ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

# Survey on Privacy Preserving in Big Data using Machine Learning

Lokini Rajesh[1], Banalaxmi Brahma[2], Priyank Jain[3], Dr. Manasi Gyanchandani[4]

Research Scholar, Department of CSE, MANIT, Bhopal, M.P, India [1, 2, 3]

Assistant Professor, Department of CSE, MANIT, Bhopal, M.P, India [4]

**ABSTRACT:** Big Data deals with large volumes of data which cannot be processed in reasonable amount of time. This data needs to be processed quickly by various tools and techniques. Whereas in machine learning a system learns from past experiences and is able to build a model which would most likely be able to comprehend future instances. One of the main reasons why big data and machine learning are used together is because big data is more likely to be a pre-processing step to machine learning. The learning comes from extensive calculations done over existing datasets to create a learning model. Big data privacy on the other hand is about making the data secure from breaches. It basically deals with protecting the data about an individual or an institution in a dataset that is not to be made public. How machine learning and big data privacy works is a new field of interest. So, this paper gives a review of the existing big data privacy mechanisms using machine learning techniques.

**KEYWORDS**: Big Data, Machine learning, classification, clustering, supervised learning, unsupervised learning, differential privacy, privacy preserving

## I. INTRODUCTION

One of the biggest reasons why we use big data is to extract some meaning out of it, so that we can make better decisions. Machine learning is the science of training systems to learn from data and output appropriate response without being explicitly programmed for that .But, on flip side without big data machine learning would be totally irrelevant, because to learn anything from data you need to have a large number of 'training examples' so that all possible scenarios are exhausted and also to avoid faulty training due to few erroneous datasets. So, they are deeply interconnected. A typical framework can't deal with substantial dataset estimation and information size is expanding step by step, thus the obtained model should be adapted accordingly. To obtain this we have to implement big data technologies and feed output to machine learning algorithms for model/ learning generation.

## II. MACHINE LEARNING

Machine leaning is a field of research that formally focuses on the theory, performance, and properties of learning systems and algorithms. It is a highly interdisciplinary field building upon ideas from many different kinds of fields such as artificial intelligence, optimization theory, information theory, statistics, cognitive science, optimal control, and many other disciplines of science, engineering, and mathematics. ML addresses the question of how to build a computer system that improves automatically through experience. An ML problem is referred to as the problem of learning from experience with respect to some tasks and performance measures. ML techniques enable users to uncover underlying structure and make predictions from large datasets. ML thrives on efficient learning techniques (algorithms), rich and/or large data, and powerful computing environments. By and large, the field of machine learning is partitioned into three subdomains: supervised learning, learning, unsupervised learning, and reinforcement learning. Briefly, supervised learning requires training with labelled data which has inputs and desired outputs. In contrast with the supervised learning, unsupervised learning does not require labelled training data and the environment only provides inputs without desired targets. Reinforcement learning enables learning from feedback received through interactions with an external environment [1].

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

## III. BIG DATA

The term Big Data is characterized as "another era of advancements and structures, intended to economically extract value from very large volumes of a wide variety of data, by enabling high-velocity capture, discovery, and/or analysis''. Based on this definition, the properties of big data are reflected by 3 V's, which are, volume, velocity and variety. Volume refers to the amount of data generated. With the emergence of social networking sites, we have seen a dramatic increase in the size of the data. The rate at which new data are generated is often characterized as velocity. A typical topic of enormous information is that the information are differing, i.e., they may contain content, sound, picture, or video etc. This diversity of data is denoted by variety [2].

### 3.1 Big data privacy

The potential risk to privacy is one of the greatest downsides of big data. It should be taken into account that big data are all about gathering as many data as possible to extract knowledge from them (possibly in some innovative ways). Information protection is centered around the utilization and administration of individual information—things like setting up arrangements in place to ensure that consumers' personal information is being collected, shared and utilized in appropriate ways. Moreover, more than often, these data are not consciously supplied by the data subject (typically a consumer, citizen), but they are generated as a by-product of some transaction (e.g., browsing or purchasing items in an online store), or they are obtained by the service provider in return for some free service (for example, free email accounts, social networks) or as a natural requirement for some service (e.g., a GPS navigation system needs knowledge about the position of an individual to supply her with information on nearby traffic conditions) [2,3].
For a privacy model to be usable in a big data environment, it must cope well with volume, variety and velocity. To determine the suitability of a privacy model for big data, we look at the extent to which it satisfies the following three properties:
Composability. A privacy model is composable if the privacy guarantees of the model are preserved (possibly to a limited extent) after repeated independent application of the privacy model.
Computational cost. The computational cost measures the amount of work needed to transform the original data set into a data set that satisfies the requirements of the privacy model.
Likability. In big data, information about an individual is gathered from several independent sources. Hence, the ability to link records that belong to the same (or a similar) individual is central in big data creation [3].

### 3.2 Machine Learning in Big Data scenario

A few recent learning methods that may be either promising or much needed for solving the big data problems [1]:
**1.Representation Learning**: A promising solution to learn the meaningful and useful representations of the data that make it easier to extract useful information when building classifiers or other predictors, has been presented and achieved impressive performance on many dimensionality reduction tasks. Representation learning aims to achieve that a reasonably sized learned representation can capture a huge number of possible input configurations, which can greatly facilitate improvements in both computational efficiency and statistical efficiency.
**2. Deep learning**: Deep learning mainly uses supervised and/or unsupervised strategies in deep architectures to automatically learn hierarchical representations. As the data keeps getting bigger, deep learning is coming to play a pivotal role in providing predictive analytics solutions for large-scale data sets, particularly with the increased processing power and the advances in graphics processors.
**3. Distributed and parallel learning:** In the framework of distributed learning, the learning is carried out in a distributed manner. With the advantage of distributed computing for managing big volumes of data, distributed learning avoids the necessity of gathering data into a single workstation for central processing, saving time and energy. With the power of multicore processors and cloud computing platforms, parallel and distributed computing systems have recently become widely accessible.
**4. Transfer learning:** Transfer learning has been proposed to allow the domains, tasks, and distributions to be different, which can extract knowledge from one or more source tasks and apply the knowledge to a target task. The advantage of transfer learning is that it can intelligently apply knowledge learned previously to solve new problems faster.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

**5. Active learning:** Frequently, learning from massive amounts of unlabeled data is difficult and time-consuming. Active learning attempts to address this issue by selecting a subset of most critical instances for labelling. In this way, the active learner aims to achieve high accuracy using as few labelled instances as possible, thereby minimizing the cost of obtaining labelled data.

**6. Kernel-based learning:** It is a very powerful technique to increase the computational capability

Based on a breakthrough in the design of efficient nonlinear learning algorithms. The outstanding advantage of kernel methods is their elegant property of implicitly mapping samples from the original space into a potentially infinite-dimensional feature space, in which inner products can be calculated directly via a kernel function.

## IV. APPROACHES TO BIG DATA PRIVACY USING MACHINE LEARNING

Few of the approaches to big data privacy using machine learning is described below.

**1. Privacy-preserving machine learning**

The training data are distributed and each shared data portion is of large volume. Specifically, the data locality property of Apache Hadoop architecture is utilized and only a limited number of cryptographic operations at the Reduce () procedures to achieve privacy-preservation. The proposed scheme is secure in the semi-honest model and use extensive simulations to demonstrate its scalability and correctness [4].

The problem of collaborative machine learning over distributed training data is addressed and proposed to use the data locality property of big data processing framework such as MapReduce to achieve privacy preservation. Generally speaking, the collaborative learning problem is decomposed into subtasks such that each subtask is only related to one share of the training data. With this decomposition, local Mappers are able to work independently to get local training results which are then summarized by a secure protocol on Reducer. The proposed framework avoids secure operations on Mappers and only uses a limited number of cryptographic operations on Reducer to achieve privacy-preservation with an affordable computation overhead. Performance of the proposed scheme is studied via theoretical analysis and extensive experiments over three real-life data sets.

**2. Supervised machine learning**

It describes the setting when labels are known for training data, and the task is to train a model to predict accurate labels given a new example [5]:

***(i)Naive Bayes Model****:* The naive Bayes model is a classifier which predicts label Y according to features in X. Given features X and a model, one can compute the conditional probability P (Y |X) for all labels Y and predict the label with largest conditional probability. A differentially private naive Bayes model mechanism relies on one additional assumption: all values for all features in the dataset are bounded by some known number. If the bound covers most of the Gaussian distribution, then both the bound assumption and Gaussian assumption hold approximately. Therefore the sensitivity of the information that is needed to compute the model can be calculated. The mechanism then adds noise to this information according to the Laplacian mechanism and computes the model.

***(ii)Linear Regression***: Linear regression is a technique for predicting numerical values in which the value is modelled as a linear combination of features. It assumes bounded sample space and proposes a differentially private mechanism for linear regression. As the loss function is analytic, the mechanism expands the function with Taylor expansion, approximates it with a low order approximation, and adds noise to the coefficients of the terms. The mechanism then finds the w that minimizes the approximate loss function. As the sensitivities of the coefficients are easy to compute, the Laplacian mechanism can ensure the differential privacy of the noisy approximation. Since no private information is used after adding noise, the output vector w here is also differentially private. Furthermore, the model that is decided by w is also differentially private.

**3. Unsupervised Machine Learning**

Unsupervised learning depicts the setting when there are no marks related with each training example. In the absence of labels, unsupervised machine learning algorithms find structure in the dataset. In clustering, for example, seeks to find distinct groups to which each data point belongs. It can be useful in many contexts such as medical diagnosis, to know of an individual's member ship in a group which shares certain specific characteristics. However, releasing the high-

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

level information about a group may inadvertently leak information about the individuals in the dataset. Thusly it is critical to grow differentially private unsupervised machine learning algorithms.

K-means clustering:K-means is an ordinarily utilized model in bunching. To prepare the model, the calculation begins with k randomly selected points which represent the k groups, then iteratively clusters samples to the closest point and updates the focuses by the mean of the examples that are grouped to the points. [6] Proposes an (ϱ, δ)-differentially private k-means clustering algorithm using the sample and aggregate framework. The mechanism is based on the assumption that the data are well-separated. 'Well separated' means that the clusters can be estimated easily with a modest number of tests. This is an essential of the example and total structure. The mechanism randomly splits the training set into many subsets, runs the non-private k-means calculation on every subset to get many yields, and afterward utilizes the smooth affectability system to publish the output from a dense region differentially privately. This step preserves privacy while the basic k-means calculation is unaltered. Any adjustments on the k-means bunching algorithm (such as k-means++) can be used in the sample step, with the sole restriction that such adjustments leave in place the property that the calculation can be evaluated with a little number of samples. Additionally, if the sample space is bounded and the number of samples surpasses a limit, there is a bound on the clamor included. However this bound is not specifically related to the number of samples in the dataset.

## V. EXISTING MECHANISM AND SYSTEM ANALYSIS

### 1. Data Clustering Algorithm

K-means is a very useful tool of data clustering, it is first used in signal processing [7]. K-means aims to partition n records into k clusters in which each record belongs to the cluster with the nearest mean. Iterative update is used in the most common algorithm. The algorithm, k-means, is named because of its extensive existence. It is also referred to as Lloyd's algorithm, particularly in computer science community. There are two steps to implement this algorithm. Let Mean be the initial set of means and Mean = fm1;m2; :::;mkg. First, it is the assignment step. In this procedure, assign each record to the cluster whose mean yields the least within- cluster sum of squares. Since the sum of squares is the squared Euclidean distance, this is instinctively the "closest" mean. Scientifically, this implies apportioning the perceptions concurring to the Voronoi diagram generated by the means.

$$S_i^{(t)} = x_p : \| x_p - m_i^{(t)} \|^2 \leqslant \| x_p - m_j^{(t)} \|^2 \forall j, 1 \leqslant j \leqslant k$$

Then, it is the update step. We calculate the new means to be the barycenter of the records in the new groups.

$$m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \subset S_i^{(t)}} x_j$$

Since the number juggling mean is a minimum squares estimator, this likewise limits the inside bunch sum of squares (WCSS) objective. The algorithm has converged when the assignments no longer change. Since both strides enhance the WCSS objective, and there just exists a limited number of such partitioning, the algorithm must converge to an (local) optimum.

The k-means data clustering algorithm is introduced to our mechanism to assign records to the nearest group by range. It has the goal to minimize the WCSS objective, and it is assigned by least sum of squares. The algorithm is often presented as assigning objects to the nearest cluster by distance. The standard algorithm aims at minimizing the WCSS objective, and thus assigns by \least sum of square", which is exactly equivalent to assigning by the smallest Euclidean distance.

The time complexity is satisfactory. If k and d are constant values, then the time complexity can be presented as:

O (ndk+1 log n);

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

Where n=jTj denotes the number of records while d denotes the dimension.

**2. K-means clustering algorithm**

Initialization:
Let n be the number of clusters
Let S be the set of feature vectors (jSj is the size of the set)
Let A be the set of associated clusters for each feature vector
Let sim(x,y) be the similarity function
Let c[n] be the vectors for our clusters
Let S0 = S
fori = 1; i 6 n; i + + do
j = rand(jS0j)
c[n] = S0[j]
S0 = S0 □ fc[n]g
end for
fori = 1; i 6 jSj; i + + do
A[i] = argmax(j = 1ton)fsim(S[i]; c[j]g)
end for
Let change = true
while change do
change = false
fori = 1; i 6 jSj; i + + do
a = argmax(j = 1tonfsim(S[i]; c[j])g
if a! = A[i] then
A[i] = a
change = true
end if
end for
if change then
fori = 1; i 6 n; i + + do
mean; count = 0
for j = 1; j 6 jSj; j + + do
if A[j] == i then
mean = mean + S[j]
count = count + 1
end if
end for
c[i] = mean=count
end for
end if
end while

**3. Shuffle Algorithm**

The Fisher-Yates shuffle is a very popular algorithm which is an in-place shuffle [8]. That means instead of creating a new shuffled copy of the records, it shuffles the records of a table in place. If the table to be shuffled is large enough, this mechanism can fit well.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

In order to initialize and shuffle a table synchronously, an advanced version is introduced to our mechanism to make it more efficient. The random algorithm can perfectly put a certain record i into a random location among the first in locations in the table, after moving the record previously taking up that location to location i. In normal conditions, which the records are meant to be shuffled by a column of number, especially the integers, this could be easy to represent by a function because the implementation will not change it.
There is another advantage of the random algorithm. It can be modified to adapt to certain situation that n=jTj remains unknown. The concrete solution is as Algorithm 2.

**4. Random shuffle algorithm**

```
Initialization:
vector stack1, stack2
while amount of shuffles do
fori = 1; i 6 n; i + + do
if rand()&1 then
stack1:pushback(array[i])
else
stack2:pushback(array[i])
end if
end for
i = 0
while stack1:empty()&&stack2:empty() == false do
if stack1:empty() == true then
array[i] = stack2:at(0)
stack2:erase(0)
i + +; continue
else if stack2:empty() == true then
array[i] = stack1:at(0)
stack1:erase(0)
i + +; continue
end if
if rand()&&1 then
array[i] = stack2:at(0)
stack2:erase(0)
i + +; continue
else
array[i] = stack1:at(0)
stack1:erase(0)
i + +; continue
end if
end while
end while
```

## VI. KEY OPPORTUNITIES AND CHALLENGES

In addition to detailed discussion of opportunities and challenges that big data present to ML throughout this section, here we highlight a few key opportunities and challenges.
ML on big data requires a new way of thinking and novel algorithms to address many technical challenges [9]. Big data is one of the key enablers of deep learning, which has improved the state-of-the-art performance in various applications. Deep learning can typically recognize at least 1000 different categories, which is at least 2 orders of magnitude higher than run of the mill number of classifications dealt with by a conventional neural system [10]. Also,

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

enormous data enables learning at multi-granularity. Furthermore, big data provides opportunities to make causality inference based on chains of sequence, to enable effective decision support.

ML on big data also highlights the importance of privacy-preserving ML. Big data may be highly personal [11]. For instance, healthcare data may be collected from multiple organizations that have different privacy policies, and may not explicitly share their data publicly. Since joint ML may sometimes become necessary, how to share big data among distributed ML entities while moderating security concerns turns into a testing issue. For example, protection conservation ML has been achieved by employing the data locality property of Hadoop architecture and only a limited number of cryptographic operations at the Reduce steps are needed [12]. A privacy-preserving solution for SVM classification has also been proposed [13]. Big data enhances the real-world impact of ML. The applications in which ML has created real world impact range from science (e.g., physical design, biological science, earthquake prediction) to business (e.g., financial systems, post-approval drug monitoring, self-driving cars), and from public platforms (e.g., social media) to organizational realms (e.g., intrusion networks, healthcare systems).

Among the existing challenges for ML on big data, one key issue is to improve the efficiency for iterations. Existing parallel frameworks are not particularly designed for ML algorithms. Usually big data tools perform computation in batch-mode and are not optimized for tasks with iterative processing and high data dependency among operations (e.g., due to heavy disk I/O). Iterative subtasks (i.e., processing steps which are executed repetitively until a convergence condition is met) dominate both categories of algorithms. Optimizing cluster resource allocations among multiple workloads of iterative algorithms often require an estimation of their runtime, which in turn requires: (a) predicting the number of iterations and (b) predicting the processing time of each iteration [14]. The Hadoop infrastructure can both avoid extremely slow, or straggler tasks and handle them at runtime (through speculative execution). Spark [15] supports not only MapReduce and fault tolerance but also cache data in memory between iterations. On a related note, methods have been developed to improve computational efficiency on big data without sacrificing ML performance, which hold only small pieces of the data rather than all data in fast memory, and build a predictor on each small piece and then combine these predictors together [16]. In addition, graph-based architectures and in-memory big data tools have been developed to minimize the I/O cost and optimize iterative processing [17].

Second challenge is to minimize the feedback/communication from/with classifiers. The problem of learning the optimal classifier chain at run-time has been modeled as a multi-player multi-armed bandit problem with limited feedback [18]. It does not require distributed local classifiers to exchange any information except limited feedback on mining performance to enable the learning of an optimal classifier chain [19].

A third challenge is to address the velocity aspect of big data in ML. Current (de-facto standard) solutions for big data analysis are not designed to deal with evolving streams [100]. A ML system must be able to cope with the influx of changing data in a continual manner. Lifelong Machine Learning is in contrast with the traditional one-shot learning [20]. To this end, online learning has been exploited to make kernel methods efficient and scalable for large-scale learning applications. For instance, two different online kernel ML algorithms – Fourier Online Gradient Descent and Nystrom Online Gradient Descent algorithms have been explored to tackle three online learning tasks: binary classification, multi-class classification, and regression [21, 22].

A fourth challenge is to address the variety aspect of big data in ML. Most traditional ML algorithms can only take certain type of input, such as numerical, text or images. In many cases, data that could be used for a single ML goal may come in different types and formats. It will result in an explosion of features to be learned and is sometimes referred as a “Big Dimensionality” challenge [23]. For instance, one ML algorithm might need learn from: (1) a mixture of large volume of data and high speed stream data, or (2) large volume of data with image, text, acoustic, and motion features.

A fifth challenge is increased problem complexity (e.g., in multiclass classification and new classes). In document and image clustering/ classification, in addition to a large number of data points and their high dimensionality, the number

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

of clusters/classes is also large. Therefore, there is a need to gradually expand the capacity of ML models to predict an increasingly large number of new classes.

ML on big data presents numerous other challenges. For instance, optimization in conventional ML focuses on average performance, but hard to prevent poor outcomes.
Most traditional ML algorithms are not designed for data that are not loaded into memory completely. Additionally, it is complex to set objective functions due to the large number of component terms and various trade-offs among performance measures. Noise is a bigger issue for big data because patterns typically reside in a small subset of data (e.g., spam and online attacks).

## VII. CONCLUSION

This paper presents an overview of opportunities and challenges of ML on big data. Big data creates numerous challenges for traditional ML in terms of scalability, adaptability, and usability, and presents new opportunities for inspiring transformative and novel ML solutions to address many associated technical challenges and create real-world impacts. These opportunities and challenges serve as promising directions for future research in this area.
Most existing work on ML for big data focused on the volume, velocity and variety aspects, but there has not been much work addressing the remaining two aspects of big data: veracity and value. To handle data veracity, one promising direction is to develop algorithms that are capable of accessing the trustworthiness or credibility of data or data sources so that untrustworthy data can be filtered during data pre-processing; and another direction is to develop new ML models that can inference with unreliable or even contradicting data. To realize the value of big data in decision support, we need to help users understand ML results and the rationale behind each system's decision. Thus, explainable ML will be an important future research area. Moreover, to support human-in-the-loop big data ML, we need to address fundamental research questions such as how to effectively acquire large amount of annotated data through crowd sourcing; how to evaluate an ML algorithm based not only on its prediction accuracy or scalability, but also on its overall capability to support end users in performing their tasks (e.g., usability-based measure). Further, additional open research issues include: (1) how to protect data privacy while performing ML; (2) how to make ML more declarative so that it is easier for non-experts to specify and interact with; (3) how to incorporate general domain knowledge into ML; and (4) how to design new big data ML architecture that seamlessly provides decision support based on real-time analysis of large amount of heterogeneous data that may not be reliable.
When private data is discussed, medical data is typically offered as an example application. However, medical datasets are often not relational. They may be temporal, and sometimes structural. Although we can transform such data, the transformation may lose some important information and increase sensitivity. Along these lines, systems uncommonly intended for such information are required. Another important question is whether privacy can be free, i.e., achieved at no cost to utility in differentially private learning. For privacy to be free, the noise required to preserve privacy might need to be smaller than noise from sample randomness. In that case, it wouldn't change the magnitude of noise to take privacy into account. For example, proves that (ϙ, δ)-differential privacy is free for learning models satisfying a certain set of conditions. The mechanism in ensures free ϙ-differential privacy for regularized logistic regression models and linear SVM models, where noise from sample randomness is $O(1/\sqrt{n})$ and the noise to preserve privacy is $O(1/n)$. The mechanism in also proves that the effect of noise brought by differential privacy is O (1/n), while the effect from sample randomness is $O(1/\sqrt{n})$. and also consider the extent to which privacy is compatible with and related to the idea of generalization in machine learning. Intuitively, machine learning algorithms seek to generalize patterns gleaned from a training set avoiding the effects of sample randomness. Ideally, these algorithms should be robust to small changes in the empirical distribution of training data. A model which fits too heavily to individual examples loses generalizability and is said to over fit. Perhaps the goals of differential privacy and generalization are compatible.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

*Website: www.ijircce.com*

Vol. 5, Issue 4, April 2017

## REFERENCES

[1] JunfeiQiu, Qihui Wu, Guoru Ding, Yuhua Xu and Shuo Feng, A survey of machine learning for big data processing, EURASIP Journal on Advances in Signal Processing (2016) 2016:67 DOI 10.1186/s13634-016-0355-x.
[2] Abidmehmood, Iynkarannatgunanathan, Yong xiang, Guanghua and Song guo, "Protection of Big Data Privacy", 2016, IEEE.
[3]Jordi Soria-Comas, Josep Domingo-Ferrer, Big Data Privacy: Challenges to Privacy Principles and Models, Data Sci. Eng. (2016) 1(1):21–28 DOI 10.1007/s41019-015-0001-x.
[4] Kaihe Xu et al, Privacy-preserving Machine Learning Algorithms for Big Data Systems, 2015 IEEE 35th International Conference on Distributed Computing Systems.
[5] Zhanglong Ji, Zachary C. Lipton, Charles Elkan, Differential Privacy and Machine Learning: a Survey and Review.
[6] KobbiNissim, SofyaRaskhodnikova, and Adam Smith. Smooth sensitivity and sampling in private data analysis. In ACM SIGACT-SIGMOD- SIGART Symposium on Principles of Database Systems, pages 75–84, 2007.
[7] M.M. Ettefagh and Mech. Eng. Dept. Bearing fault diagnosis using hybrid genetic algorithm k-means clustering. In Innovations in Intelligent Systems and Applications (INISTA) Proceedings, pages 84{89. IEEE, 2014.
[8] PE Black. Dictionary of algorithms and data structures. 2005.
[9] X.-w. Chen, X. Lin, Big data deep learning: challenges and perspectives, Access, IEEE 2 (2014) 514–525.
[10] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, 2016.
[11] K. Xu, H. Yue, L. Guo, Y. Guo, Y. Fang, Privacy-preserving machine learning algorithms for big data systems, in: Proceedings of the 2015 IEEE 35th International Conference on Distributed Computing Systems (ICDCS), 2015, pp. 318–327.
[12] J. Vaidya, H. Yu, X. Jiang, Privacy-preserving SVM classification, Knowledge Inf. Syst. 14 (2008) 161–178.
[13] A.D. Popescu, A. Balmin, V. Ercegovac, A. Ailamaki, PREDIcT: towards predicting the runtime of large scale iterative analytics, Proc. VLDB Endow. 6 (2013) 1678–1689.
[14] M.Zaharia, M.Chowdhury, M.J.Franklin, S.Shenker, I.Stoica, Spark: cluster computing with working sets, presented at in: Proceedings of the 2nd USENIX conference on Hot topics in Cloud Computing, Boston, MA, 2010.
[15] L. Breiman, Pasting small votes for classification in large databases and On-Line, Machine Learn. 36 (1999) 85–103.
[16] H. Kashyap, H.A. Ahmed, N. Hoque, S. Roy, D.K. Bhattacharyya, Big Data Anal. Bioinforma.: A Mach. Learn. Perspect. (2015).
[17] J.Xu, C.Tekin, M.van der Schaar, Learning optimal classifier chains for real-time big data mining, in Proceedings 51st Annu. Allerton Conference Comm., Control and Comput. (Allerton'13), 2013.
[18] G.DeFrancisci Morales, SAMOA: a platform for mining big data streams, in: Proceedings of the 22nd International Conference on World Wide Web, 2013, pp. 777–778.
[19] Q.Yang, Big data, lifelong machine learning and international conference on Web search and data mining, 2013, pp. 505–506.
[20] J. Lu, S.C. Hoi, J. Wang, P. Zhao, Z.-Y. Liu, Large scale online kernel learning, J. Mach. Learn. Res. 17 (2016) 1–43.
[21] Z. Wang, K. Crammer, S. Vucetic, Breaking the curse of kernelization: budgeted stochastic gradient descent for large-scale SVM training, The J. Mach. Learn. Res. 13 (2012) 3103–3131.
[22] Y. Zhai, Y.S. Ong, I.W. Tsang, The emerging big dimensionality, IEEE Comput. Intell. Mag. 9 (2014) 14–26.
[23] T.Xiao, J.Zhang, K.Yang, Y.Peng, Z.Zhang, Error-Driven Incremental Learning in Deep Convolutional Neural Network for Large-Scale Image Classification, in: Proceedings of the ACM International Conference on Multimedia, 2014, pp. 177– 186.

## BIOGRAPHY

**Lokini Rajesh is** completed B.Tech from NIT Silchar in 2014, presently pursuing M.Tech Research Scholar in MANIT Bhopal and **Area of Interest**: Area of research interest is Big Data analytics, Privacy and Security, Big Data privacy, Data Mining, Hadoop, High Performance computing, Internet of things. I am currently doing experimental work in Privacy and Security Concerns on Big data.

**Banalaxmi Brahma**is a research scholar in MANIT, Bhopal. She is pursuing M.Tech in Computer Networking from CSEdepartment. Her areas of interest are Computer Networking, Big Data Privacyand Internet of Things. She received her B.Tech degree from Tezpur University,Assam in Computer Science and Engineering.

**Priyank** Jain is working as a PhD Research Scholar. He is having 8 years' Experience as an Assistant professor & in research field. Mr. Priyank Jain has experience From Indian Institute of Management Ahmedabad India (IIM A) in research field. His Educational Qualification is M.Tech & BE in Information Technology. Mr. Priyank Jain's areas of specialization are Big data, Big Data Privacy & Security, data mining, Privacy Preserving, & Information Retrieval. Mr. Priyank Jain has publications in various International Conference, International Journal & National Conference. He is a member of HIMSS.

**Mansi Gyanchandani** working as Assistant Professor in MANIT Bhopal. She is having 20 years' experience, Her Educational Qualification is PhD in Computer Science & Engineering. Dr. Manasi Gyanchandani area of Specialization in Big data, Big Data Privacy & Security, data mining, Privacy Preserving, Artificial Intelligence, Expert System, Neural Networks, Intrusion Detection & Information Retrieval. Dr. Manasi Gyanchandani, publications in 04 International Conference, 04 International Journal & 04 National Conference. She is Life member of ISTE.