Privacy vs Utility analysis when applying Differential Privacy on Machine Learning Classifiers

Mathuranthagaa Selvarathnam
Department of Engineering
School of Science & Technology
City, University of London
London, United Kingdom
mathuranthagaa.selvarathnam@city.ac.uk

Roshan Ragel
Department of Computer Engineering
University of Peradeniya
Peradeniya, Sri Lanka
roshanr@eng.pdn.ac.lk

Constantino Carlos Reyes-Aldasoro
Department of Computer Science
School of Science & Technology
City, University of London
London, United Kingdom
constantino-carlos.reyes-aldasoro@city.ac.uk

Muttukrishnan Rajarajan
Department of Engineering
School of Science & Technology
City, University of London
London, United Kingdom
r.muttukrishnan@city.ac.uk

***Abstract*—In this paper we present how Differential Privacy (DP), which is the state of the art in privacy preserving technologies in recent times, plays a role with different ML classifiers. Preserving privacy while serving the utility needs it is supposed to do is a challenge for each ML implementation. In order to study the effects of different DP implementations , we perturb data at different phases of the ML cycle such as perturbing data at its origin (Differential Privacy Method 1 - DPM1), during the training process (DPM2) or the parameters of the ML model generated (DPM3) by adding noise at these phases and see the effect of privacy preservation on ML model utility. Further we have tested with different perturbation methods such as Laplace mechanism, Gaussian mechanism, Analytic Gaussian mechanism, Snapping mechanism and Staircase mechanism for DPM1 and analysed the results to know which one works better for numerical data. We have used privacy attacks such as Membership Inference Attack (MIA) and Attribute Inference Attack (AIA) to see the effect of DP in protecting the privacy of data. Based on the results of our experiment we found perturbing at later stages of an ML method provides better utility and when considering different DPM1 mechanisms, improved versions of Laplace and Gaussian works better in terms of utility while preserving privacy.**

***Index Terms*—Differential Privacy, Machine Learning, Privacy Vs Utility**

I. Introduction

Machine Learning (ML) is heavily used these days to make use of vast amounts of data available by advances in technology such as processing, storage and networking. ML uses data and algorithms to deliver a decision process. It involves supervised or semi-supervised or unsupervised learning mechanisms. In supervised learning, labelled datasets are used. Classification, regression techniques come under this category. In addition to serving their purposes such as classifying correctly or predicting most closely, ML methods need to safeguard themselves and the data used to train them from different security and privacy attacks [1] . Under this umbrella, this paper covers preserving privacy of classification algorithms namely Naïve Bayes, Random Forest, Logistic Regression and Neural Networks.

ML can be used for different purposes in general and [2] considers ML in three different views such as ML as a target for privacy attacks, ML as a privacy tool and ML as an attack tool with appropriate attack and threat models for all these three views. The first category covers our focus area emphasizing the importance of having privacy preserving mechanisms along with ML. The privacy preservation methods of using cryptography methods such as Fully Homomorphic Encryption (FHE) [3] and Secure Multi-party Computation (SMC) [4] along with ML are mentioned with the drawback of them having high communication overhead. This survey states that the Differential Privacy (DP) methods are widely used with Deep Learning (DL) methods and obfuscation with data is not extensively studied, which makes sense of our work. Aggregation methods are mentioned for collaborative learning scenarios (In the sense of aggregating from multiple parties) such as Federated Learning (FL) [5] where DP and encryption methods could be combined to overcome bottlenecks such as communication cost. Another survey on Big Data Privacy Protection [6] first explains the importance of Data Mining (DM) and among four major stages of DM process (Data Preprocessing, Data transformation, DM and Evaluation and knowledge representation), considers Preprocessing and DM task stages for privacy protection. Privacy preservation techniques and DM tasks such as Association rule mining, clustering, and classification are explained but it lacks in mapping between DM tasks and privacy preservation methods.

Existing literature on privacy preservation focuses on how to get the use of data without accessing it in its original form [7]. When the original data is modified (LDP), Knowledge discovery from individual data is restrained and results in greater utility loss. But still ML methods such as clustering and classification which use data distribution could make use

of LDP. [8]. Reference [8] have added that finding solutions for the privacy - utility trade-off is required for personalized privacy. Reference [9] mentions that DP has become the standard way of privacy preservation but how to choose the privacy budget is a remaining question. Reference [10] mentions the issue of utility loss when using DP and they decide on a threshold for classification error for their ensemble classifier and modify the DP noise added to dataset, using Laplace Mechanism until the threshold is not violated. They have provided experimental results that even when they change the number of weak decision tree learners in their ML method with the expectation to improve utility, still the results for perturbed dataset remains the same. Reference [11] has tried an implementation of privacy preserving K- nearest neighbour not with DP but using bloom filters on dataset attributes and have studied the effect of using different bit size bloom filters, using Jaccard similarity measure instead of standard Euclidean and bloom filter for individual attribute vs whole record. They have used recall, precision and F-score on prediction accuracy as metrics but no clear measure on privacy is mentioned. Three different ways of handling utility vs privacy is mentioned in literature as given in [12]. Predetermining the expected utility via measures such as accuracy and finding optimal privacy level for it, in other way, deciding on privacy leakage first and tuning learning parameters to maximize utility and the third way of adding feature based noise according to the relevance of that feature to the output.

Our work closely follows the work of [9] and [13], where both have used MIA and AIA to measure privacy leaks. MIA is more relevant here as it says if a record is used in ML training or not and hence serves as a way to measure privacy preservation. Reference [9] have considered relaxed definitions of DP such as Advanced composition, Zero-concentratedDP and RenyiDP against naïve composition on ML methods: Neural Network and Logistic Regression and have shown how increasing privacy budget for greater utilization opens the door for privacy attacks with relaxed versions of DP. Reference [13] has considered DPM1, but only using Laplace mechanism, DPM2 and DPM3 on different datasets including synthetic data. They have produced results with varying privacy budgets and suggest using their work to select ML methods based on data complexity and privacy utility needs. We have expanded their work in DPM1 mechanisms and ML methods considered.

We found that considering different ML methods with different datasets and evaluating Utility Vs Privacy when using DP would be an appropriate area to explore as it will provide an idea on which mechanism works better with which ML method and how the utility is affected by perturbation phases for a ML method.

This paper is organized as follows: Section II provides preliminaries used in this paper such as DP, privacy attacks. Section III gives details about the dataset, which type of DP perturbation is done with different ML methods and the metrics used. Section IV discusses the results we have obtained. Section V concludes the paper and briefs future work we aim to work on.

II. Preliminaries or Background

A. Differential Privacy

DP is a technology to provide useful information from a dataset without revealing any information of individuals in that dataset. If you consider a dataset with a particular record and another without that record and otherwise similar, when you perform a query on these two databases, DP confirms that the probability of producing the results will be the same (or nearly same) for both databases. DP assists in learning about the population but nothing about an individual [14]. A function which satisfies DP is commonly known as a DP mechanism in this paper.

Definition 1. (ϵ, δ) privacy - Given a Mechanism *M* with domain $N^{|x|}$ and possible range *S*, (ϵ, δ) DP is given [14] for neighbouring datasets x,y by

$$Pr[M(x) \in S] \leq e^{\epsilon} \cdot Pr[M(y) \in S] + \delta \tag{1}$$

Here ϵ is called a privacy budget and needs to be as small as possible for stronger privacy. Parameter δ represents failure probability which means that the mechanism is ϵ-differential private with $1 - \delta$ probability. Hence δ needs to be as low as possible like in the order of inverse of dataset size. When $\delta = 0$ the mechanism is called as ϵ-differential private. The Laplace mechanism is an example for ϵ-DP and Gaussian is for (ϵ, δ) DP. These mechanisms are explained under II-A0b. DP has some properties listed below.

- Composition properties
  Composition properties hold when you combine the privacy cost for the same dataset with more than one DP mechanism applied on them.
  - Sequential composition - If you apply a DP mechanism *M1* with ϵ_1 and another *M2* with ϵ_2 on same data, then the total results satisfies $\epsilon_1 + \epsilon_2$ DP.
  - Parallel composition - If a DP mechanism *M* satisfies ϵ DP and if you split the data into non overlapping chunks and apply same *M* mechanism on these disjoint data, the total results will satisfy ϵ DP.
- Post processing property - If a function satisfies ϵ DP, then any other function applied on top it also provides minimum ϵ DP. Hence it's safe to operate on a DP output

These properties are useful when applying DP in ML methods. For example, for DL methods where there are iterations in training, accounting them in privacy cost is vital.In random forest you could split data into disjoint data and use parallel composition or use same data and use sequential composition [15].

a) Global Differential Privacy (GDP) and Local Differential Privacy (LDP):

DP could be applied globally or locally and the most commonly studied and implemented version of DP [16] is GDP where noise is added while providing the query results. This method is called Central Differential Privacy (CDP) as well and a trusted aggregator or curator is employed in this case. When perturbation is done on data itself before sending it to curator or aggregator, it provides better privacy and is called

LDP. Though LDP is better in privacy, doesn't need a third party aggregator, GDP provides better accuracy in general. Since each user adds noise in LDP, it in turn would produce a larger sum of noise at the end and turns accuracy down.

b) DP mechanisms:
Laplace and Gaussian mechanisms are mainly used in literature and there are variants available as needs and issues with these base implementations arise. We obtain DP applied version F(x) by applying a DP mechanism on a function f(x) and for Laplace mechanism it is given by

$$F(x) = f(x) + Lap\left(\frac{s}{\epsilon}\right) \tag{2}$$

Here *s* is the sensitivity of the function *f*, which is obtained by calculating the output change for a unit change in input. In a simple example of a sum query, when a single row changes, the output changes by 1 and hence the sensitivity will be 1.
In Gaussian mechanism, Gaussian noise could be added using

$$F(x) = f(x) + N\left(\sigma^2\right) \tag{3}$$

where

$$\sigma^2 = \frac{2s^2 log\left(\frac{1.25}{\delta}\right)}{\epsilon^2} \tag{4}$$

Here *s* is the sensitivity of the function *f*.

Both Laplace and Gaussian could be extended to be used on vector valued functions and sensitivity values should be modified accordingly. L1 or L2 norm could be used as sensitivity value for vector valued function. The L1 norm is the sum of the absolute differences of two vectors and known as the Manhattan distance as well. The L2 norm is calculated by taking the square root of the sum of the squared differences of two vectors and known as the Euclidean distance as well. Though the noise spectrum of Laplace is narrow comparing to Gaussian, Gaussian works better with application which has lower L2 sensitivity than L1 as it allows L1, L2 norms to be used for sensitivity where Laplace only allows L1 norm [14]. Further there are improvements suggested for both these Laplace and Gaussian Mechanisms and we have tested three of those versions in this paper as listed below.

- Analytic Gaussian mechanism is proposed in [17] as an improvement to tighter the privacy bounds of Gaussian mechanism and much helpful for high dimensional dataset.
- Snapping mechanism is proposed as a defence mechanism to the security risk possessed by floating point implementation of Laplace Mechanism [18].
- Staircase mechanism is suggested in [19] to optimize the utility of Laplace mechanism by using a staircase shaped probability density function for noise addition.

c) DP mechanisms and data types: Laplace mechanism, Gaussian mechanism are proposed for numerical data and exponential for categorical data. Applying DP on unstructured data, set-valued data and image data are studied in literature [20]. There are also mechanisms available for binary data (Binary mechanism), discrete values (Geometric mechanism) and convex optimization functions (Vector mechanism) [21]. We have considered numerical data in this paper.

d) Stages of DP noise addition in ML methods: DP noise could be added at different stages of a ML method. Data could be perturbed at its origin (DPM1) and it gives more privacy as there is no reliance on any centralized authority or third party. Further perturbation could be done during the training process (DPM2) or later after the training is completed, on the model parameters it produced (DPM3). For DL, DPM2 can be further considered as adding noise to gradients and objective function. Here we have considered perturbing gradients only. Composition properties of DP need to be considered as DL involves multiple iterations and different tighter bounds on privacy are available such as zero-concentrated DP [22], Moments account [23] and Rényi DP [24]. We have used naïve composition with a Gaussian mechanism which has looser bounds.

B. Privacy attacks

Measuring privacy has been often application specific and it brings in a challenge to provide a common measure [8]. There are ways to measure privacy using Information Loss and Mutual Information. Information loss is defined as "Lack of precision in estimating the original dataset is known as information loss which can lead to the failure of the purpose of data mining" by [25]. Mutual Information privacy provides a notion of how much one random variable is related to another [26]. But recently privacy attacks are used widely [27] and we follow the same. Membership Inference Attack (MIA) and Attribute Inference Attacks (AIA) are two attacks we have used and details of them are given below.

a) MIA: MIA is performed with the purpose of knowing if a given data record is used in a model's training or not. There are methods proposed in literature to serve this purpose. Knowing whether a record was used in training leads to information leakage and privacy violation. It also serves as a good measure to know how effective a DP mechanism is on a dataset. Reference [7] assumes a black-box access to the target ML model and creates several shadow models to train attack models per each possible prediction class. Shadow models are needed to imitate the target model and the attack model is a binary classifier deciding if a particular record is used in the target model's training or not. Proposed method here is generic to datasets used and ML model, i.e. it doesn't depend on these factors.

Following above MIA, [27] comes up with three different adversary models and relaxes several assumptions made by previous work. Adversary one uses a single shadow model and an attack model and still shows that they achieve similar outcome as [7]. Using several shadow models increases cost for construction and operation of attack. In adversary two, they show that even if the shadow model is trained using a different model and even if the data used for this training is of different distribution, still MIA could be successful. It is called as 'data transferring attack' and has the benefits of being able to be applied on dataset of any kind rather than dataset with binary features which was the case for [7] and removing the number

of queries required to generate a synthetic dataset against a black-box ML service provider as in [7].

Adversary three only uses the posterior probabilities provided by the target model and does not use any shadow models. It performs model and data independent membership inference attacks. Whether the maximum posterior is higher than a threshold is the deciding factor here to predict against members and non-members and they assume ML model is more confident on a record that it has seen already than a new one. They have shown this works as expected yet being a simple method. This is the method used in this paper to study utility against privacy attacks when applying DP techniques. Threshold could be chosen based on the need. Authors have proposed a way of selection by randomly generating data, using those as non member queries and using the maximum posteriors from them as threshold. Area under the Receiver Operating Characteristic (ROC) curve (AUC) is used to serve the same purpose without setting a threshold.

b) AIA: AIA is performed to know the exact value of an attribute in a record. Reference [13] has used [28] and we do the same. For a record, all values except that particular attribute is assumed to be known and for that attribute, all possible values it could take is considered to find the result. With all possible values that an attribute could take, loss of the model is calculated and the attribute value which gives value close to training loss is considered as its original value.

c) Defence mechanisms for privacy attacks: Overfitting is identified as the main measure or one of the main reasons supporting MIA. Other than it, model structure and type also contribute to leaks as some try to remember information. Further DP mechanisms could be applied to reduce leakage [7]. Overfitting could be measured using the difference between the training accuracy and test accuracy. Ways preventing overfitting are listed below

- Regularization - Refers to methods which prevent overfitting by introducing some penalty values to the loss function.
- Dropout - It could be used with Neural Networks by dropping a portion of layers in input or hidden layers.
- Model stacking - Since dropout works for Neural Networks, for models other than it, model stacking can be used. Here parts of data are used to train multiple ML models and combination of them is used as target model to prevent overfitting.

III. Methodology

A. Datasets considered

We have used the two dataset used in [13] and as provided in [29]. The Netflix [30] dataset contains movie ratings from 1 to 5 and here they have considered the top 1000 rated movies and users who have rated them. Similar user groups are identified and put in the same classes using the k-means clustering algorithm. Class values as shown in table I are chosen by changing the value of k here. Prediction task here is if the users labelled as belonging to the same group are predicted correctly or not.

Purchase dataset [31] contains the user details with binary features - if they have bought or not any products among 599 products. Here similar to Netflix data, users with similar purchase behaviour are classed together using k-means clustering algorithm.

TABLE I
DATASETS USED IN EXPERIMENTS

Dataset name	Rows	Attributes	Classes
Netflix	100000	1000	2,10,20,50,100
Purchase	200000	599	10,20

B. ML methods with their perturbation stages

In this paper we consider ML methods with given stages of perturbation as given in table II.

For DPM1, noise was directly added to the dataset, as a kind of pre-processing step and ML models were trained and tested on these perturbed data. Mechanisms mentioned in Section II-A0b were considered for this purpose.

Except for the Neural Network method, IBM DP library [21] is used in either DPM2 or DPM3 with other ML methods.

1) Logistic Regression: Logistic regression is used from the $Scikit-learn$ library and DPM3 is based on [32]. DPM3 implements an objective function perturbed classifier with Laplace mechanism.

2) Naïve Bayes: Basic implementation is used from the $Scikit-learn$ library. DPM3 version is based on the work in [33] where they have provided ways to calculate sensitivity values for numerical and categorical attributes and add Laplacian noise for mean and standard deviation in case of neumerical attribute or to the counts in case of categorical attribute.

3) Neural Network: Neural Network model structure of input size * 256 * 256 * output classes are used with kernel regularizer of ratio 0.0001 using keras library. For the DP version, tensorflow_privacy [34] is used with DPM2 which has the implementation of Differentially Private Stochastic Gradient Descent (DP-SGD) algorithm from [23].

4) Random forest: Non-private implementation is used from the $Scikit-learn$ library. A custom implementation is used by authors of [13] based on the work [15] for DPM2. But because of the addendum published by the author of [15] in August 2021, we skipped using it and have used IBM DP library [21] as it has a fix covering the addendum, in its latest version. [15] has implemented it using a low sensitivity query to select the frequent label using an exponential mechanism with high probability and hence reducing the privacy budget. Further, they have used parallel composition on disjoint data while building the random forest.

C. Privacy attacks

Here an equal number of train and test sets are used for privacy attacks to maximize the uncertainty of the model predictions and to keep the baseline accuracy for this case as 0.5 as explained in [9]. A training set member is considered as

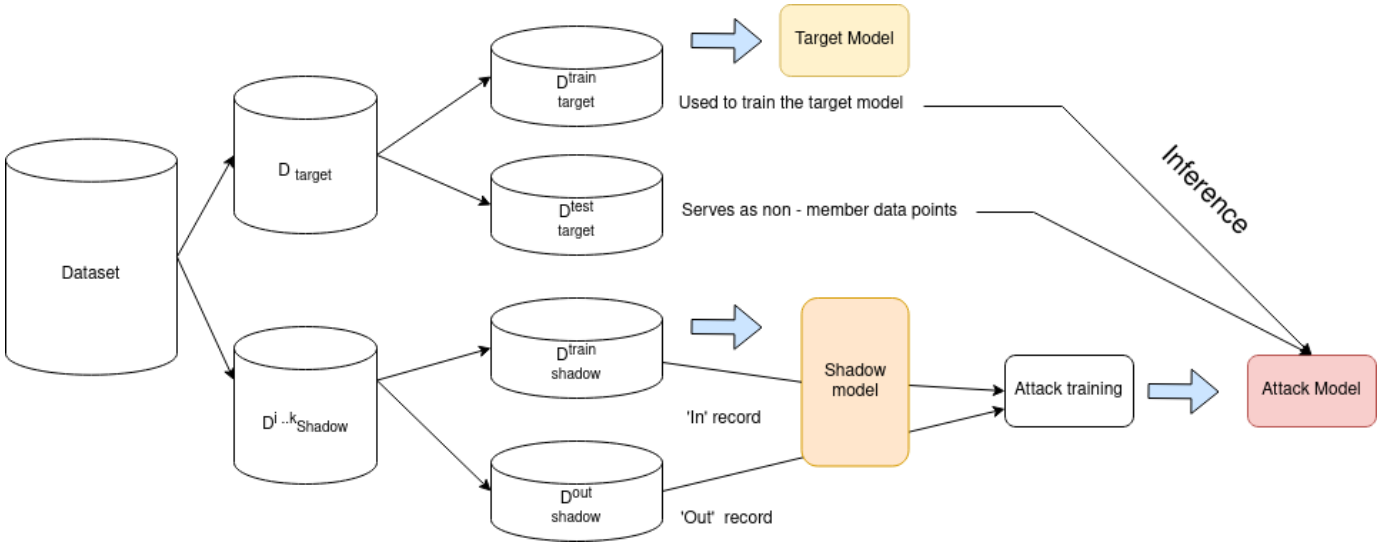

Fig. 1. A sample MIA privacy attack model is depicted in this figure, which is useful to measure the impact DP has on a ML model.

TABLE II
ML METHODS AND PERTURBATION PLACES

ML method	Perturbation places		
	DPM1	DPM2	DPM3
Logistic Regression	√		√
Naïve Bayes	√		√
Neural Network	√	√	
Random forest	√	√	

TABLE III
CONFUSION MATRIX FOR A TWO CLASS CLASSIFICATION PROBLEM

		Predicted class		
		Positive	Negative	Total
Actual class	Positive	True Positives (TP)	False Negatives (FN)	TP + FN
	Negative	False Positives (FP)	True Negatives (TN)	FP+TN
	Total	TP+FP	FN+TN	All

a member for the attack and a test set member is considered as non member for both MIA and AIA. In MIA, TPR (equation 9) and FPR (equation 10) are calculated using the ROC curve passing the predicted values and membership details. Membership advantage is calculated using the difference between both TPR and FPR. In AIA, a number of randomly selected attributes are tested for attribute membership. Difference between predicting member correctly (TPR) vs non-member incorrectly (FPR) is taken to calculate the advantage here. Average of this value among n attributes is considered in graphs.

D. Metrics used

a) Measuring utility: ML models are used to decide on unseen data using the training it had on the available training dataset. For classification there are many evaluation metrics available and accuracy could be used when the dataset has class-balance and not skewed. Accuracy was measured on test set data after training the model using a training set data. A confusion matrix could be used for calculating accuracy and other metrics such as sensitivity, specificity, precision, recall and F-score used in ML. An example confusion matrix for a two class classification problem is given in table III.

Accuracy is defined as

$$Accuracy = (TP + TN)/(TP + FN + FP + TN) \quad (5)$$

, it can be put as correct predictions / total predictions as well.

Precision states that how much predicted as positive are actually positive and could be calculated using the equation

$$Precision = TP/(TP + FP) \quad (6)$$

Recall denotes the portion of positives which are classified correctly.

$$Recall = TP/(TP + FN) \quad (7)$$

We have used accuracy loss when comparing privacy vs utility [13], where it is defined as the loss in accuracy when DP is present against DP is not present.

$$Accuracy\ Loss(ACL) = 1 - \frac{ACC_{(m,\epsilon)}}{ACC_{(m,\epsilon=inf)}} \quad (8)$$

b) Measuring privacy via privacy attacks: AUC provides a measure to differentiate between positive classes and negative classes by plotting the relationship between the TPR and FPR in the ROC curve.

$$\begin{aligned} True\ Positive\ Rate(TPR) &= Sensitivty \\ &= TP/(TP + FN) = Recall \end{aligned} \quad (9)$$

$$\begin{aligned} False\ Positive\ Rate(FPR) &= 1 - Specificity \\ &= FP/(TN + FP) \end{aligned} \quad (10)$$

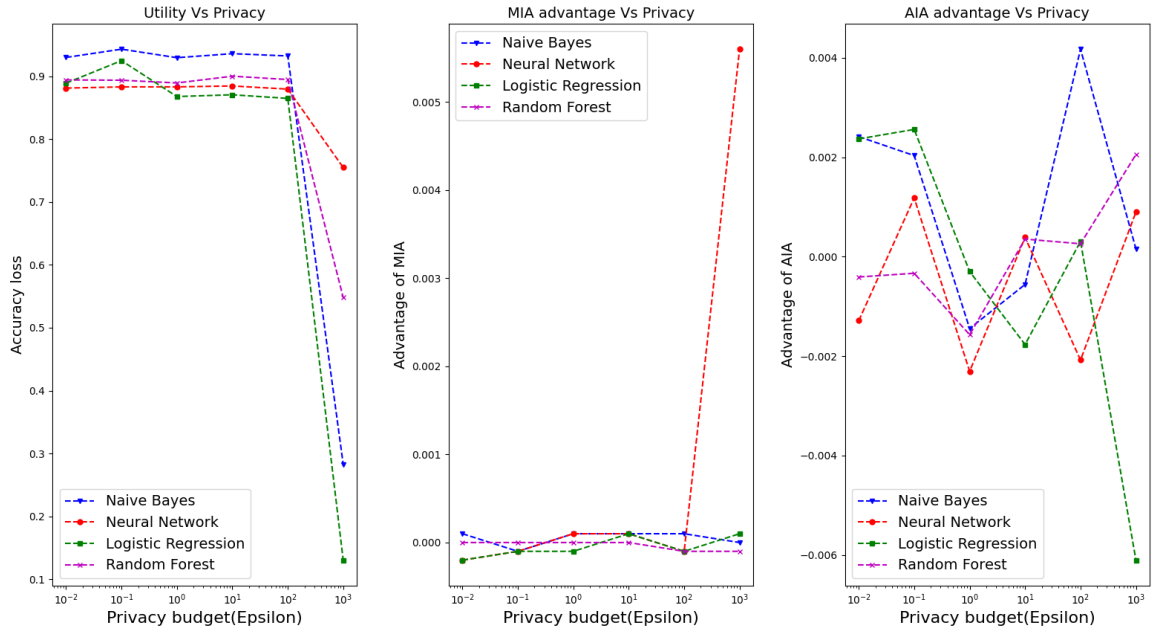

Fig. 2. Comparison among different ML methods with Netflix data of class size 20 and DPM1 using Laplace mechanism.

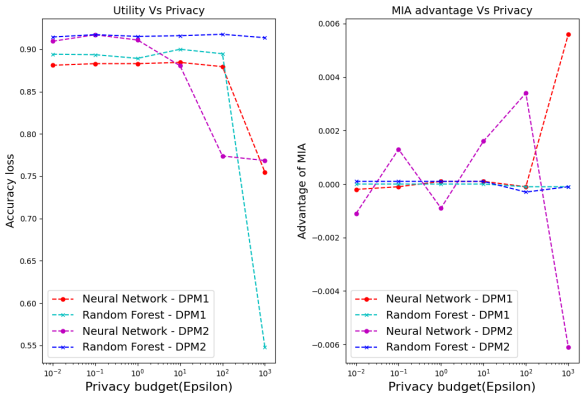

Fig. 3. Comparison of Neural Network and Random forest with Netflix data of class size 20 perturbed using DPM1 Vs DPM2

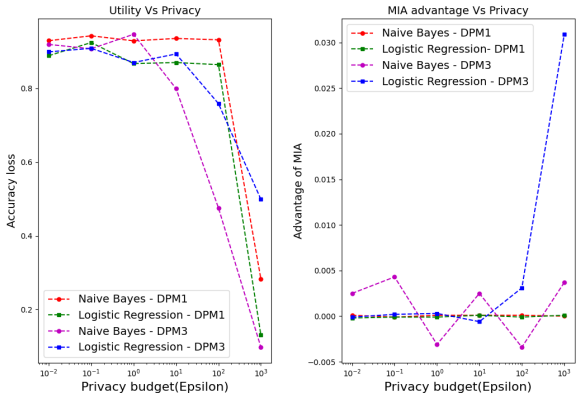

Fig. 4. Comparison of Naïve Bayes and Logistic Regression with Netflix data of class size 20 perturbed using DPM1 Vs DPM3

Difference between True Positive Rate (TPR) (Actual true or sensitivity) and False Positive Rate (FPR) (Actual false) is calculated as adversary advantage in privacy leak attacks here. Both these measures are used to measure privacy leaks.

IV. Results and Discussion

In this study, DP is applied on different ML classifiers at their different stages, as in table II and their accuracy is compared against the relevant model results without applying DP, to know the utility measure using equation 8. Further, Different mechanisms for DPM1 were considered. MIA and AIA are performed on these models to know the privacy impact when adding different levels of noise or privacy budget. Privacy budget was changed by varying ϵ in a range from 0.01 to 1000.

A set of experiments were performed on an equally split data set, where the first part is used for training and the other part is for testing. Using the model obtained from training data, accuracy was calculated on the test dataset.

Netflix dataset with different class sizes were tested to know the effect of number of classes on DP and further Purchase dataset with size 10000 was compared against size 80000 to

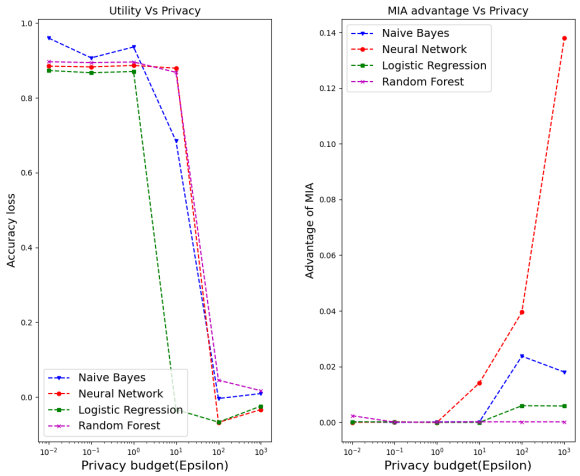

Fig. 5. DPM1 with Gaussian noise for Netflix data of class size 20

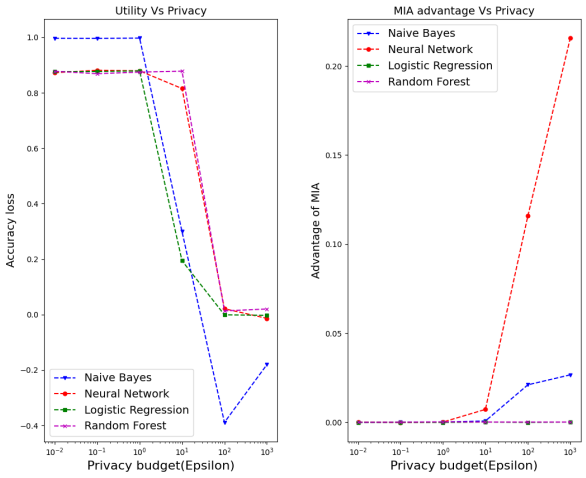

Fig. 6. Gaussian DPM1 on Purchase data of class size 20

know the effect of dataset size. But the size of the dataset change was not significant enough to observe considerable difference in results for the purchase dataset. In addition, AIA results didn't show much correlation with the privacy budget where MIA did it. Hence we skipped plotting it for the cases except in Fig 2.

Fig 2 provides metrics when DPM1 of Laplace mechanism is used with a Netflix dataset of 20 classes for different ML methods. It could be seen that there is considerable accuracy loss for all ML methods with low privacy budgets. Logistic Regression performs better once the budget is increased to the order of 1000. MIA for Neural Networks is badly affected by the increased privacy budget while other methods fail to provide much correlation for both MIA and AIA against privacy.

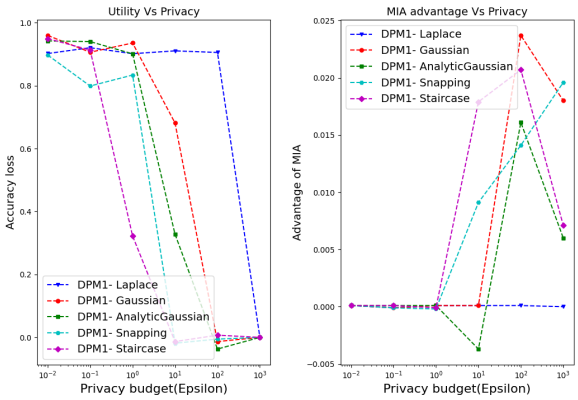

Fig. 7. Comparison of different DPM1 mechanisms for Netflix data of class size 20 with Naïve Bayes classifier

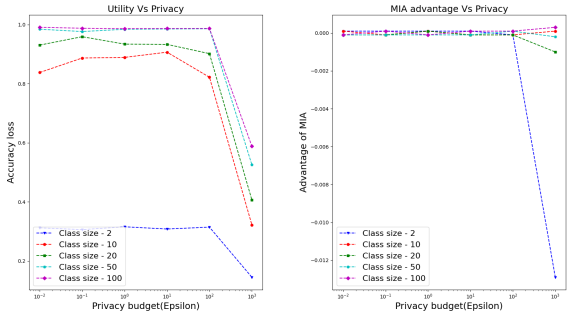

Fig. 8. Comparison of Naïve Bayes classifier with Netflix data of different class sizes using DPM1

From Fig 3 it is clear that DPM2 performs better when comparing to DPM1 with low privacy budget for Neural Network. Similarly DPM3 performs better than DPM1 for Naïve Bayes and Logistic regression, which could be seen in Fig 4.

We performed DPM1 using Gaussian noise instead of Laplace and results for Netflix data with 20 classes are given in Fig 5. It could be compared against Fig 2 which uses Laplace noise and it could be seen that Gaussian noise starts to provide more utility than Laplace, in the order of 10 in epsilon. The reason for this is because Laplace obeys ϵ DP but the Gaussian (ϵ, δ) DP. Fig 6 provides results for DPM1 with Gaussian for Purchase dataset with 20 classes.

Fig 7 provides results when different DPM1 mechanisms are used on Netflix data with 20 classes using Naïve Bayes classifier. Staircase mechanism shows more utility with low privacy budget and Laplace mechanism performs worst. In general, all improved versions provides better utility with low privacy budget when comparing to Laplace or Gaussian mechanism.

Fig 8 compares the metrics when varying the number of classes with DPM1 using Naïve Bayes on Netflix data. When the variation in class size increases, accuracy loss increases.

DP is becoming the quintessential mechanism to preserve privacy, and reflected in the results here as well. It is observed in almost all the graphs that increasing the privacy budget or epsilon value results in lesser accuracy loss and in turn increases the privacy risks by having higher MIA and AIA advantage. When comparing results among four models while keeping the noise, data , classes same, Random Forest classifier performs comparably better. When comparing DPM1 using Laplace Vs Gaussian, Gaussian provides better utility (accuracy) and Laplace provides better privacy. Improved versions of DPM1, notably the staircase mechanism provide better utility with low privacy budget and considerable protection against privacy attacks.

V. Conclusion and future work

In this work, we have conducted experiments of applying DP on different ML classification methods at different stages, i.e noise addition to data itself, during the training and to the ML parameters after the training and studied the utility Vs privacy relationship. Further how privacy withstands MIA, AIA is studied. How a ML method could be used for a particular task considering the privacy budget, Which stage of perturbation works better for a ML model, how a DPM1 method could be used for LDP tasks can be driven from this work.

Expanding the mechanisms across different data types and sizes is one of the next steps we aim to take. Further, FL is a promising area which supports privacy to an extent by its architecture itself [12]. We are interested in experimenting LDP with FL though there are some considerable work going on this area in recent times [35] [36] [37], they are in their initial stages and lacks the coverage in ML methods or data types considered.

References

[1] Avi Mendelson. Security and privacy in the age of big data and machine learning. *Computer*, 52(12):65–70, 2019.

[2] Bo Liu, Ming Ding, Sina Shaham, Wenny Rahayu, Farhad Farokhi, and Zihuai Lin. When machine learning meets privacy: A survey and outlook. *ACM Computing Surveys*, 54(2):1–36, Mar 2022.

[3] Frederik Armknecht, Colin Boyd, Christopher Carr, Kristian Gjøsteen, Angela Jäschke, Christian A Reuter, and Martin Strand. A guide to fully homomorphic encryption. *Cryptology ePrint Archive*, 2015.

[4] Chuan Zhao, Shengnan Zhao, Minghao Zhao, Zhenxiang Chen, Chong-Zhi Gao, Hongwei Li, and Yu-an Tan. Secure multi-party computation: theory, practice and applications. *Information Sciences*, 476:357–372, 2019.

[5] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020.

[6] Mohammed Binjubeir, Abdulghani Ali Ahmed, Mohd Arfian Bin Ismail, Ali Safaa Sadiq, and Muhammad Khurram Khan. Comprehensive survey on big data privacy protection. *IEEE Access*, 8:20067–20079, 2020.

[7] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, page 3–18, May 2017.

[8] Ricardo Mendes and João P. Vilela. Privacy-preserving data mining: Methods, metrics, and applications. *IEEE Access*, 5:10562–10582, 2017.

[9] Bargav Jayaraman and David Evans. Evaluating differentially private machine learning in practice. In *Proceedings of the 28th USENIX Conference on Security Symposium*, SEC'19, page 1895–1912, USA, 2019. USENIX Association.

[10] Kato Mivule, Claude Turner, and Soo-Yeon Ji. Towards a differential privacy and utility preserving machine learning classifier. *Procedia Computer Science*, 12:176–181, 2012. Complex Adaptive Systems 2012.

[11] M. R. Gorai, K.S. Sridharan, T. Aditya, R. Mukkamala, and S. Nukavarapu. Employing bloom filters for privacy preserving distributed collaborative knn classification. In *2011 World Congress on Information and Communication Technologies*, page 495–500, Dec 2011.

[12] Ahmed El Ouadrhiri and Ahmed Abdelhadi. Differential privacy for deep and federated learning: A survey. *IEEE Access*, 10:22359–22380, 2022.

[13] Benjamin Zi Hao Zhao, Mohamed Ali Kaafar, and Nicolas Kourtellis. Not one but many tradeoffs: Privacy vs. utility in differentially private machine learning. In *Proceedings of the 2020 ACM SIGSAC Conference on Cloud Computing Security Workshop*, page 15–26, Nov 2020. arXiv:2008.08807 [cs].

[14] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2013.

[15] Sam Fletcher and Md Zahidul Islam. Differentially private random decision forests using smooth sensitivity. *Expert Systems with Applications*, 78:16–31, Jul 2017. arXiv:1606.03572 [cs].

[16] Emre Yilmaz, Mohammad Al-Rubaie, and J. Morris Chang. Locally differentially private naive bayes classification. (arXiv:1905.01039), May 2019. arXiv:1905.01039 [cs, stat].

[17] Borja Balle and Yu-Xiang Wang. Improving the gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. (arXiv:1805.06530), Jun 2018. arXiv:1805.06530 [cs, stat].

[18] Ilya Mironov. On significance of the least significant bits for differential privacy. In *Proceedings of the 2012 ACM conference on Computer and communications security*, page 650–661, Raleigh North Carolina USA, Oct 2012. ACM.

[19] Quan Geng and Pramod Viswanath. The optimal mechanism in differential privacy. (arXiv:1212.1186), Oct 2013. arXiv:1212.1186 [cs].

[20] Mariana Cunha, Ricardo Mendes, and João P. Vilela. A survey of privacy-preserving mechanisms for heterogeneous data types. *Computer Science Review*, 41:100403, Aug 2021.

[21] Naoise Holohan, Stefano Braghin, Pól Mac Aonghusa, and Killian Levacher. Diffprivlib: the IBM differential privacy library. *ArXiv e-prints*, 1907.02444 [cs.CR], July 2019.

[22] Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. (arXiv:1605.02065), May 2016. arXiv:1605.02065 [cs, math].

[23] Martín Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, page 308–318, Oct 2016. arXiv:1607.00133 [cs, stat].

[24] Ilya Mironov. Renyi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, page 263–275, Aug 2017. arXiv:1702.07476 [cs].

[25] Majid Bashir Malik, M. Asger Ghazi, and Rashid Ali. Privacy preserving data mining techniques: Current scenario and future prospects. In *2012 Third International Conference on Computer and Communication Technology*, page 26–32, Nov 2012.

[26] Weina Wang, Lei Ying, and Junshan Zhang. On the relation between identifiability, differential privacy, and mutual-information privacy. *IEEE Transactions on Information Theory*, 62(9):5018–5029, Sep 2016.

[27] Ahmed Salem, Yang Zhang, Mathias Humbert, Mario Fritz, and Michael Backes. Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models. *CoRR*, abs/1806.01246, 2018.

[28] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st computer security foundations symposium (CSF)*, pages 268–282. IEEE, 2018.

[29] Privateutility/privateutility. [Online]. Available: https://github.com/PrivateUtility/PrivateUtility, Aug 2021. Accessed: 2022-10-18.

[30] Netflix prize dataset. [Online]. Available: https://www.kaggle.com/datasets/netflix-inc/netflix-prize-data. Accessed: 2022-10-18.

[31] Acquire valued shoppers challenge. [Online]. Available: https://www.kaggle.com/c/acquire-valued-shoppers-challenge/data. Accessed: 2022-10-18.

[32] Kamalika Chaudhuri, Claire Monteleoni, and Anand D. Sarwate. Differentially private empirical risk minimization. (arXiv:0912.0071), Feb 2011. arXiv:0912.0071 [cs].

[33] Jaideep Vaidya, Basit Shafiq, Anirban Basu, and Yuan Hong. Differentially private naive bayes classification. In *2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, page 571–576, Atlanta, GA, USA, Nov 2013. IEEE.

[34] Tensorflow privacy. [Online]. Available: https://github.com/tensorflow/privacy. Accessed: 2022-10-18.

[35] Kang Wei, Jun Li, Ming Ding, Chuan Ma, Howard H. Yang, Farhad Farokhi, Shi Jin, Tony Q. S. Quek, and H. Vincent Poor. Federated learning with differential privacy: Algorithms and performance analysis. *IEEE Transactions on Information Forensics and Security*, 15:3454–3469, 2020.

[36] Stacey Truex, Ling Liu, Ka-Ho Chow, Mehmet Emre Gursoy, and Wenqi Wei. Ldp-fed: federated learning with local differential privacy. In *Proceedings of the Third ACM International Workshop on Edge Systems, Analytics and Networking*, page 61–66, Heraklion Greece, Apr 2020. ACM.

[37] Pathum Chamikara Mahawaga Arachchige, Dongxi Liu, Seyit Camtepe, Surya Nepal, Marthie Grobler, Peter Bertok, and Ibrahim Khalil. Local differential privacy for federated learning. In Vijayalakshmi Atluri, Roberto Di Pietro, Christian D. Jensen, and Weizhi Meng, editors, *Computer Security – ESORICS 2022*, page 195–216, Cham, 2022. Springer International Publishing.