IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A Fundamental Study for Artificial Intelligence, Machine Learning and Big Data Analytics

Narender Singh, *Former Assistant Professor, Dept. of Computer Application,*
Rajiv Gandhi Mahavidhyalya, Uchana, Jind, Haryana, India, 126115

***Abstract-** In today's scenario, both Sciences and Industry are moving towards a data revolution. Therefore, this can result in complete facts of new formats and unparalleled data bases. These increments in amount of data have given rise to an opportunity for Machine Learning and Big data to come simultaneously and to increase Machine Learning methods that have the ability to carry present data types and for steering of large amount of information with minimal or no human interference. For processing of data implementing fast and effective algorithms and information driven models, Machine Learning is capable to give faultless results. Nowadays, as we find the utilisation result of Machine learning than it is positive as it is being vigorously used in a broad range of areas than we foresee. We can depict a pure Machine Learning process can be presented when more data is provided to the system, the more it can learn from it, returning the outcomes that are looking for, and due to this fact it works well with Big data. The Machine Learning can't keep running at its at most level without it and this is because of the way that with less information, the machine has less examples to gain from, and consequently its outcomes may be influenced. Popularity is gained by the ML based Big Data Processing and also several new developments are on the mount for efficient data processing. In order, to discover interestingness for decision making, this field is witnessing unparalleled emergence of new methods and approaches for efficient data processing. Therefore, we can conclude that now more and more ML based data processing approaches are being used for Big Data Processing. Firstly, we check the machine learning techniques and then highlight a few promising learning methods, and then we centre over the analysis and discussions with regards to the challenges and different possible solutions of machine learning for big data.*

***Index Terms*—Big Data, Machine Learning, Supervised Learning, Un-Supervised Learning, Artificial Intelligence.**

I. Introduction

For all business concern data is considered as their backbone. The amount of data is increasing at an unusual rate as per the result that we get out of the social media, Mobiles, Web technologies and Sensing services. Such as: it is truly exciting to see the amount of data that we deliver every day. At current rate, 2.5 quintillion bytes of data are created every day. Rapidly data is increasing day by day at a rapid pace. The approximate amount of 1.7 mega bytes is the estimated limit for the new data generated for every individual per second by 2020. According to the prediction from 2020, the accumulated information of Big data will boost up from 4.4 Zetta bytes to roughly 44 Zetta bytes or 44 trillion giga bytes [1]. With the help of this Big data there can be massive supplement in terms of business cost in variety of fields like medical field, financial services, health care, transportation and online advertising [2]. But if we consider traditional methods than we find that they are facing difficulty with this huge amount of data.

According to [3], in the year 2011, it can be depicted that the digital information has develop nine times in volume within 5 years and its value around the world will reach up to 35 trillion gigabytes by 2020. Hence, we can conclude that the term “Big Data” was tossed to confine the deep meaning of this data detonation trend.

Presently, various good surveys have been presented to elucidate what the big data refers to, and each of them depicts the big data from several perspectives, consisting several challenges and opportunities, background and research status, and analytics platforms [4]. McKinsey Global Institute (M, GI) has presented a comprehensive overview of the big data from three different angles, i.e., innovation, competition, and productivity with the help of above surveys. A number of more recent studies have investigated big data under particular context besides describing about these fundamental techniques and technology of big data. Such as, [5] gave a concise re-evaluation of the features of big data from Internet of Things (IoT). In wireless networks, there are some authors who have analyzed about the new quality of big data in wireless networks, i.e., in terms of 5G. In [6], as per the data mining prospective the authors offered several big data processing models and algorithms.

Since, several decades, machine learning techniques have been adopted widely in a number of immense and multifaceted data-intensive fields like medicine, astronomy, biology, and so on, for these techniques it also gives many possible solutions to extract the information that is hidden in the data. However, while the time for big data is coming, the set of data sets is so big and multifaceted that it has become difficult to deal with it by using conservative learning methods from the time when the established process of learning from conventional datasets was not framed towards and will not effort well with elevated volumes of data. For example, if we consider most of the established machine learning algorithms than they are planned for the data that would be entirely weighed down into reminiscence, which does not hold any more in the context of big data. Consequently, while learning from these several data is accepted that they will bring an important science and engineering advances beside with developments in superiority of our life, it brings excellent challenges at the same point of time.

II. MACHINE LEARNING

Learning to advance in the future later reliant on what was well-informed previously is termed as Machine Learning. Its main goal is to prepare learning algorithms that can do the learning as expected without any human support or administration. For artificial intelligence Machine learning is considered as its sub area which empowers software applications to get keen on a situation of self-learning devoid of being explicitly programmed [7]. When it is offered towards new data, than these systems are empowered to study, modify, develop and execute by themselves. Over the advancement of systems, machine learning mainly focuses by which it can get in to the information and utilize it from themselves [8]. With the help of Machine Learning data and trends can be identified. Its main objective is to help the computers to learn subsequently and change actions fittingly.

In the field of computing, the concept of Machine Learning is not considered new, though due to ever varying nature of needs of today's world as it has come up in a new 'Avatar' all jointly. Generally, we find that presently everyone is talking about ML based solution strategies for a given problem set. As we all know, that ML is considered as a subset of Artificial Intelligence, where computer algorithms are utilized to originally learn from facts and information. As the utilisation of Internet is increasing day by day a lot of digital information is being formed - which means that there is more data available for machines to analyse and 'learn' from. Thus, we can reach to a conclusion that the renaissance of Machine Learning. Nowadays, with the help of machine learning algorithms computers can communicate with humans, autonomously drive cars, write and publish sport match reports, and find terrorist suspects. In computer science, Machine learning (ML) is considered as the most growing field [9].

Some of the popular Machine learning techniques/methods are as follows: Classification, regression, topic modelling, time series analysis, cluster analysis, association rules, collaborative filtering, and dimensionality reduction [10]. With the help of these techniques one can perform analytics and predict the future trends based resting on the existing patterns and correlations between data in the known dataset.

A maturity model for describing advanced analytics and it also distinguishes analytical tools into three generations of machine learning and it was proposed by Agneeshwaran [11]. Following are the three generation of Machine learning are as follows:

1st Generation Machine Learning (1GML) needs the facts workload to adjust into memory of a single machine. Towards vertical scaling these tools are restricted which is considered as a negative aspect while considering Big Data. Generally, tools in this crowd were innovated earlier than Hadoop and are referred to as traditional analytical tools. (R, RapidMiner, KNIME, SAS, WEKA are some of the examples of 1GML tools).

2nd Generation Machine Learning (2GML) helps to develop 1GML with capabilities for distributed meting out diagonally Hadoop clusters. In 1GML context, facts remnants at its place whereas the code implementation is separated and processed resting on each necessary data node in parallel (Mahout (MapReduce) is an example).

3rd Generation Machine Learning (3GML) helps 2GML with abilities to powerfully implement distributed processing of iterative algorithms. This class is referred to as afar Hadoop (Mahout (Spark/H2O/Flink), MLlib, H2O ML, Flink-ML SAMOA, MADlib are some of the examples).

III. MACHINE LEARNING FOR BIG DATA

For artificial intelligence, Machine learning is regarded as a significant area. Whereas the main aim of machine learning is to invent information and develop intelligent decisions. Supervised, unsupervised, and semi-supervised are regarded as various categories of Machine learning algorithms. It is necessary to scale up machine learning algorithms if it focuses on big data [12]. Classification, regression, clustering, and density estimation, etc are included into another categorisation of machine learning according to the output of a machine learning systems. Decision tree learning, association rule learning, artificial neural networks, support vector machines (SVM), clustering, Bayesian networks, and genetic algorithms, etc these all techniques are included in Machine learning approaches.

Naïve Bayes, boosting algorithm, support vector machines (SVM), and maximum entropy method (MaxENT), etc these are few examples of supervised learning algorithms. Unlabelled data is utilised by unsupervised learning and classifies it by comparing data features. Unsupervised learning algorithms examples consist: clustering (*k*-means, density-based, and hierarchical, etc.), self-organizing maps (SOM), and adaptive resonance theory (ART) (Jaswant and Kumar, 2015) [13].

For both structured and unstructured data is a massive volume as it is so large that it is not easy to process using traditional database and software techniques. There are several impacts of Big data on various processes like scientific discoveries and value creation Massive parallel-processing (MPP), distributed file systems, and cloud computing, etc. support Big Data (Zaslavsky et al., 2012) [14]. Hadoop, Databases/Servers SQL, NoSQL, and MPP databases, etc. are also used to support Big Data despite of general cloud infrastructure services, technologies (Turk, 2012) [15].

A. *Machine Learning Tools for Big Data Analysis*

1. **Hadoop Ecosystem:** It is considered as an open source software structure for storing data and successively applications on clusters of product machine is having distributed file system (DFS). In an enormous web of projects related to each step of a big data workflow, it has risen up. Data collection, storage, processing, and much more are integrated in Hadoop. The area of structures that have been invented to either one reinstate or harmonize these novel rudiments has made the current description of Hadoop blurred. We need to consider both the framework itself and the environment that supports it

as to abundantly realize Hadoop. Some modules are already involved in the Hadoop framework [16].

Hadoop Common: Hadoop segments termed as the set of Java libraries and utilities that are required by other Hadoop segments. Apache Hadoop Framework is considered as the important model, beside with the Hadoop Distributed File System (HDFS), Hadoop Map Reduce and Hadoop YARN. Hadoop Common is having one more name that is Hadoop Core. *

Hadoop distributed file system (HDFS): Through several knobs of commodity machine or hardware a file outline deliberate to hoard huge quantities of data. It affords elevated throughput contact to function data. Integral fault tolerance capabilities are possessed inside this scheme. In case of disk failure, three or more copies of each data block are mainly maintained. Same as Google File System, HDFS has designed goals [17]. Both target at facts intense computing requests where massive data files are shared and get improved in duty of amazing continuous bandwidths in spite of less potential or latency, for improved maintain batch-processing style workloads.

Map Reduce: Map Reduce is regarded, as a software framework for applications, which is based on the LISP map and reduces primitives. It processes a huge quantity of data in parallel on the large cluster (millions of nodes) of commodity hardware. This entire process must be done in a reliable as well as fault tolerant way. Google introduced a connected execution of Map Reduce having parallel programming model [18]. Data mining, data analytics are some data analysis in which we use Map Reduce. On various parameters like efficiency, performance and other it is explored consistently. Generally, communication between nodes and distribution of tasks is not deal by Map Reduce programming. It refers to inscription a Map task function and Reduce task function. Hadoop program utilize these functions. Simultaneously, Several Map task functions can be performed. It is able to generate output as a list of intermediate values with its key as by taking key value pair as a source of data. This is considered as an element of the process that breaks the tasks. As an input data Reduce-task function takes the output of the Map-task function.

2. **APACHE HIVE:** HIVE is regarded as a data warehouse organizing tool that resides its position on top of Hadoop with regards to processing data query, analysis, and summarization and specifically used to process structured data. A SQL-type interface is provided by Hive to query data stowed in several file systems and databases that incorporate with Hadoop. To assimilate SQL-type Queries without implementing queries in the low-level Java API, HIVE provides an essential SQL abstraction. It also provides a Hive Query Language (HQL) is regarded as a SQL-type language. For Online Transaction Processing (OLAP) it is designed. It includes several features as it is familiar, scalable, fast, and extensible.
3. **APACHE PIG:** Usually it is used with Hadoop; Pig [19] is regarded as a notion over Apache Spark, Tez or Map Reduce. Moreover, with the help of APACHE PIG we can get a high-level platform to form programs run on Apache Hadoop. For this platform to write data analysis function, a high-level language known as Pig Latin is provided by Pig. By using Pig Latin, manipulation in operations on data in Hadoop can be performed. With the help of it we get the allowance for writing a data flow program by which data can be transformed easily. To perform an operation Pig Latin provides various operators such as sort, filter, join, and many other operations. Enormous operators to the user for developing their own function for reading, writing, and processing data are provided by the Pig Latin language. By using Ruby, Python, Java, or other scripting languages, users can extend Pig Latin. Pig can be run by the help of two commands one is 'Pig' command and second is 'java'.

For accessing Hadoop cluster, Map Reduce Mode- Map Reduce mode is regarded. Local Mode- By using file system all files are installed and run and a local host having contact to the sole machine in Local Mode.

4. **HBase:** HBase is replicated as a distributed database which is developed towards board prepared data in tables that ought to have billions of row and millions of columns.
5. **HCatalog:** A storage management layer for Hadoop which stores data in table format is termed as HCatalog. It provide pillar to diverse modules which are existing in Hadoop such as Hive, Map Reduce, and Pig which helps easily read and write data from the knot. For data archiving and cleaning tools it provide visibility. It also supports different types of file formats such as RC File, ORC, CSV, JSON these are different type of file formats that HCatalog supports.

B. Advanced Machine Learning Methods for Big Data

Many Machine Learning techniques are not comprehensive with regards to big data processing. In different words, according to different data it is regarded as a regular need to utilize specific learning methods. It is in the need to scale up Machine Learning algorithms when focused are on big data. Following are the different learning methods.

1. **Deep Learning:** We can call Deep Learning as Hierarchical Learning as it mainly utilizes Supervised and unsupervised Learning inside deep architectures to learn hierarchical representations. Deep Learning upside is considered as the program which assembles the data set by itself or by unsupervision [20]. The two mainstream deep learning approaches are Deep Belief Networks (DBNs) and Convolutional Networks (CNNs) are regarded as two basic mainstream approaches of deep learning. As the data is increasing, for large scale data sets with the increased computational and processing usage it also provide predictive analytics solutions.
2. **Feature learning:** If we consider High dimensional datasets than they have turned out to be augmented which has challenge the present learning to detach and classify the significant data from the information. Feature learning a solution is regarded as that can learn the useful representations of the data that makes uncomplicated and undemanding to acclimatize significant information. We regard both Feature learning or Representation learning

as a collection of mechanisms that helps a computer program to unsurprisingly discover the representations utilize for feature detection or arrangement from data that is collected and to study the features and use them to solve a specific problem [21]. The three main types of Representation Learning are Feature selection, Feature extraction, and Distance metric learning.

3. **Active Learning:** There are several conditions in which unlabeled information is high in comparison with manually labelling and getting labeled data is costly. In such situations, for labeled data learning algorithms can question the user. This is regarded as a sort of iterative supervised learning and it can be known as Active learning. This learning method is considered as the learning which can interact and ask the user to receive the impulse results at new data points. Hence, as the client chooses the samples, the amount of samples calculation concept can be much less than the amount needed in normal supervised learning.
4. **Ensemble Learning:** Business, social networks and other domains created amount of information which has increased extremely. The use of all these data can be useful only when if it is precisely performed by which the users can create correct resolution based on them. We can make, ensemble methods take a group of models into account, and will aggregate those models to give a final model despite of making one model and desiring that model as the best and realistic predictor.

   A collective whole is comprised inside the Ensemble learning where several techniques that is more than an isolated learning method. Four types of Ensemble Learning are Bagging, Boosting, Stacking, and Error-correcting.
5. **Distributed Learning:** In Distributed learning there is frequently exhilarating information that is covered up in enormous volumes of data. To utilize all the data to learn within a specific amount of time, learning from these new data has the incompetence of learning techniques. Distributed learning goals to provide a positive solution inside this position, whereas in allotting the learning process among different applications is a characteristic technique for escalate up the learning Algorithms. For enormous sum of data, disseminated and parallel learning methods have more stranded central points.

C. *Critical Problems of Machine Learning for Big Data*

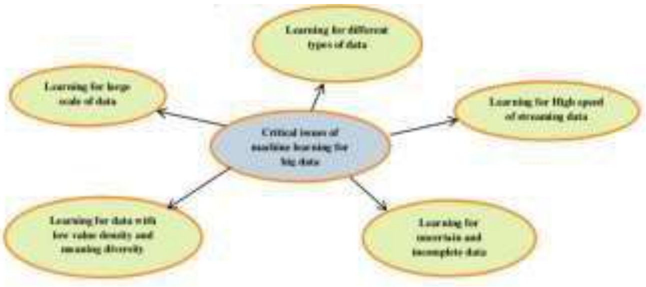
Learning for different types of data
Learning for large scale of data
Learning for High speed of streaming data
Critical issues of machine learning for big data
Learning for data with low value density and meaning diversity
Learning for uncertain and incomplete data

Figure 2. Critical problems of machine learning for big data

1) **Learning for large Scale of Data:** Critical problems-Big data main quality is the data size or volume. Those altruism main question or resist for machine learning. From our phones, computers, trains, buses, planes, parking meters, and social sites, large extent of data flooding occurred. Replicating example of digital or social media data, like: every day, roughly 24 petabytes (petabyte = 1024 × 1024 × 1024 × 1024 × 1024 bytes) of data as if we see Google than it alone required being processed [66]. If we talk about Facebook that it procedures and processes up to one million amount of photographs in every single moment. Some main prediction suggests that Facebook stowed 260 billion photos by using storage space of more than 20 petabytes. In adding up, if we maintain concentration over the former data sources, due to which the data scale will be very big. Beneath recent development tendencies, data collected by large organizations and analysed then it will definitely reach Petabyte to Exabyte (Exabyte = 1024 petabyte) or even more scale speedily.
2) **Learning for Different Types of Data:** Critical problem- Another feature of big data that makes it attractive as well as challenging is huge variation in data. Combine and generate non- linear, heterogeneous, and high dimensional data with various representation forms these several source from which structured, semi-structured and unstructured. Learning with this type of data set, the degree of complexity is not even conceivable before reaching here for this type of data set of learning’s.
3) **Learning for Fervid of Flooding Data Speed:** Critical problem- For big data velocity or speed truly matters. For machine learning this is one more emerging challenge. It is important for the system that they have to complete a work in real-world applications within the definite time otherwise outcome become not as much valuable or even they can be worthless. On data novelty the potential worth of data depends. For example, Agent-based autonomous exchange systems, stock market prediction, and earthquake prediction. In real-time in time- sensitive cases data needed to be handled.
4) **Learning for Incomplete and Uncertain Data:** Critical problem- Nowadays, with comparatively precise data from quite limited as well as renowned sources machine learning techniques usually nourished. The outcomes that we get from the learning have a tendency to be certain also. Hence, we can conclude that reliability has not ever been a harsh problem for apprehension. This can also be highlighted that the accuracy and faith of the input data hastily become a problem for the stark size of data which is accessible today. From various different ancestries data inputs are coming and data value is not fully confirmable. Hence, one other major difficulty with big data is reliability.
5) **Learning for Data with Meaning Diversity and Low Value Density:** Critical problem- To evaluate big data sets there are a variety of training methods that can be used. To extract pleasing information from massive volumes of data in the form of salable settlement as well as deep discernment is the main goal. As a leading big data feature its value is also categorised [22]. To originate considerable value from massive volumes of data a small worth density is not straight forward.

D. Examples of Machine Learning Applications in Big Data

The mixture of supervised and unsupervised machine learning techniques for creatively analyzing a large amount of crime data was projected. Three steps are included in the combination: dimensionality reduction, clustering, and classification. As being a powerful tool to deal with big data, R statistical software was used. Following are specific work that are outlined are as follows *(Nasridinov, 2014) [23]:

This method helps to measure the correlation between crime and social attributes. And help to reduce the dimensionality of the crime data.

Helps to divide crime data into several groups, it utilises the unsupervised machine learning method; *-*-to cluster the crime data into risky it utilises j- means clustering algorithm, average, and safe regions.

To predict whether a particular region is dangerous or safe it uses supervised machine learning technique; to perform predictions it uses decision tree classification algorithm. For society issues analysis and mining of social network was conducted by using Big Data. The process of analyzing, representing as well as extracting actionable patterns z.0 is termed as Social data mining.

+- from social network data. To classify the tweets it uses Machine learning and stemming algorithms. In the pattern of big data, Tweets can be replicated. From a collection of tweets the predicting features could be extracted; help to remove the stopping words; and also help to select all the keywords. The meaning of the tweets may be ambiguous as tweets are very short and may contain incomplete sentences. In machine learning, we can analyze all the data can be analysed by the help of support vector machines (SVM) as they are supervised models with related learning algorithms which are used for classification of the tweets. In text mining, Stemming algorithm uses a pre-processing task and it can also be utilize as a common need of natural language processing functions. Stemming algorithm was used to extract the main keywords or root words from the tweets, the Stemming algorithm is utilised. To predict the keywords from the tweets, the Stemming algorithm can be use. By the use of SVM algorithm we can classify all the keywords.

E. Challenges of Machine Learning Applications in Big Data

Machine learning General challenges are as follows:

i. Firstly it face a big challenge of designing scalable and flexible computational architectures for machine learning;

ii. Secondly, the capability to recognize the description of data before applying machine learning algorithms and tools; and

iii. At last but not the least, the capability to construct, learn and infer with the growing sample size, dimensionality, and categories of labels (Sukumar, 2014) [24]. Following are the main problems that depict the machine learning (ML) methods inappropriate for solving out the big data classification problems are as follows:

i. Firstly, an ML method is a method which is qualified on a scrupulous labeled datasets and might not be appropriate for another dataset – that the arrangement may not be vigorous over diverse datasets;

ii. Usually, an ML method is qualified by utilising a firm number of class types and hence, a great varieties of class types found in a vigorously growing dataset and it will lead to an inexact classification marks; and

iii. On the basis of single learning task an ML method is developed, and as a result they are considered not fit for today's several learning tasks and knowledge transfer requests of Big data analytics (Suthaharan, 2014)[25].

IV. ISSUES, CHALLENGES AND OPPORTUNITIES IN BIG DATA PROCESSING USING MACHINE LEARNING

Nowadays, as business requirements are changing day by day, and ever growing & diversified data poses innovative challenges towards the researchers. As the world has realised the prospective of discovering meaningful insights from unstructured data due to it Big Data Processing is also gaining prominence, over the past few years now. Hence, it is depicting a true result that for Machine Learning algorithms as which has been pressed to the front position and are aiding within making more timely & accurate predictions. To apprehend the value of Big Data, ML algorithms are used and to process the large facts volume at high velocity than ever before witnessing tremendous & unprecedented changes. In the field of ML, there is a continuous development as well but still the Models Scalability and Distributed Computing are some of the important challenges to ML implementations at some stage in Big Data Processing.

a) **Redundancy in Data:** Data duplication is termed as Redundancy in Data and it leads to incompatible data which can be disadvantageous to ML based system. For identifying duplicates some techniques do exist in a given data set [26] however these conventional methods are not so successful in case of Big Data. Like Dynamic Time Warping techniques are much more proficient than conventional Euclidian Distance algorithms.

b) **Noisy Data:** One of the primary source of noise in data consist missing or incorrect values, which may severely hamper the outcome of applying analytics over the data set containing noise. In case of Big Data Processing, Traditional mechanisms of removing noise from the data set fails due to their want for of scalability, and we cannot merely dispose of piercing data by deleting them as several very attractive insights that may be a part of them. To amplify scalability of outlier detection efforts are made for efficiently exploring anomalies in large data sets.

c) **Heterogeneous Nature of Data:** It is the Variety characteristics of Big Data that gather & present data collected from various sources, in different formats and are thus essentially heterogeneous in nature. Unstructured, text, audio and video data formats are regarded as diverse formats with regards to heterogeneous data that poses challenges to ML algorithms vis–vis their learning rate. It is not possible for us to treat all the sort of a data set uniformly important and concatenate them into one as it won't give a most favourable learning conclusion and most favourable performance. To learn from multiple views in parallel, Big Data is regarded as an opportunity and then learn the significance of attribute views w.r.t. the mission to be accomplished. Consequently, it will automatically boost towards the data outliers to deal with some optimization and data convergence issues. The collection and storage of mixed data based on different patterns or rules can be challenging in analysis of large scale data is regarded as the heterogeneous mixture data. The authors [27] present some solution to deal with such data where they formulate a state of 'heterogeneous mixture learning' – which is regarded as a complex form of analysis technology that was developed by NEC.

d) **Discretization of Data:** With the help of this process we can translate the quantitative data into qualitative data which result in a non-overlapping partition of continuous field. ML algorithms include some example like Decision Trees and Naïve Bayes which can only deal with discrete data. Attribute Discretization leads to classification of data which are efficient for learning task. Though at that point of time when it is trading with Big Data like conventional approaches are not regarded efficient.

e) **Data Labelling:** In data understanding, Annotations are considered significant but the process is pretty monotonous as data increases in size/dimension. For data labelling several alternative methods are been offered when dealing with Big Data, For example, Online Crowd-generated repositories which can serve up as a source for complimentary annotated training data. To address human-level concept learning, Probabilistic program induction is regarded as another approach. It will result in diminished performance if the user-specific context issue has not been addressed.

f) **Imbalance of Data:** As stratified random sampling methods is a Traditional method and it can be time consuming and also cannot efficiently support user-specified data set for value-based sampling. Hence, they also suffer failure in addressing Big Data and the solution is parallel to data sampling, which are based on multiple distributed index files.

g) **Feature Representation and Feature Selection:** ML algorithms performance is affected by the way the data is represented or features are selected (prominent feature identification). To handle Big Data the current algorithms for the above purpose are not sufficiently equipped.

V. TRENDS AND OPEN ISSUES IN BIG DATA PROCESSING USING MACHINE LEARNING

As we know at this point that ML based methods and their applications are a vital part of Big Data Processing, and it is regarded as a hot research area with many new developments occurrence in this direction. Even though by the help of research in ML based application development has achieved important results that are boosting deriving meaningful insights from Big Data, however much more is yet to be accomplished, in this imperative domain. Qiu et al. [28] illustrates the following future trends from diverse perspectives inside ML based applications for Big Data Processing.

1. **Data Meaning Perspective:** To achieve context-awareness, it implies as to how to make ML more intelligent.
2. **Pattern Training Perspective:** During the process of training patterns it implies how to evade the over fitting.
3. **Technique Integration Perspective:** With ML for Big Data Processing it deals with integrating other related techniques. For Big Data Processing, it is developing, a composite, integrated and seamless platform which have a great research potential.
4. **Privacy & Security Perspective:** For ensuring security and privacy in Big Data, it provides a research direction which is processing using ML techniques.
5. **Realization and Application Perspective:** To gain optimal results, how and where one must apply ML research in Big Data. To real world problems carries huge potential as research area by applying and utilizing the developed ML techniques. (8)

VI. CONCLUSION

In all science and engineering domains, Big Data is rapidly expanding nowadays. To bring significant opportunities and trans- formative potential for various sectors this was expected with the help of learning from these huge data. Conversely, to handle the data with the characteristics of large volume, different types, high speed, uncertainty and incompleteness, and low value density most of the traditional machine techniques are not inherently efficient or scalable enough. In reply, there is need for machine learning to reinvent itself for big data processing. With a short review of conventional machine learning algorithms this paper began, as followed by several current advanced learning methods. Then, a discussion concerning the challenges of learning with the big data and also some significant issues which are related to the big data processing by using the machine learning is being measured. The researchers can directly pick the work for further research and it will also help them in providing the concise review of the machine learning, big data and a variety of phases that are connected with the machine learning for big data.

REFERENCES

[1] Kairan Sun, Xu Wei, Gengtao Jia, Risheng Wang, and Ruizhi Li, "Large-scale Artificial Neural Network:MapReduce-based Deep Learning", arXiv:1510.02709v1 [cs.DC] 9 Oct 2015

[2] Ming Ke, Yuxin Shi, "Big Data, Big Change: In the Financial Management", Open Journal of Accounting, 2014, 3, 77-82

[3] J Gantz, D Reinsel, Extracting value from chaos (EMC, Hopkinton, 2011)

[4] D Che, M Safran, Z Peng, From big data to big data mining: challenges, issues, and opportunities, in Proceedings of the 18th International Conference on DASFAA (Wuhan, 2013), pp. 1–15

[5] Q Wu, G Ding, Y Xu, S Feng, Z Du, J Wang, K Long, Cognitive internet of things: a new paradigm beyond connection. IEEE Internet Things J 1(2), 129–143 (2014)

[6] X Wu, X Zhu, G Wu, W Ding, Data mining with big data. IEEE Trans Knowl Data Eng 26(1), 97–107 (2014)

[7] Lidong Wang, Cheryl Ann Alexander, "Machine Learning in Big Data", International Journal of Mathematical, Engineering and Management Sciences Vol. 1, No. 2, 52–61, 2016

[8] M. Rouse, "Machine Learning Definition," 2011. http://whatis.techtarget.com/definition/machine-learning

[9] Jordan, M.I., Mitchell, T.M.: Machine learning: trends, perspectives, and prospects. Science 349(6245), 255–260 (2015)

[10] Twardowski, B., Ryzko, D.: Multi-agent architecture for real-time big data pro- cessing. In: 2014 IEEE/WIC/ACM International Joint Conferences on Web Intel- ligence (WI) and Intelligent Agent Technologies (IAT), vol. 3, pp. 333–337. IEEE (2014)

[11] Agneeswaran, V.S., et al.: Big-data-theoretical, engineering and analytics perspec- tive. In: BDA, pp. 8–15. Springer (2012)

[12] Chen, C. L. P. & Zhang, C.-Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences*, *275*(10), 314-347.

[13] Jaswant, U. & Kumar, P. N. (2015). Big data analytics: a supervised approach for sentiment classification using mahout: an illustration. *International Journal of Applied Engineering Research*, *10*(5), 13447- 13457.

[14] Zaslavsky, A., Perera C. &Georgakopoulos, D. (2012). Sensing as a service and big data. *International Conference on Advances in Cloud Computing (ACC)*, Bangalore, India, July 2012, 1-8.

[15] Turk, M. (2012). A chart of the big data ecosystem, take 2. http://mattturck.com/2012/10/15/a-chart-of-the- big-data-ecosystem-take-2/

[16] Tom White, Hadoop The Definitive Guide, OREILLY, 2009.

[17] P. R. A. Preethi and P. J. Elavarasi, “Big Data Analytics Using Hadoop Tools – Apache Hive Vs Apache Pig,” vol. 24, no. 3, pp. 16–20, 2017.

[18] J. Talbot, R. M. Yoo, and C. Kozyrakis, “Phoenix++: Modular MapReduce for Shared-Memory Systems,” MapReduce ’11 Proc. Second Int. Work. MapReduce its Appl., pp. 9–16, 2011.

[19] G. Engelberg, O. Koren, and N. Perel, “Big data performance evaluation analysis using apache pig,” Int. J. Softw. Eng. its Appl., vol. 10, no. 11, pp. 429–440, 2016.

[20] PariwatOngsulee, “Artificial Intelligence, Machine Learning and Deep Learning”, 2017 Fifteenth International Conference on ICT and Knowledge Engineering

[21] Zhong G, Wang LN, Ling X, Dong J, “An Overview on Data Representation Learning: From Traditional Feature Learning to Recent Deep Learning”, The Journal of Finance and Data Science (2017), doi: 10.1016/j.jfds.2017.05.001

[22] H. Hu, Y. Wen, and X. Li, “A Framework for Big Data Analytics as a Scalable Systems,” IEEE Access, vol. 2, pp. 652–687, 2014.

[23] Nasridinov, A. (2014). Combining unsupervised and supervised machine learning to analyze crime data. *International Journal of Applied Engineering Research*, *9*(23), 18663-18669.

[24] Sukumar, S. R. (2014). Machine learning in the big data era: are we there yet? Conference: *ACM Knowledge Discovery and Data Mining: Workshop on Data Science for Social Good*, Oak Ridge National Laboratory, December 8, 2014, 1-5.

[25] Suthaharan, S. (2014). Big data classification: problems and challenges in network intrusion prediction with machine learning. *Performance Evaluation Review*, *41*(4), 70-73.

[26] C. Li, M. Georgiopoulos, and G. C. Anagnostopoulos, “A unifying framework for typical multitask multiple kernel learning problems,” IEEE Trans. Neural Networks Learn. Syst., vol. 25, no. 7, pp. 1287–1297, 2014.

[27] P. Le Callet, C. Viard-Gaudin, and D. Barba, “A convolutional neural network approach for objective video quality assessment,” IEEE Trans. Neural Networks, vol. 17, no. 5, pp. 1316–1327, 2006.

[28] K. Dwivedi, K. Biswaranjan, and A. Sethi, “Drowsy driver detection using representation learning,” Souvenir 2014 IEEE Int. Adv. Comput. Conf. IACC 2014, pp. 995–999, 2014.