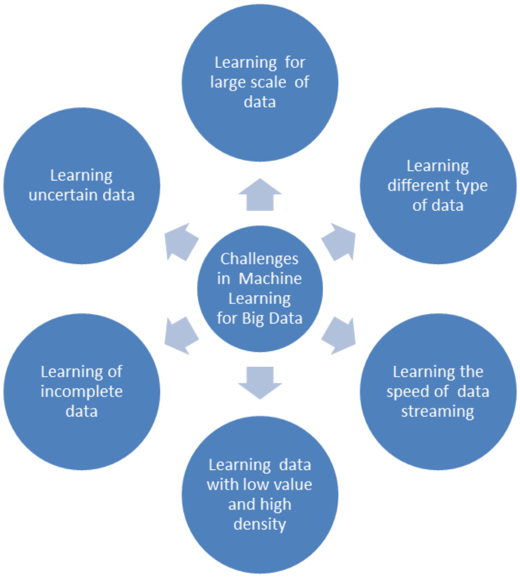

Challenges in Machine Learning for Big Data
Learning for large scale of data
Learning different type of data
Learning the speed of data streaming
Learning data with low value and high density
Learning of incomplete data
Learning uncertain data

Fig.2 Challenges in ML for BD

The fig.2 shows the issues faced by the M-L for the B-DA. Some of the issue that are related to M-L for B-DA are integration of data analytics that is fast and huge, the security and the privacy preservation, shortage in the big data, on device intelligence, context awareness [12], data uncertainty, data heterogeneity real time processing/streaming etc. these challenges prevails as the interruptions in machine learning to provide the deeper insights. The promising method to improve the challenges faced by the M-L in B-D is addressed as a future work of the paper.

5. CONCLUSION

The review in the paper provides the outline of the exploitation of the machine learning algorithms in the B-DA to gain deep insight for the information's gathered. The paper proceeds with the explanation describing the involvement of the M-L in B-DA application, the tool kits of machine learning available for the B-D applications and further provides the details of the trials and the issue associated in M-L for the B-D. The paper in the future is to continue with the promising methods to overcome the challenges faced in the M-L for B-D.

References

[1] Kashyap, Hirak, Hasin Afzal Ahmed, Nazrul Hoque, Swarup Roy, and Dhruba Kumar Bhattacharyya. "Big data analytics in bioinformatics: A machine learning perspective." *arXiv preprint arXiv:1506.05101* (2015).

[2] Bibault, Jean-Emmanuel, Philippe Giraud, and Anita Burgun. "Big data and machine learning in radiation oncology: state of the art and future prospects." *Cancer letters* 382, no. 1 (2016): 110-117.

[3] Kaur, Beant, and Williamjeet Singh. "Review on heart disease prediction system using data mining techniques." *International journal on recent and innovation trends in computing and communication* 2, no. 10 (2014): 3003-3008.

[4] Gandomi, Amir, and Murtaza Haider. "Beyond the hype: Big data concepts, methods, and analytics." *International journal of information management* 35, no. 2 (2015): 137-144.

[5] Kaur, Prableen, Manik Sharma, and Mamta Mittal. "Big data and machine learning based secure healthcare framework." *Procedia computer science* 132 (2018): 1049-1059.

[6] Koh, Hian Chye, and Gerald Tan. "Data mining applications in healthcare." *Journal of healthcare information management* 19, no. 2 (2011): 65.

[7] Bhardwaj, Ashu, and Williamjeet Singh. "Systematic review of big data analytics in governance." In *2017 International Conference on Intelligent Sustainable Systems (ICISS)*, pp. 501-506. IEEE, 2017.

[8] Alsheikh, Mohammad Abu, Dusit Niyato, Shaowei Lin, Hwee-Pink Tan, and Zhu Han. "Mobile big data analytics using deep learning and apache spark." *IEEE network* 30, no. 3 (2016): 22-29.

[9] Chen, Xue-Wen, and Xiaotong Lin. "Big data deep learning: challenges and perspectives." *IEEE access* 2 (2014): 514-525.

[10] Mohammadi, Mehdi, Ala Al-Fuqaha, Sameh Sorour, and Mohsen Guizani. "Deep learning for IoT big data and streaming analytics: A survey." *IEEE Communications Surveys & Tutorials* 20, no. 4 (2018): 2923-2960.

[11] He, Ying, F. Richard Yu, Nan Zhao, Victor CM Leung, and Hongxi Yin. "Software-defined networks with mobile edge computing and caching for smart cities: A big data deep reinforcement learning approach." *IEEE Communications Magazine* 55, no. 12 (2017): 31-37.

[12] Mohammadi, Mehdi, and Ala Al-Fuqaha. "Enabling cognitive smart cities using big data and machine learning: Approaches and challenges." *IEEE Communications Magazine* 56, no. 2 (2018): 94-101.

[13] Xu, Chenhan, Kun Wang, Peng Li, Rui Xia, Song Guo, and Minyi Guo. "Renewable energy-aware big data analytics in geo-distributed data centers with reinforcement learning." *IEEE Transactions on Network Science and Engineering* (2018).

[14] Otoum, Safa, Burak Kantarci, and Hussein Mouftah. "Empowering reinforcement learning on big sensed data for intrusion detection." In *ICC 2019-2019 IEEE International Conference on Communications (ICC)*, pp. 1-7. IEEE, 2019.

[15] Ma, Chuang, Hao Helen Zhang, and Xiangfeng Wang. "Machine learning for big data analytics in plants." *Trends in plant science* 19, no. 12 (2014): 798-808.

[16] nationalinterest.in/big-data-analytics-usingmachinelearningalgorithmsc33ef8488638#:~:targetText=Machine%20Learning%20is%20used%20to,past%20experience%20i.e.%20data%20models.

[17] Hussain, Amir, and Erik Cambria. "Semi-supervised learning for big social data analysis." *Neurocomputing* 275 (2018): 1662-1673.

[18] Wang, Lidong, and Cheryl Ann Alexander. "Machine learning in big data." *International Journal of Mathematical, Engineering and Management Sciences* 1, no. 2 (2016): 52-61.

[19] Condie, Tyson, Paul Mineiro, Neoklis Polyzotis, and Markus Weimer. "Machine learning on big data." In *2013 IEEE 29th International Conference on Data Engineering (ICDE)*, pp. 1242-1244. IEEE, 2013.

[20] Harrington, Peter. *Machine learning in action*. Manning Publications Co., 2012.

[21] Landset, Sara, Taghi M. Khoshgoftaar, Aaron N. Richter, and Tawfiq Hasanin. "A survey of open source tools for machine learning with big data in the Hadoop ecosystem." *Journal of Big Data* 2, no. 1 (2015): 24.

[22] Zhou, Lina, Shimei Pan, Jianwu Wang, and Athanasios V. Vasilakos. "Machine learning on big data: Opportunities and challenges." *Neurocomputing* 237 (2017): 350-361.

[23] Madden, Sam. "From databases to big data." *IEEE Internet Computing* 16, no. 3 (2012): 4-6.

[24] Zhang, Qingchen, Laurence T. Yang, and Zhikui Chen. "Deep computation model for unsupervised feature learning on big data." *IEEE Transactions on Services Computing* 9, no. 1 (2015): 161-171.

[25] Kanevsky, Jonathan, Jason Corban, Richard Gaster, Ari Kanevsky, Samuel Lin, and Mirko Gilardino. "Big data and machine learning in plastic surgery: a new frontier in surgical innovation." *Plastic and reconstructive surgery* 137, no. 5 (2016): 890e-897e.

[26] Mayhew, Michael, Michael Atighetchi, Aaron Adler, and Rachel Greenstadt. "Use of machine learning in big data analytics for insider threat detection." In *MILCOM 2015-2015 IEEE Military Communications Conference*, pp. 915-922. IEEE, 2015.

[27] Lei, Yaguo, Feng Jia, Jing Lin, Saibo Xing, and Steven X. Ding. "An intelligent fault diagnosis method using unsupervised feature learning towards mechanical big data." *IEEE Transactions on Industrial Electronics* 63, no. 5 (2016): 3137-3147.

[28] Assefi, Mehdi, Ehsun Behravesh, Guangchi Liu, and Ahmad P. Tafti. "Big data machine learning using apache spark MLlib." In *2017 IEEE International Conference on Big Data (Big Data)*, pp. 3492-3498. IEEE, 2017.

[29] Hajj, Nadine, Yara Rizk, and Mariette Awad. "A mapreduce cortical algorithms implementation for unsupervised learning of big data." *Procedia Computer Science* 53 (2015): 327-334.

[30] Veeramachaneni, Kalyan, Ignacio Arnaldo, Vamsi Korrapati, Constantinos Bassias, and Ke Li. "AI^ 2: training a big data machine to defend." In *2016 IEEE 2nd International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing (HPSC), and IEEE International Conference on Intelligent Data and Security (IDS)*, pp. 49-54. IEEE, 2016.

[31] Park, Seongwook, Kyeongryeol Bong, Dongjoo Shin, Jinmook Lee, Sungpill Choi, and Hoi-Jun Yoo. "4.6 A1. 93TOPS/W scalable deep learning/inference processor with tetra-parallel MIMD architecture for big-data applications." In *2015 IEEE International Solid-State Circuits Conference-(ISSCC) Digest of Technical Papers*, pp. 1-3. IEEE, 2015.

[32] Zhao, Ying, Doug MacKinnon, and Shelley P. Gallup. "Big data and deep learning for understanding DoD data." *CrossTalk* 28, no. 4 (2015): 4-10.

[33] Richter, Aaron N., Taghi M. Khoshgoftaar, Sara Landset, and Tawfiq Hasanin. "A multi-dimensional comparison of toolkits for machine learning with big data." In *2015 IEEE International Conference on Information Reuse and Integration*, pp. 1-8. IEEE, 2015.

[34] Suthaharan, Shan. "Big data classification: Problems and challenges in network intrusion prediction with machine learning." *ACM SIGMETRICS Performance Evaluation Review* 41, no. 4 (2014): 70-73.

[35] L’heureux, Alexandra, Katarina Grolinger, Hany F. Elyamany, and Miriam AM Capretz. "Machine learning with big data: Challenges and approaches." *IEEE Access* 5 (2017): 7776-7797.