Machine Learning for Big Data Processing: A Literature Review

Bharat Kumar Padhi[1], Dr. S S Nayak[2], Dr. B N Biswal[3]
[1]*PhD. Scholar, Centurion University of Technology & Management, Odisha*
[2]*Professor, Centurion University of Technology & Management, Odisha*
[3]*Director, Bhubaneswar Engineering College, Odisha*

***Abstract*- Big data has come from many years ago and due to the exponential growth of data from different sources, it becomes a challenge for find the significant sense out of it. Due to the Massive increasing of data which made difficult to store, process, analyze, interpret, consume and to make better decisions by overcoming the challenges in big data. Machine learning is a kind of artificial intelligence method to discover knowledge for making better decision intelligently. The further discussing about the issues and challenges present in big data. This paper presents an extensive literature study and review of machine leaning for big data processing. Then we have presented a road map of machine learning for processing big data. Finally, we have stated our future research direction.**

Index Terms- Machine Learning, Big Data.

I. INTRODUCTION

In the recent years there has been exponential growth of data from different sources from internet, smart phones or smart sensors, which has lead to big data. The term big data can be referred as the data which is massive, high speed, different categories and with lots unwanted noises that are very difficult to store, process, analyze, interpret, consume and make better decisions in the field of healthcare, finance, business or in multiple industries. Massive data has come from individuals use of computer, smart phones, gadgets which are used to share message & videos with friends in social media such face book, histogram,whats's up , for sharing short clips online, share their views and buy online products. The rise of Big Data applications where data collection has grown tremendously and is beyond the ability of commonly used software tools to capture, manage, and process within a "tolerable elapsed time" [1].

Even every individuals movements and their activities are been recorded by smart sensors which are placed in part of the cities and also in different public places.

The most fundamental challenge for Big Data applications is to explore the large volumes of data and extract useful information or knowledge for future actions. In many situations, the knowledge extraction process has to be very efficient and close to real time because storing all observed data is nearly infeasible [1].

Big Data starts with large-volume, heterogeneous, autonomous sources with distributed and decentralized control, and seeks to explore complex and evolving relationships among data which is known as HACE Theorem. These characteristics make it an extreme challenge for discovering useful knowledge from the Big Data [1].

One of the fundamental characteristics of the Big Data is the huge volume of data represented by heterogeneous and diverse dimensionalities. This is because different information collectors prefer their own schemata or protocols for data recording, and the nature of different applications also results in diverse data representations [1].

Though, there are many more traditional strategies struggle while dealing with massive data and learning from these large data brings significant opportunities for numerous sectors.

Machine learning (ML) is a kind of artificial methods which is used for discovering knowledge from massive data for making better intelligent decisions [10].

Machine learning algorithms categorize the learning task in three types such as supervised, unsupervised and reinforcement learning.

II. LITERATURE REVIEW

[1] Xindong Wu, Fellow et .al.[1] presented a HACE theorem that characterizes the features of the Big Data revolution, and proposes a Big Data processing model, from the data mining perspective. This data-driven model involves demand-driven aggregation of information sources, mining and analysis, user interest modeling, and security and privacy considerations.

[2] Desamparados Blazquez et .al.[2] developed a Big Data architecture that properly integratesmost of the non-traditional information sources and data analysis methods in order to provide a specifically designed system for forecasting social and economic behaviors, trends and changes.

[3] Junfei Qiu et. al.[3] presented a literature survey of the latest advances in researches on machine learning for big data processing and highlight some promising learning methods in recent studies, such as representation learning, deep learning, distributed and Also, analysis and discussions about the challenges and possible solutions of machine learning for big data.

[4] Philip Chen et. al. [4] aimed to demonstrate a close-up view about Big Data, including Big Data applications, Big Data opportunities and challenges, as well as the state-of-the-art techniques and technologies to deal with the Big Data problems. And also discussed several underlying methodologies to handle the data deluge.

[5] Ming Ke et. al. [5] makes an analysis about what change that "Big Data" brings to Accounting Data Processing, Comprehensive Budget Management, and Management Accounting through affecting the idea, function, mode, and method of financial management and also stated the challenges that "Big Data" brings to enterprise aiming to illustrate that only through fostering strengths and circumventing weaknesses can an enterprise remain invincible in "Big Data" era.

[6] Jianwu Wang et. al [6] discussed the challenges and opportunities of Big Data provenance related to the veracity of the datasets themselves and the provenance of the analytical processes that analyze these datasets and explained tracking and utilizing Big Data provenance using workflows as a programming model to analyze Big Data.

[7] Khine et. al. [7] presented the nature of big data and how organizations can advance their systems with big data technologies. By improving the efficiency and effectiveness of organizations, people can benefit the can take advantages of a more convenient life contributed by Information Technology.

[8] Zoila Ruiz et. al. [8] analyzed different the use of Machine Learning (ML) for processing large volumes of data (Big Data).

[9] Al-Jarrah et. al [9] reviewed the theoretical and experimental data-modeling literature, in large-scale data-intensive fields, relating to: model efficiency, including computational requirements in learning, and data-intensive areas' structure and design, and introduces new algorithmic approaches with the least memory requirements and processing to minimize computational cost, while maintaining/improving its predictive/classification accuracy and stability.

[10] D. Saidulu et. al. [10] discussed various types of data types, learning methods, vital issues in big data processing and application of machine learning approaches in big data.

III. MOTIVATION TO THE PROBLEM

Big Data problems involves in most of the scenarios e.g. global economy, society administration, national security and traditional strategies struggle when deal with this large data, varying types, high speed and uncertainty. Learning from massively large data can brings significant opportunities for different sectors in business, health, climate, bio, medicine and many more [9].

Most of the traditional machine learning techniques are good for processing structured data but they are lacking in processing unstructured data which requires massive computational efficiency, more scalability to handle the data with massive volume, varying types, great speed, uncertainty, inconsistency and incompleteness [9]. Therefore to design more optimal techniques which can solve huge sized unstructured data efficiently.

IV. MACHINE LEARNING

Machine Learning is an idea to learn from examples and experience, without being explicitly programmed. Instead of writing code, you feed data to the generic algorithm, and it builds logic based on the data given [35].

Life without machine Learning-when we search information on any topics on google,it will collect all the information what we want and present it to us accordingly but if there was no goggle , then we have to go through different books ,articles and even we could find the relevant answer.

Life with Machine Learning- we are getting easily connects with our old friends by associating one friend with others is like Facebook and shopping product from different vendors online like from amazon, flip kart etc which has made shopping easy.

Machine learning is a data analytics technique that teaches computers to do what comes naturally to humans and animals: learn from experience. Machine learning algorithms use computational methods to "learn" information directly from data without relying on a predetermined equation as a model. The algorithms adaptively improve their performance as the number of samples available for learning increases. Deep learning is a specialized form of machine learning [36].

Machine Learning is a field which is raised out of Artificial Intelligence (AI). Applying AI, we wanted to build better and intelligent machines. But except for few mere tasks such as finding the shortest path between point A and B, we were unable to program more complex and constantly evolving challenges. There was a realisation that the only way to be able to achieve this task was to let machine learn from itself. This sounds similar to a child learning from its self. So machine learning was developed as a new capability for computers. And now machine learning is present in so many segments of technology, that we don't even realise it while using it [35].

Example:

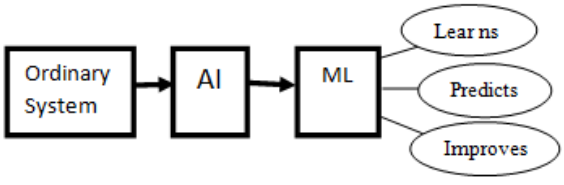

Fig1. Application of Machine Learning

Suppose we provide a system with input data that contains the photo of the students of a university. Then we do:

- First analyze data. Then it tries to find patterns such as dress color, height, size...etc.
- Based on these patterns, the system tries to predict different types of student belong to particulars course, fittest student...etc and partition them.
- Finally, it keeps all tracks of the decisions; it took to make sure that it is learning. Then next time when we ask the machine to predict and segregate the different types of student. Then it does not go through the entire processes again. That's the machine learning works.

With the rise in big data, machine learning has become a key technique for solving problems in areas [36], such as:

1. Automotive, aerospace, and manufacturing, for predictive maintenance.
2. Computational biology, for tumor detection, drug discovery, and DNA sequencing.
3. Computational finance, for credit scoring and algorithmic trading.
4. Energy production, for price and load forecasting.
5. Image processing and computer vision, for face recognition, motion detection, and object detection
6. Natural language processing, for voice recognition applications.

Machine leaning is a field of research that formally focuses on the theory, performance, and properties of learning systems and algorithms [3].

Generally, the field of machine learning is divided into three sub domains:

1. Supervised learning.
2. Unsupervised learning.
3. Reinforcement learning.

1. Supervised learning:

The machine learns from the training data that is labeled. So we have to supervise machine learning while training it to work by its own. Supervised learning requires training with labeled data which has inputs and desired outputs.

Supervised learning uses classification and regression techniques to develop predictive models.

Classification techniques predict discrete responses—for example, whether an email is genuine or spam, or whether a tumor is cancerous or benign. Classification models classify input data into categories. Typical applications include medical imaging, speech recognition, and credit scoring.

Regression techniques predict continuous responses—for example, changes in temperature or fluctuations in power demand. Typical applications include electricity load forecasting and algorithmic trading.

Use regression techniques if you are working with a data range or if the nature of your response is a real number, such as temperature or the time until failure for a piece of equipment.

2. Unsupervised learning

The machine learns from the training data but without labeled. Unsupervised learning finds hidden patterns or intrinsic structures in data. It is used to draw inferences from datasets consisting of input data without labeled responses [36].

Clustering is the most common unsupervised learning technique. It is used for exploratory data analysis to find hidden patterns or groupings in data. Applications for cluster analysis include gene sequence analysis, market research, and object recognition [36].

For example, if a cell phone company wants optimize the locations where they build cell phone towers, they can use machine learning to estimate the number of clusters of people relying on their towers. A phone can only talk to one tower at a time, so the team uses clustering algorithms to design the best placement of cell towers to optimize signal reception for groups, or clusters, of their customers.

3. Reinforcement Learning:

The machine learns on its own i.e. by its mistake and experiences.

Suppose a new born baby put the finger to the burning candle flame which hurts, so next time when the baby sees the candle burning, then it recalls what has happen last time and would repeat again. This is how reinforcement learning works.

Table2. Comparison of machine learning methods

Learning Methods	Data processing tasks	Learning algorithm
Supervised	Classification/ Regression/Estimation	Support Vector Machine
		Naive Bayes
		Hidden Markov Model
		Bayesian network
Unsupervised	Clustering/ Prediction	Neural Networks
		K-means
		Gaussian Mixture Model
Reinforcement	Decision making	Q-learning
		R-learning
		TD learning

The right kind of machine learning solution depends on:

1. The problem statement: If the problem is to predict the future stock market price, the supervised learning would work best.
2. The size, quality & nature of the data: If the data is clotted, then we can go for unsupervised learning. If the data is categorised manner, then we go for supervised learning.
3. Complexity of the algorithm: If we are going for predicting stock market price ,then we can go for reinforcement learning which would be vary time consuming then go for supervised learning.

V. MACHINE LEARNING ALGORITHM

There are various algorithms in machine learning but key algorithms are k- Nearest Neighbors, linear regression, decision tree and naive bayes.

1. k- Nearest Neighbors: k-Nearest Neighbors (kNN) algorithm works in a way that a new data point is assigned to a neighboring group to which it is most similar.

It can be used for both classification and regression problems. However, it is more widely used in classification problems in the industry. K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases by a majority vote of its k neighbors. The case being assigned to the class is most common amongst its K nearest neighbors measured by a distance function [37].

In k-Nearest Neighbors, 'k' can be an integer greater than 1.So , for every new data point we want to classify , we compute to which neighboring group it is closest to it or similar to it [37].

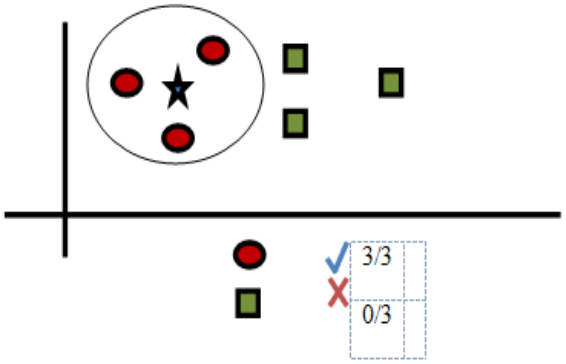

Fig 2. kNN

Things to consider before selecting kNN:

- kNN is computationally expensive.
- Variables should be normalized else higher range variables can bias it.
- Works on pre-processing stage more before going for kNN like outlier, noise removal.

2. Linear Regression

Linear regression is a process used for estimating the relationships among variables. Here, one of the variables is dependent on one or more variables.

It is used to estimate real values (cost of houses, number of calls, total sales etc.) based on continuous variable(s). Here, we establish relationship between independent and dependent variables by fitting a best line. This best fit line is known as regression line and represented by a linear equation:

Y= a *X + b ……..eq.(1)

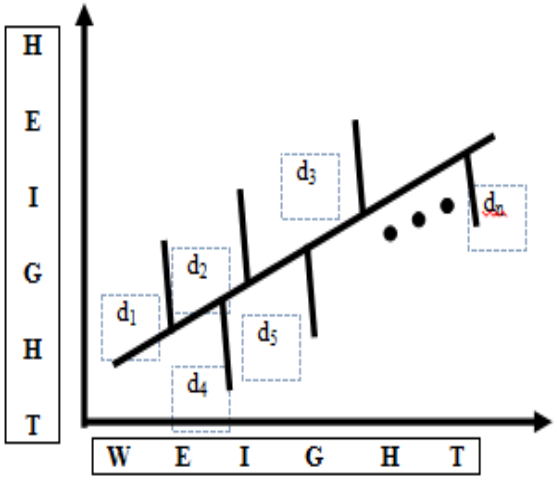

Fig 3., Linear Regression

D=d12 +d22+.....+dn2 ……..eq.(2)

The regression line has the least value of D.

3. Decision Tree

This is one of my favorite algorithm and I use it quite frequently. It is a type of supervised learning algorithm that is mostly used for classification problems. Surprisingly, it works for both categorical and continuous dependent variables [37].

A decision tree is a graph that uses a branching method to illustrate every possible outcomes of a decision. By using branching method, it realizes the problem and makes the decision based on the conditions.

Suppose, I am sitting at home and thinking for swimming, then so

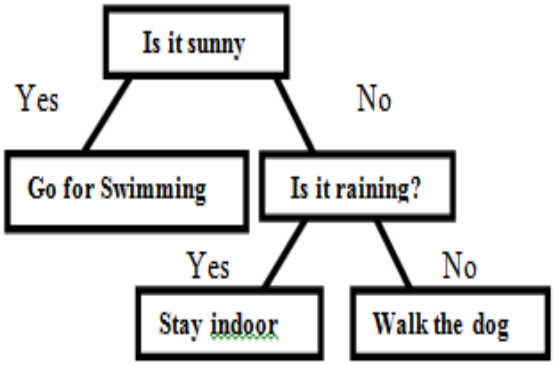

Fig 4. Decision Tree

Therefore we follow the decision tree everyday and realise the problem and take the decision accordingly.

4.Naive Bayes

Naive Bayes is a classification technique based on Bayes' theorem with an assumption of independence between predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature [37].

Naive Bayes classification is done when there is larger dataset. The naive Bayes classifier technique is based on conditional probability and it particularly suited when the complexity of the inputs is high.

$$P(C|A) = \frac{P(C|A)\ P(C)}{P(C)} \quad \text{.....eq.(3)}$$

Therefore, algorithms are not the solution to the particular problem but they are methods to solve the particular problem.

VI. BIG DATA

Big data is a term for data sets that are so large or complex that traditional data processing application software is inadequate to deal with them. Big data challenges include capturing data, data storage, data analysis, search, sharing, transfer, visualization, querying, updating and information privacy [10].
Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making. ("Gartner IT Glossary, n.d.")

Here we describe the 8 V's below.

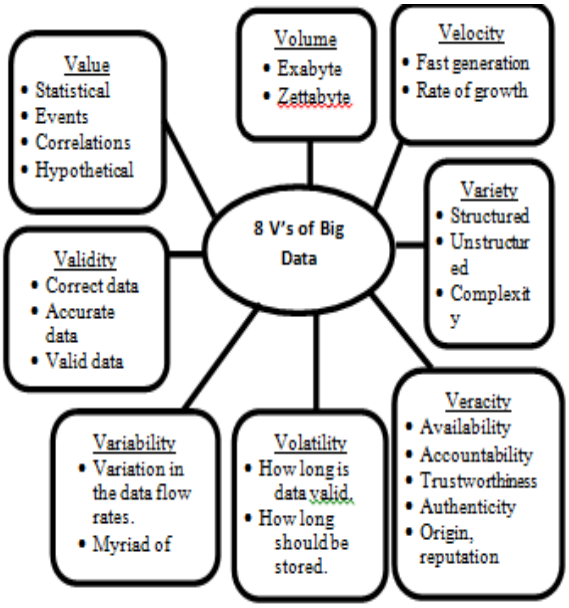

Fig 5. 8 V's of Big data

1. Volume:
Volume refers to the magnitude of data. Big data sizes are reported in multiple terabytes and petabytes. A survey conducted by IBM in mid-2012 revealed that just over half of the 1144 respondents considered datasets over one terabyte to be big data (Schroeck, Shockley, Smart, Romero-Morales, & Tufano, 2012). One terabyte stores as much data as would fit on 1500 CDs or 220 DVDs, enough to store around 16 million Facebook photographs. Beaver, Kumar, Li, Sobel, and Vajgel (2010) report that Facebook processes up to one million photographs per second. One petabyte equals 1024 terabytes. Earlier estimates suggest that Facebook stored 260 billion photos using storage space of over 20 petabytes [24].

2. Variety:
Variety refers to the structural heterogeneity in a dataset. Technological advances allow firms to use various types of structured, semi-structured, and unstructured data. Structured data, which constitutes only 5% of all existing data (Cukier, 2010), refers to the tabular data found in spreadsheets or relational databases. Text, images, audio, and video are examples of unstructured data, which sometimes lack the structural organization required by machines for analysis. Spanning a continuum between fully structured and unstructured data, the format of semi-structured data does not conform to strict standards. Extensible Markup Language (XML), a textual language for exchanging data on the Web, is a typical example of semi-structured data. XML documents contain user-defined data tags which make them machine-readable [24].

3. Velocity:
Velocity refers to the rate at which data are generated and the speed at which it should be analyzed and acted upon. The proliferation of digital devices such as smart phones and sensors has led to an unprecedented rate of data creation and is driving a growing need for real-time analytics and evidence-based planning. Even conventional retailers are generating high-frequency data. Wal-Mart, for instance, processes more than one million transactions per hour (Cukier, 2010). The data emanating from mobile devices and flowing through mobile apps produces torrents of information that can be used to generate real-time, personalized offers for everyday customers. This data provides sound information about customers, such as geospatial location, demographics, and past buying pat-terns, which can be analyzed in real time to create real customer value [24].

4.Veracity:
IBM coined Veracity as the fourth V, which represents the unreliability inherent in some sources of data. For example, customer sentiments in social media are uncertain in nature, since they entail human judgment. Yet they contain valuable information. Thus the need to deal with imprecise and uncertain data is another facet of big data, which is addressed using tools and analytics developed for management and mining of uncertain data.

5.Variability (and complexity):
SAS introduced Variability and Complexity as two additional dimensions of big data. Variability refers to the variation in the data flow rates. Often, big data velocity is not consistent and has periodic peaks and troughs. Complexity refers to the fact that big data are generated through a myriad of sources. This imposes a critical challenge: the need to connect, match, and transform data received from different sources [24].

6.Validity:
Validity of data may sound similar to veracity of data. However, they are not the same concept but similar. By validity, we mean the correctness and accuracy of data with regard to the intended usage. In other words, data may not have any veracity issues but may not be valid if not properly understood. Critically speaking, same set of data may be valid for one application or usage and then invalid for another application or usage. Even though, we are dealing with data where relationship may not be defined easily or in initial stages, but it is very important to verify relationship to some extent between elements of data, which we are dealing with to validate it against intended consumption, as possible. As an example given in, Can a physician simply take a data from clinical trial that is related to patient's disease symptoms without validating them? The answer is No. Another example from, we can verify or validate the storm potential in some areas predicated by Weather Satellites along with Tweets to see how much impact is going to be on individuals [36].

7.Volatility:
Speaking of volatility of big data, we can easily recall the retention policy of structured data that we implement every day in our businesses. Once retention period expires, we can easily destroy it. As an example: an online ecommerce company may not want to keep a 1 year customer purchase history. Because after one year and default warranty on their product expires so there is no possibility of such data restore ever. Big data is no exception to this rule and policy in real world data storage. Such issue is very much magnified in big data world and not as easy as we have dealt with it in traditional data world. Big data retention period may exceed and storage and security may become expensive to implement. Actually, Volatility becomes significant due to Volume, Variety and Velocity of data [36].

8.Value:
Oracle introduced Value as a defining attribute of big data. Based on Oracle's definition, big data are often characterized by relatively "low value density". That is, the data received in the original form usually has a low value relative to its volume. However, a high value can be obtained by analyzing large volumes of such data [24].

VII. CHALLENGES OF BIG DATA PROCESSING

The challenges in big data handing are [22]:

1. Heterogeneous data sources
   Data required for analytical and computational purpose is strongly heterogeneous which possess typical integration problem.
2. Unstructured nature of data sources
   Data from different sources are unstructured and it is a big challenge to handle. Transformation of unstructured data into a suitable and structured format in order to extract meaningful data.
3. High scalability
   Data is scaling at an unprecedented rate and it is a challenging issue as data volume is increasing faster than compute resources and CPU speed is static.
4. Timeliness
   With the increasing volume of data the time to analyze the data will also increase.
5. Privacy & Confidentiality
   The privacy of data is another major and important concern in the context of big data. Big data contains large amount of sensitive and personal information that may be exposed to privacy and confidentiality.
6. Data Integration
   Big data integration is multidimensional and multidisciplinary which requires multi-technology methods that solve the big data challenges.

VIII. MACHINE LEARNING FOR BIG DATA PROCESSING

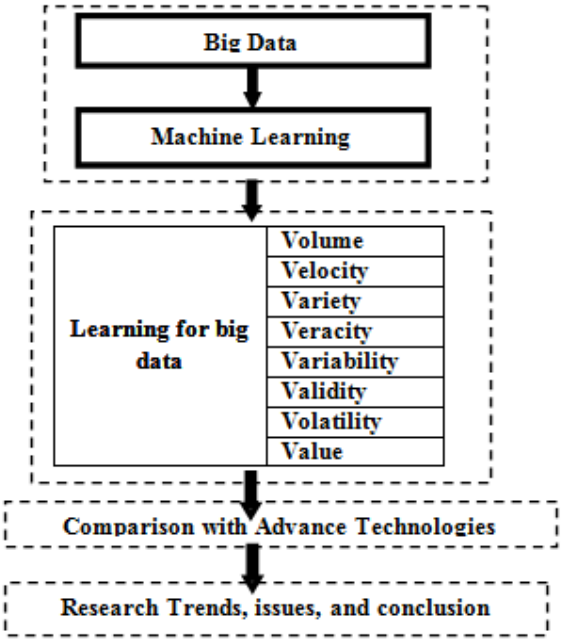

Fig 6. Road Map of Machine learning for big data processing

The figure provide a comprehensive study for processing the challenges of big data by machine learning approaches and mainly eight aspects of big data. It also gives comparison with the advance technologies for big data. Find out the open issues and research trends. Conclusions are drawn.

IX. ISSUES OF MACHINE LEARNING FOR BIG DATA PROCESSING

A review about the critical concerns of machine learning procedures for big data processing is presented in fig. it includes:

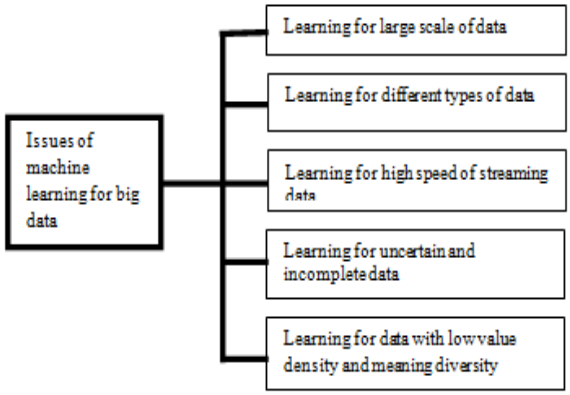

Fig 7. Issues of ML for Big data processing

X. CONCLUSION

Now we are in an era of big data, processing and analysis of massive sized unstructured, inconsistent, incomplete and imprecise data by computing machine is a big challenging task. To perform operation and analysis incompleteness in a data, present in higher dimensions may be very complicated or complex. In future, we will present a new efficient model of machine leaning for big data processing which will solve the challenges.

REFERENCES

[1] Xindong Wu, Fellow, IEEE, Xingquan Zhu, Senior Member, IEEE,Gong-Qing Wu, and Wei Ding, Senior Member, IEEE. "Data Mining with Big Data", IEEE transactions on knowledge and data engineering, vol. 26, no. 1, january 2014.

[2] Desamparados Blazquez, Josep Domenech*. "Big Data sources and methods for social and economic analyses". Technological Forecasting & Social Change 130 (2018) 99–113.https://www.sciencedirect.com/science/article/pii/S0040162517310946.

[3] Junfei Qiu, Qihui Wu, Guoru Ding*, Yuhua Xu and Shuo Feng. " A survey of machine learning for big data Processing". Qiu et al. EURASIP Journal on Advances in Signal Processing (2016) 2016:67 DOI 10.1186/s13634-016-0355-x.

[4] Philip Chen, C. L., & Zhang, C.-Y. (2014). "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data". Information Sciences, 275, 314–347. DOI:10.1016/j.ins.2014.01.015.

[5] Ming Ke, Yuxin Shi, Beijing Wuzi University, Beijing, China." Big Data, Big Change: In the Financial Management". Open Journal of Accounting, 2014, 3, 77-82 Published Online October 2014 in SciRes. http://www.scirp.org/journal/ojacct and http://dx.doi.org/10.4236/ojacct.2014.34009.

[6] Jianwu Wang, Daniel Crawl, Shweta Purawat, Mai Nguyen, Ilkay. " Big Data Provenance: Challenges, State of the Art and Opportunities". 2015 IEEE International Conference on Big Data (Big Data)

[7] Khine, P.P. and Shun, W.Z. (2017) "Big Data for Organizations: A Review". Journal of Computer and Communications, 5, 40-48.https://doi.org/10.4236/jcc.2017.53005.

[8] Zoila Ruiz, Jaime Salvador, and Jose Garcia-Rodriguez "A Survey of Machine Learning Methods for Big Data". Springer International Publishing AG 2017, J.M. Ferr´andez Vicente et al. (Eds.): IWINAC 2017, Part II, LNCS 10338, pp. 259–267, 2017.DOI: 10.1007/978-3-319-59773-7 27.

[9] Al-Jarrah, Omar Y. et al. "Efficient Machine Learning for Big Data: A Review." Big Data Research 2 (2015): 87-93.

[10] D. Saidulu , Dr. R. Sasikala. "Machine Learning and Statistical Approaches for Big Data: Issues, Challenges and Research Directions". International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 21 (2017) pp. 11691-11699.

[11] Andrea De Mauro Marco Greco and Michele Grimaldi. "What is Big Data? A Consensual Definition and a Review of Key Research Topics". Conference Paper • September 2014 DOI: 10.13140/2.1.2341.5048.

[12] D. P. Acharjya , Kauser Ahmed P. "A Survey on Big Data Analytics: Challenges, Open Research Issues and Tools". (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 2, 2016.

[13] Andrea De Mauro, Marco Greco and Michele Grimaldi . "A formal definition of Big Data based on its essential features". March 2016 DOI: 10.1108/LR-06-2015-0061.

[14] Ver_onica Bol_on-Canedo, Beatriz Remeseiro, Konstantinos Sechidis et al. "Algorithmic challenges in Big Data analytics". ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1. http://www.i6doc.com/en/.

[15] Wullianallur Raghupathi and Viju Raghupathi ." Big data analytics in healthcare: promise and Potential". Health Information Science and Systems 2014,2:3http://www.hissjournal.com/content/2/1/3.

[16] Puneet Singh Duggel,Sanchita Pual, "Big Data Analytics: challenges and solution". International conference on Cloud, Big data and Trust 2013, Nov 13-15 RGPV.

[17] Chanchal Yadav, Shullang Wang,Manoj Kumar, "Algorithm and Approaches to handle large Data- A Survey " IJCSN,Vol-2 issue 3, 2013, ISSN: 2277-5420.

[18] Richa Gupta ,Sunny Gupta ,Anuradha Singhal , " Big Data : Overview ". IJCTT, Vol 9,Number 5,March 2014.

[19] Che D., Safran M., Peng Z. (2013) " From Big Data to Big Data Mining: Challenges, Issues, and Opportunities". In: Hong B., Meng X., Chen L., Winiwarter W., Song W. (eds) Database Systems for Advanced Applications. DASFAA 2013. Lecture Notes in Computer Science, vol 7827. Springer, Berlin, Heidelberg. https://link.springer.com/chapter/10.1007/978-3-642-40270-8_1.

[20] Jaseena K.U, Julie David. "Issues, Challenges and Solutions: Big Data Mining". December 2014, DOI, 10.5121/csit.2014.41311,Sixth International Conference on Networks & Communications.https://www.researchgate.net/publication/301468821_Issues_Challenges_and_Solutions_Big_Data_Mining.

[21] Khushboo Wadhwani , Dr. Yun Wang . "Big Data Challenges & Solutions". February 2017, DOI:10.13140/RG.2.2.16548.88961.https://www.researchgate.net/publication/313819009_Big_Data_Challenges_and_Solutions.

[22] Neelam Singh, Neha Garg and Varsha Mittal . " Big Data –insights, motivation and challenges". IJSER,Vol4,Issue12,December2013.https://www.ijser.org/researchpaper/Big-Data-insights-motivation-and-challenges.pdf.

[23] M. K.Kakhani, S. Kakhani and S. R.Biradar, "Research issues in big data analytics". International Journal of Application or Innovation in Engineering & Management, 2(8) (2015), pp.228-232.

[24] A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics". International Journal of Information Management, 35(2) (2015), pp.137-144.

[25] R. Nambiar, A. Sethi, R. Bhardwaj and R. Vargheese, "A look at challenges and opportunities of big data analytics in healthcare". IEEE International Conference on Big Data, 2013, pp.17-22

[26] Alexandra L'Heureux, Katarina Grolinger, Hany F. ElYamany, Miriam A. M. Capretz. " Machine Learning with Big Data: Challenges and

Approaches".DOI10.1109/ACCESS.2017.2696365, IEEE Access.
[27] M. A. Beyer and D. Laney, "The Importance of 'Big Data': a Definition," Gartner Research Report, 2012.
[28] H. V Jagadish, J. Gehrke, A. Labrinidis, Y. Papakonstantinou, J. M. Patel, R. Ramakrishnan, and C. Shahabi, "Big Data and its Technical Challenges," Communications of the ACM, vol. 57, no. 7, pp. 86–94, 2014.
[29] M. James, C. Michael, B. Brad, and B. Jacques, "Big Data: The Next Frontier for Innovation, Competition, and Productivity," The McKinsey Global Institute, 2011.
[30] M. Rouse, "Machine Learning Definition," 2011. [Online].Available:http://whatis.techtarget.com/definition/machine-learning.
[31] S. R. Sukumar, "Machine Learning in the Big Data Era: Are We There Yet?," in Proceedings of the 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining : Workshop on Data Science for Social Good (KDD 2014), 2014.
[32] J. Qiu, Q. Wu, G. Ding, Y. Xu, and S. Feng, "A Survey of Machine Learning for Big Data Processing," EURASIP Journal on Advances in Signal Processing, vol. 67, pp. 1–16, 2016.
[33] M. A. u. d. Khan, M. F. Uddin, and N. Gupta, "Seven V's of Big Data understanding Big Data to extract value," in Proceedings of the 2014 Zone 1 Conference of the American Society for Engineering Education, 2014, pp. 1–5.
[34] P. Domingos, "A Few Useful Things to Know About Machine Learning," Communications of the ACM, vol. 55, no. 10, p. 78, 2012.
[35] https://towardsdatascience.com/introduction-to-machine-learning-db7c668822c4.
[36] https://in.mathworks.com/discovery/machine-learning.html.
[37] https://stats.stackexchange.com/questions/144154/supervised-learning-unsupervised-learning-and-reinforcement-learning-workflow.
[38] Cornelia L. Hammer,Duane C. Kostroch,Gabriel Quiros, and STA internal group, "Big Data: Potential, Challenges, and Statistical Implications".https://www.imf.org/en/Publications/Staff-DiscussionNotes/Issues/2017/09/13/Big-Data-Potential-Challenges-and-Statistical-Implications-45106.