*Engineering and Scientific International Journal (ESIJ)*
*Volume 8, Issue 3, July – September 2021*
ISSN 2394-7187(Online)
ISSN 2394 - 7179 (Print)

# Big Data on Machine Learning – A Review

K. Balasree[*1], Dr. K. Dharmarajan[2]

[1]*Research Scholar, Department of Computer Science, VISTAS Pallavaram, Chennai, India*
*Email id: bala1996shree@gmail.com*
[2]*Associate Professor, Associate Professor, Department of Computer Science, VISTAS-Pallavaram, Chennai, India*

**Abstract**— In rapid development of Big Data technology over the recent years, this paper discussing about the Machine Learning (ML) playing role that is based on methods and algorithms to Big Data Processing and Big Data Analytics. In evolutionary fields and computing fields of developments that both are complementing each other. Big Data: The rapid growth of such data solutions needed to be studied and provided to handle then to gain the knowledge from datasets and extracting values due to the data sets are very high in velocity and variety. The Big data analytics are involving and indicating the appropriate data storage and computational outline that enhanced by using Scalable Machine Learning Algorithms and Big Data Analytics then the analytics to reveal the massive amounts of hidden data's and secret correlations. This type of Analytic information useful for organizations and companies to gain deeper knowledge, development and getting advantages over the competition. When using this Analytics we can predict the accurate implementation over the data. This paper presented about the detailed review of state-of-the-art developments and overview of advantages and challenges in Machine Learning Algorithms over big data analytics.

***Keywords*** *— Big Data; Analytics; Machine Learning; SVM; DecisionTree; Naïve Baye's; Random Forest.*

## 1. Introduction

In the massive world, the internet has been exponential growth of data includes smart phones or smart sensors leads to big data. The term big data can be referred as the data has massive, high speed, different categories and with lots unwanted noises that are very difficult to store, process, analyse, interpret, consume and make improved decisions in the field of healthcare, business, finance or in any other industries.

This paper provides the overview review about Machine Learning and explains the working procedure using different Algorithms efficiently and we predicted some accuracy using algorithms that are included by Machine Learning Literature review, includes theoretical, empirical and experimental studies relating to the various needs and recommendation have analysed [10]. In section I represents about the overview of Big Data and Machine Learning, Section II represents about the Techniques in Machine Learning. In Section III, the role of Machine Learning on Big Data and the Different types of Algorithms available in Machine Learning and Section IV describes the Popular Language in Machine Learning (Python), then section V explains the Framework of Big Data in Machine Learning.

### 1.1 Big Data

In Large-volume, autonomous and heterogeneous sources with the distributed and decentralized control that is to evolve dealing among with the data and to explore the complex situations. The important characteristics in Big Data is a large volume of data that represented by heterogeneous and varied dimensionalities. The Machine Learning (ML) is an artificial approach that is used for learning knowledge in massive data for making it better to take intelligent decisions. This techniques that generated a huge volume of social range applications such as understanding, Computer vision, Neuroscience, Health, Internet of Things and Speech recognition. The arrival of Big Data urgedas broad benefits in Machine Learning [3].

### 1.2 Machine Learning Techniques

#### *1.2.1 Supervised Learning*

In Supervised Learning, The machine learning is a task of learning function that maps the input and output on both example of output-input pairs. In this learning, each pair of example contains an input object i.e., typically a vector and a desired output (supervisory signal). In supervised learning algorithm, the analysis trained about the data and then produce the indirect function that used for mapping. An optimal scenario that resolves the algorithm to control it suitably and the labels class for unseeded instances.
Supervised learning practices labelled training data that Maps the function to turns both the variables such as the input variables (X) into the output variable (Y). In additional words, it solves for f in the subsequent equation:

$$Y = f \quad (1)$$



DOI: 10.30726/esij/v8.i3.2021.83018

ISSN 2394-7187(Online)
ISSN 2394 - 7179 (Print)

Supervised learning algorithms:

- k-Nearest Neighbor (Classification)
- Naive Bayes (Classification)
- Decision Tress/Random Forest (Classification and Regression)
- Support Vector Machine (Classification)
- Linear Regression (Regression)
- Logistic Regression (Classification)
- Deep learning algorithms:
- Neural Network
- Convolution Network
- Recurrent Network

*1.2.2 Unsupervised Learning*

The unsupervised learning Algorithm decided to perform the more complexion process that tasks when related to supervised learning. So, the learning, it is able to be random to predict the associated with the other usual learning methods. Unsupervised learning aim is to find the fundamental structure of data set groups the data permitting the correspondences and indicates the dataset in compressed format.

*1.2.3 Application of Unsupervised Machine Learning*

Clustering is divided into the groups that based on their similarity of data. Anomaly detecting controls the unusual data facts in dataset that is valuable for the assumption regarding fraudulent transactions. Association mining that classified the set of substances regularly often together in dataset. Latent variable models are widely used in data pre-processing.

Unsupervised Learning Algorithms:

- Neural Networks.
- Anomaly detection.
- KNN (k-nearest neighbors).
- K-means clustering.
- Hierarchal clustering.

*1.2.4 Semi-supervised Learning*

In semi-supervised Learning, the approach of Machine Learning that combines a small amount of data labelled in a large amount of data unlabelled throughout the training. The Semi-Supervised learning falling the supervised learning-With only labelled data and unsupervised learning-With no labelled data training. Once used the unlabelled data, the combination with the smaller amount of data labelled and it can also produce the significant development in accuracy of learning.

Semi Supervised Learning Algorithms:

- Graph-Based Methods
- Manifold Assumption
- Continuity Assumption
- Generative Models
- Heuristic Approaches
- Low-Density Separation
- Cluster Assumption

*1.2.5 Reinforcement Learning*

Reinforcement Learning (RL) is a type of learning that troubled with the smart agents to take an arrangements in the environment. In order to exploit the idea of increasing reward. This learning is one of the three basic Machine Learning patterns that collected with the supervised and unsupervised learning. As in alternative, the balance between to find and focus on the exploitation which is to the current knowledge and exploration tends to uncharted knowledge.

Types of Reinforcement Learning

- Value-based
- Model-based
- Policy-based

**1.3 Role of Machine Learning in Big Data**

In Machine Learning role, there are such always to the analytics each resolve the change on depending the business type and understanding about the organizations gain. The big data has clearly growing in popularly with this variety of recent survey [11]. The approaches in traditional analytical that have a trouble to manage a vast volumes of businesses in data and now it can collect as a significance, the results are not always mostly accurate.

To conflict with these issues, Organizations are revolving with the big data in machine learning techniques. The Machine Learning is one of the automated process that one is next to the impossible humans that were to try in their own. In view of considering the computer programming and statics that are evolved, human programmer could not create a predictive model at the same rate.

**2.4 Popular Machine Learning Language - Python**

Python is used in,

- Mathematics
- Web Development
- Software Development
- System Scripting


DOI: 10.30726/esij/v8.i3.2021.83018

ISSN 2394-7187(Online)
ISSN 2394 - 7179 (Print)

Difference between Python Syntax to other programming languages:

- Python, it is designed for readability and has some correspondences to the English language by influence from mathematics.
- Python uses new lines to complete a command, as opposed to other programming languages which often use semicolons or parentheses.
- Python depends on the indentation, whitespace and to define scopes ie., the loops, functions and classes. The other programming languages uses curly-brackets for the certain places in some purpose but in this language no need.

### 2.5 Framework of Big Data on Machine Learning

In a machine learning big data framework they have illustrated as in Figure.1. This framework explains the elaborate picture of machine learning, it deals with the big data problems. In this representation, User, Domain, Big Data and Systems are below diagram. The Machine learning it helps Pre-processing, Evaluation and Learning.

The Big Data input to the learning component and produces it to outputs, it becomes an amount of big data. The user can interrelate with the knowledge by providing the modules and personal preferences also in usability comments, by leveraging learning results that improves decision making, domain both can attend the a source of knowledge to central the learning process and the background of applying learned models. We concluded finally each and every components separately [13].

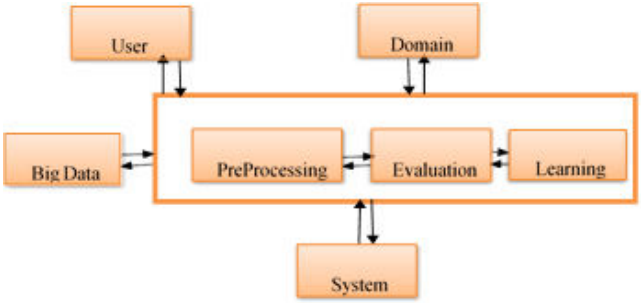


**Fig.1: The Framework of Big Data on Machine Learning**

## 2. Literature Review

**Table 1. Survey on Literature Review**

| Sl. No. | Paper And Author Details | Objective | Method |
|---|---|---|---|
| 1 | Machine Learning with Big Data: In Challenge and approaches, Alexandra L'Heureux, 2017. | To summarize the organizing relationship by challenges according to the Big Data. | Machine Learning |
| 2 | Droid Dolphin: The Dynamic Android Malware Detection Framework by Big Data and Machine Learning, Wein-Chieh Wu, et al.,2014. | Detection of Malware and to Protect Smartphone users using Big Data. | Android, Machine Learning, Big Data, Dynamic Analysis. |
| 3 | Big Data and Machine Based Secure Healthcare Framework, Prableen Kaur, et al., 2018. | Analysing the exact privacy weightage to and the security of Healthcare data roles. | Masking Encryption, Activity Monitoring control, Granular, Access Control, Dynamic Data Exchange. |
| 4 | Emergency Department patients with sepsis: A local Big Data-Driven, Machine Learning Approaches, R.Andrew Taylor, et al.,2015. | To Predict the Mortality Emergency Department in Hospital with sepsis. | Random Forest Model, Classification and Regression Tree (CART) Mode l. |
| 5 | Application of Big Data Machine Learning in Smart Grid and related security concerns: A Review, Elkas Hussai, et al., 2017. | Application of Big Data on Machine Learning in Electrical Power Grid then it Intending next Generation Power Grid. | Big Data Analysis, Cyber Security, IOT, Machine Learning, Smart Grid. |
| 6 | Harnessing the Power of Big Data: Infusing the scientific Method with the Machine Learning to convert ecology, Debra.P.C.Peter, et al., 2014. | To Find and focus on the data and Metadata sharing in Shortening timelag between individual discoveries that | KLAS, Hypothesis- driven, Data-intensive Method. |
| 7 | The Big Data Newsvendor: A Practical insights from the Machine Learning, Gah- Yi Ban, et al., 2019. | Handle large number of feature information and Investigate the data-driven Newsvendor problem. | ERM (Empirical Risk Minimization), KO(Kernel Weights Optimization) |


DOI: 10.30726/esij/v8.i3.2021.83018

ISSN 2394-7187(Online)
ISSN 2394 - 7179 (Print)

| | | | |
|---|---|---|---|
| 8 | Social Big Data: The Recent Achievements and New Challenges, Gema Bello- orgaz, et al., 2015. | Finding a problem in large number of research areas such as data storage, data representation, data processing. | Apache Hadoop, Apache Spark. |
| 9 | Applying a Spark founded Machine Learning Model on streaming Big Data for health status prediction, Lekha R.Nair, et al., 2015. | To develop a real time remote health remote status predictive system. | Streaming the data processing, Apache spark, Health Informatics and Tweet Processing. |
| 10 | A Survey on Architectures in big data and Machine Learning Algorithms in Healthcare, Gunasekaran Manogaran, et al., 2017. | To provide an overview of data architecture and state-of-the- art Machine Learning Algorithm. | Machine Learning, Healthcare and Big Data Architecture. |
| 11 | Machine Learning for Internet of things: A survey, Md. Saeid Mahadavinejad, et al., 2017. | Presentation of Taxanomy of Machine Learning Algorithms explaining different techniques. | IOT (Internet of Things), Smart Data. |
| 12 | Machine Learning with Big Data: An Efficient Electricity Generation Forecasting System, Md. Naimur Rahman, et al., 2016. | To prediction amount of power forecasting technique to improve the efficiency in Electricity Generation. | Artificial Neural Network, Electricity Forecast Generation, Back propogation, map Reduce, Hadoop. |
| 13 | A scalable Machine Learning Online service for Big data Real-time, Alejandro Baldominos, etal., 2014. | To develop and to experiment the scalable Machine Learning architecture. | Apache Spark and MapReduce. |
| 14 | Machine Learning on Big Data, Tyson Condie, et al., 2013. | Aim to discover the crops fertilizing research among the database and Machine Learning Community. | Tera Scale Learning. |
| 15 | From Machine Learning to Deep Learning: Progress in Machine Intelligence for rational drug discovery, Luzhang, et al., 2017. | To find and deal with the powerful early stage drug design and discover using Big Data | QSAR, ANN, Deep Learning. |
| 16 | A Machine Learning Approach to integrate Big Data for precision medicine in acute myeloid leukemia, paul B.de.Laat, et al., 2018. | Review about the different techniques are applied to the data and how the data publishing increasing nowadays. | SVM(Support Vector Machine), Merge Algorithm, MATLAB. |
| 17 | Pentuum: A new Platform for Distributes Machine Learning on Big Data, Eric P.Xiang, et al., 2015. | To propose a general purpose framework. | Distributed System, Model-Parallelism, BigModel. |

## 3. Algorithm using Accuracy Prediction

Naïve Bayes is a type of supervised learning algorithm. It is based on the Bayes theorem and is used for resolving the classification difficulties. This algorithm classifies one of the simplest effectiveness in classification algorithm and used for fast machine learning models that are able to take predictions quickly and make it to the ability. It is mainly used to the classifications in text that may contains the training data set as high in dimensionally. The probabilistic classifier, predicts the scheduled basis of probability of an object. Some example are Sentimental Analysis, Spam Filtration and Classifying Articles.

$$P(A|B)=\frac{P(B|A)P(A)}{P(B)}$$

- P(B|A): Likelihood probability-evidence assumed that the possibility of a hypothesis is true.
- P(A): Prior Probability-hypothesis previously detecting the evidence.
- P(B): Marginal Probability-Evidence.


DOI: 10.30726/esij/v8.i3.2021.83018

ISSN 2394-7187(Online)
ISSN 2394 - 7179 (Print)

### 3.1 Support Vector Machine

In Supervised Learning algorithms, it is one of the popular algorithm. In this, the classification and regression problems are used. The problems that primarily used in the Machine Learning classification. The aim of SVM Algorithm is to create a decision boundary and can split in n-dimensional space into the class. It is easily to put the data point newly and correct the category in future. The best of this type of boundary is called as hyper plane. Two

Types of SVM:
- Linear
- Non-Linear

### 3.2 Decision Tree Algorithm

Decision Tree Algorithm, it is a type of Supervised Algorithm used to solve the Regression and Classification. The goal in this is used to create a model training and to predict the value or class of target variable by a rule learning making decision that prior simply to thedata. Types of

Decision Trees are,
- Categorical Variable Decision Tree
- Continuous Variable Decision Tree

### 3.3 Random Forest Algorithm

Random Forest Algorithm, one of the popular Algorithm in Machine Learning is a supervised technique that is used for Regression and Classification problems in Machine Learning. The process of multiple classifiers combining to solve the complex problem and to improve the performance of models then based on collective Learning of concept. The Random Forest Classifier has a number of decision trees and it contains a several subsets of dataset assumes to take and to improve the predictive accuracy. The high accuracy that leads to the number of greater trees and prevents the problem for over fitting.

**Table 3. Predictive Analysis of Accuracy algorithm**

| Decision Tree | Random Forest | Naïve Bayes | Neural Network | SVM |
|---|---|---|---|---|
| 0.8114 | 0.7030 | 0.7457 | 0.7821 | 0.7890 |
| 0.8171 | 0.8143 | 0.8934 | 0.7000 | 0.7891 |
| 0.7914 | 0.7890 | 0.7534 | 0.7985 | 0.8390 |
| 0.7900 | 0.7909 | 0.7312 | 0.7123 | 0.8753 |
| 0.8115 | 0.8756 | 0.8665 | 0.8342 | 0.8001 |
| **0.8076** | **0.8764** | **0.7309** | **0.7601** | **0.7912** |

In Table 3 the accuracy of algorithm have been analysed and the prediction based on below the figure shows the Machine Learning such as Naïve Bayes, Neural Network, Random Forest, Support Vector Machine and Decision Tree [4].

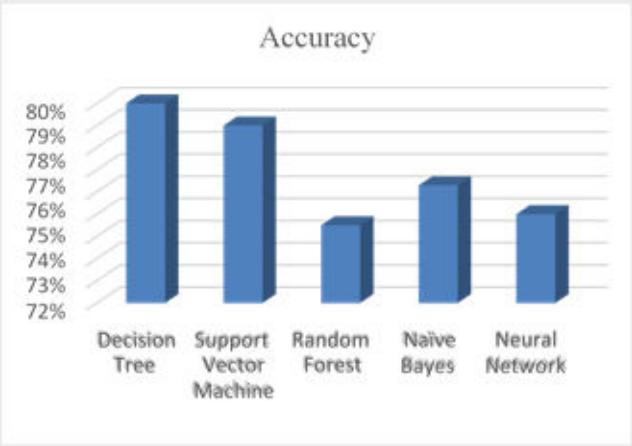


**Fig. 2: Accuracy Matrices Prediction using Machine Learning Algorithm**

## 4. Conclusion

In this review, we discussed about the big data processing, implementation and analysis are provided and then the overview of researches used in suitable areas also discussed. In Machine Learning, Tools, Types and Languages are used based on the content, scope, samples, advantages, methods and privacy are used for Learning of big data's concern that have existed and reviewed.

The algorithmic comparison given to predict the percentage level of classification in this paper. The results consumes revealed about the obtainable data, tools and techniques in the literature, which is presented. Moreover this techniques enhances in various sectors.

## References

[1] Al-Jarrah, Omar Y., et al. "Efficient machine learning for big data: A review." Big Data Research 2.3 (2015): 87-93.
[2] Hossain, Eklas, et al. "Application of big data and machine learning in smart grid, and associated security concerns: A review." IEEE Access 7 (2019): 13960-13988.
[3] Bhatnagar, Roheet. "Machine Learning and Big Data processing: a technological perspective andreview." International Conference on Advanced Machine Learning Technologies and Applications. Springer, Cham, 2018.
[4] Gupta, Preeti, Arun Sharma, and Rajni Jindal. "Scalable machine-learning algorithms for big data analytics: a comprehensive review." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 6.6 (2016): 194- 214.
[5] Sagiroglu, Seref, and Duygu Sinanc. "Big data: A review." 2013 international conference on collaboration technologies and systems (CTS). IEEE, 2013.
[6] George, Gerard, Martine R. Haas, and Alex Pentland. "Big data and management." (2014): 321-326.
[7] Ngiam, Kee Yuan, and Wei Khor. "Big data and machine learning algorithms for health-care delivery." The Lancet Oncology 20.5 (2019): e262-e273.
[8] Qiu, Junfei, et al. "A survey of machine learning for big data


DOI: 10.30726/esij/v8.i3.2021.83018

ISSN 2394-7187(Online)
ISSN 2394 - 7179 (Print)

processing." EURASIP Journal on Advances in Signal Processing 2016.1 (2016): 67.
[9] Angra, Sheena, and Sachin Ahuja. "Machine learning and its applications: A review." 2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC). IEEE, 2017.
[10] Al-Jarrah, Omar Y., et al. "Efficient machine learning for big data: A review." Big Data Research 2.3 (2015): 87-93.
[11] Hossain, Eklas, et al. "Application of big data and machine learning in smart grid, and associated security concerns: A review." IEEE Access 7 (2019): 13960-13988.
[12] Ma, Chuang, Hao Helen Zhang, and Xiangfeng Wang. "Machine learning for Big Data analytics in plants." Trends in plant science 19.12 (2014): 798-808.
[13] Miklosik, Andrej, and Nina Evans. "Impact of big data and machine learning on digital transformation in marketing: A literature review." IEEE Access (2020).
[14] L'heureux, Alexandra, et al. "Machine learning with big data: Challenges and approaches." IEEE Access 5 (2017): 7776-7797.
[15] Salkuti, Surender Reddy. "A survey of big data and machine learning." International Journal of Electrical & Computer Engineering (2088-8708) 10 (2020).
[16] Zhou, Lina, et al. "Machine learning on big data: Opportunities and challenges." Neurocomputing 237 (2017): 350-361.
[17] Divya, K. Sree, Peyakunta Bhargavi, and S. Jyothi. "Machine learning algorithms in big data analytics." International Journal of Computer Sciences and Engineering 6.1 (2018): 64-70.
[18] Beam, Andrew L., and Isaac S. Kohane. "Big data and machine learning in health care." Jama 319.13 (2018): 1317-1318.
[19] Ning, Chao, and Fengqi You. "Optimization under uncertainty in the era of big data and deep learning: When machine learning meets mathematical programming." Computers & Chemical Engineering 125 (2019): 434-448.
[20] Tu, Chunming, et al. "Big data issues in smart grid–A review." Renewable and Sustainable Energy Reviews 79 (2017): 1099-1107.
[21] Suthaharan, Shan. "Big data classification: Problems and challenges in network intrusion prediction with machine learning." ACM SIGMETRICS Performance Evaluation Review 41.4 (2014): 70-73.
[22 Zuo, Renguang, and Yihui Xiong. "Big data analytics of identifying geochemical anomalies supported by machine learning methods." Natural Resources Research 27.1 (2018): 5-13.
[23] Voyant, Cyril, et al. "Machine learning methods for solar radiation forecasting: A review." Renewable Energy 105 (2017): 569-582.
[24] Cabitza, Federico, Angela Locoro, and Giuseppe Banfi. "Machine learning in orthopedics: a literature review." Frontiers in bioengineering and biotechnology 6 (2018): 75.
[25] Manogaran, Gunasekaran, and Daphne Lopez. "A survey of big data architectures and machine learning algorithms in healthcare." International Journal of Biomedical Engineering and Technology 25.2-4 (2017): 182-211.


DOI: **10.30726/esij/v8.i3.2021.83018**