COLLEGE OF ENGINEERING
COMPUTER SCIENCE
VIRGINIA TECH

Project:
Collaborative Learning Through Multiple Private Datasets Ensuring Data Privacy

Author:
Tanmoy Sarkar Pias
PhD Student
CS, VT

Instructor:
Professor Daphne Yao
Elizabeth and James E. Turner Jr. '56 Faculty Fellow
CACI Faculty Fellow
CS, VT

Course Project

of

AI Technologies for Cybersecurity Defenses

CS 6804 (Spring 2021)

Contents

1 Abstract

The performance of a machine learning algorithm mostly depends on the nature of the data on which it is trained on. Sometimes, these models require sensitive and personal information to make themselves useful. To ensure the privacy of the data, only the trained models are published. However, model inversion attack and membership linkage attacks can effectively reconstruct the training dataset by only accessing the public model as a black-box. This is a serious privacy concern. In this project, I have implemented a privacy-preserving machine learning framework that can effectively mitigate the mentioned attacks. In this framework, the available private datasets are used to a set of train private models. Then these models make predictions to a similar publicly available unlabeled dataset. At the time of making a prediction, a certain level of noise is added to make the process differentially private. Finally, a public model is trained on the newly annotated public dataset. In this project, it has been established that both high data utility and strong data privacy can be achieved at the same time by increasing the number of teacher models.

2 Introduction

The machine learning (ML) models are being used in many areas nowadays. Sometimes, ML models require sensitive data i.e. healthcare data, credit card data, personal information, and so on. To protect the privacy and to keep the data secured, only the model is published and applied in real-life. However, keeping the training data in a safe place is not enough for ensuring complete privacy. The training data can be extracted from the model [2], [3]. So, necessary steps should be taken when the model is trained on sensitive data.

Many attempts have been made to keep sensitive data safe and private while maintaining the data utility. Federated machine learning is one kind of distributed machine learning approach where the global model on the server continuously learns from the client edge device. The client-side edge device interacts with the user and most of the time edge device learns from the user's personal information and preferences. The edge model learns from these user-data and to keep the user-data private, only the client model's weights are sent to the server model. However, this approach is not completely safe. Wei et al. proposed a framework to improve the privacy of federated machine learning by using the differential privacy [1]. Hao et al. proposed a similar framework by combining federated learning with differential privacy [8]. In this paper, the effectiveness is analyzed with real-world experimentation. Zhao et al. [6] proposed a privacy-preserving block-chain based federated learning. Hu et al. [7] proposed a personalized version of federated learning using differential privacy.

Another semi-supervised differentially private framework called Private Aggregation of Teacher Ensembles (PATE) has been proposed by Papernot et al. [4], [5]. The framework is used to annotate a public dataset with the help of multiple private datasets. The whole process is differentially private. In this paper, the author demonstrated the performance of their framework on datasets like MNIST. However, MNIST doesn't contain any sensitive data and works as a dummy dataset.

So, based on the concepts of PATE, I have implemented a similar framework to evaluate the performance using real-life medical data which has sensitive information like age, marital status, and other personal data. Moreover, the relationship between data utility and data privacy has also been experimentally analyzed. Another important factor in this framework is the teacher model or private machine learning model. Interestingly, the number of teacher models can effectively control data privacy and data utility. In this project, the impact of the teacher model has also been demonstrated empirically. Moreover, this framework prevents two major threats which are discussed in the following section.

3 Threat Model

The machine learning model entirely depends on the dataset. However, this ML model can be a threat to the privacy of the data by which it is trained. Today data cannot even be published with anonymization because of the data linkage attack [10]. So, sensitive data is better to keep private. However, even though private data can be exploited by only accessing the trained model.

There are two main threat models for publicly available machine learning model
1. Model Inversion Attack
2. Membership Inference Attack

3.1 Model Inversion Attack

This attack was introduced by Fredrikson et al. [2] back in 2015. In this experimental study, it has been established that the adversary can retrieve the training data points only if it has access to the machine learning model. It's like inverting the ML model to extract the data points which were used to train the model. This study shows that a machine learning model which is used for facial recognition can be used to reconstruct the real subject's facial image with a high confidence value. However, this attack can be extended to extract a significant number of data points only by inspecting the model. This is a serious privacy concern. In this project, I have applied the concept of differential privacy to protect the data against this model inversion attack. Detailed implementation is explained in the following sections.

3.2 Membership Inference Attack

Membership inference attack determines if a particular data-point exists in a dataset on which the machine learning model is trained on. Under this attack, the adversary treats the machine learning model as a black-box and builds an inference model. The adversary uses this inference model to recognize the difference in the predictions of the target model on two different sets of data points. This threat model assumes that the prediction result will be different on the data points on which it is trained versus on the data points on which it is not trained. Shokri et al. [3] established this hypothesis by empirical evaluation. Therefore, this thread model can effectively determine the existence of a certain data point in a private dataset. This is also a serious threat to the privacy of the dataset. Interestingly, this threat can also be prevented by leveraging the power of differential privacy.

4 Novelty Statement

I have implemented a private machine learning framework which is a follow-up work of Private Aggregation of Teacher Ensembles (PATE) by Papernot et al. [4; 5]. A major drawback of the existing similar models is that those approaches are evaluated with some dummy datasets like MNIST. However, medical datasets are different in nature. So, in this project, I have used a real medical dataset and analyzed the scalability, and evaluated the performance empirically. The noble contribution of these projects are:

1. Implementation of a collaborative machine learning framework where an unlabeled dataset can be annotated by using differentially private local models trained on private datasets.

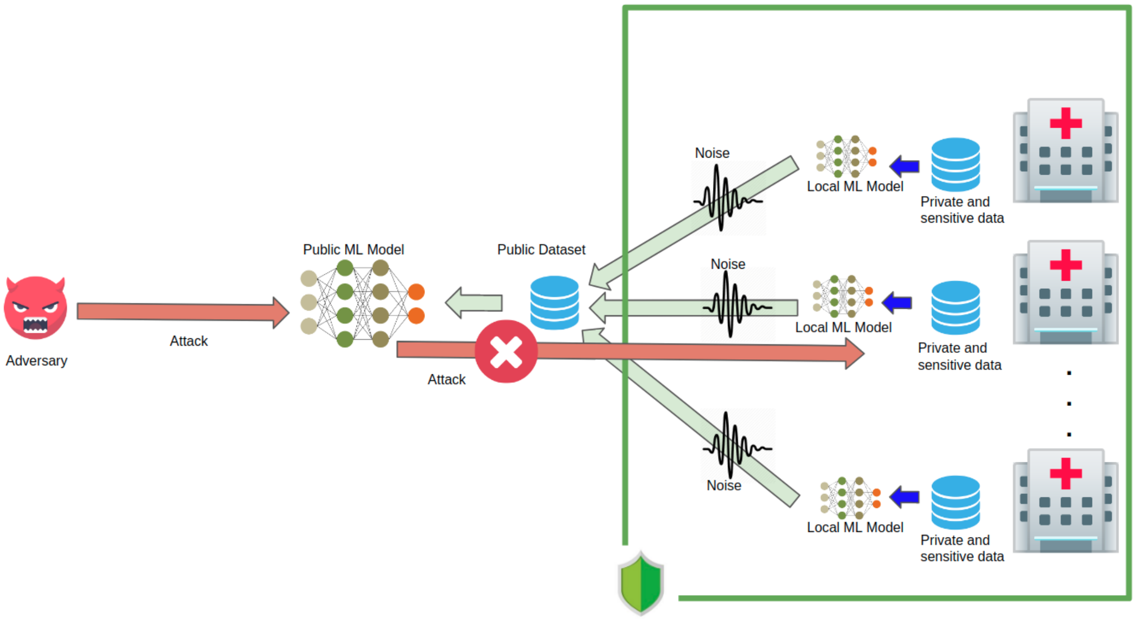
2. Empirical analysis of different attributes to understand the trade-off between data privacy and data utility under collaborative environment.

3. Achieve a very high data utility (accuracy) under strong privacy constraints (low privacy budget epsilon).

5 Work Performed

I have implemented the collaborative machine learning framework and evaluate the performance on a stroke prediction dataset [11]. This dataset has 12 columns and 4909 rows. At first, the categorical attributes are converted into numerical values by one-hot-encoding. The ID attribute is removed because it doesn't contribute to the stroke prediction. The dataset is split into a train (70%) and test (30%) sets.

This framework is divided into two main stages. At first, each of the privately available datasets is used to train a private machine learning model which is not accessible by the adversary. In this scenario, there can be n (>1) number of machine learning models if n number of private datasets are available. These private ML models are called teacher models. In this experiment, the training set is split into n equal subsets and n number of artificial neural network models is built. Each of the n models are trained on one of the n subsets.



After that, the test set (30%) is again split into a student-train set (70%) and a student-test set (30%). The trained n-teacher models make prediction for each of the data points of the student-train set. At the time of making a prediction, Laplacian noise is added. This makes this whole framework differentially private.

$$\textit{Lapalican Noise: } f(x|\mu, b) = \frac{1}{2b}\exp\left(-\frac{|x-\mu|}{b}\right) \tag{1}$$

So, there are n labels for each data-points and one label is selected based on the majority. This is how the whole student-train set which is representative of a public dataset, is annotated by the teacher models. Then this public dataset is used to train a student artificial neural network model which represents the public model. Though the model is public, the adversary can use neither the model inversion attack nor membership inference attack to exploit the privacy of the private dataset because of differential privacy. The public student model is evaluated using the student-test dataset.

In this experiment, the impact of two main key attributes which are the privacy budget (epsilon = 1/b) and the number of teacher models, are analyzed. The simulated results are discussed in the following sections.

6 Findings

There are two main findings of this experiment.

1. Epsilon means the privacy budget. Privacy budget gives a sense of how much information leakage can be allowed. The smaller value of epsilon, the more data protection is guaranteed. However, smaller epsilon is equivalent to adding more randomized noise to the predicted label. In the left graph, it is evident that accuracy increase with the increment of the privacy budget but the privacy budget is inversely proportional to data protection. Therefore, it can be concluded that data utility and privacy are inversely proportional.

2. Interestingly, the right graphs show that the student model accuracy increases with the number of teacher models. This graph is generated under a constant epsilon of 0.5. The graph shows that the model can achieve almost perfect accuracy with 20 teacher models. Therefore, both good data utility and high data privacy can be achieved with a sufficient number of teacher models.

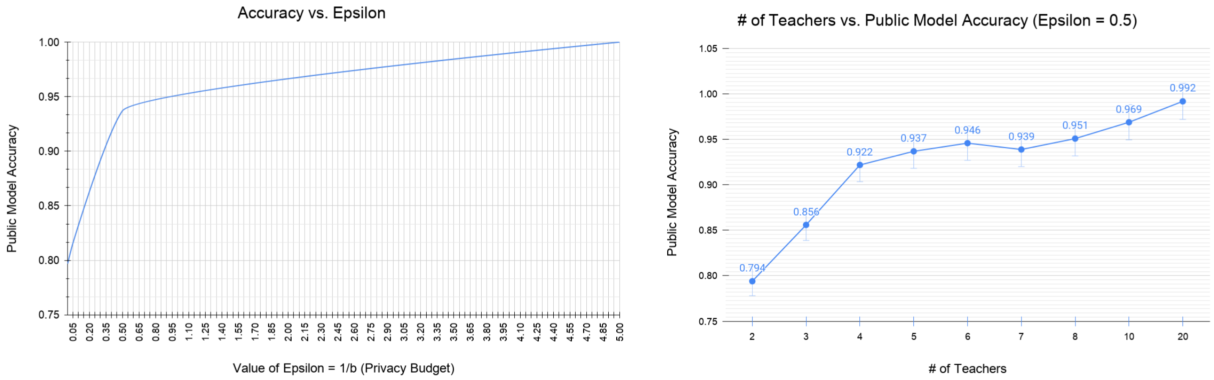


Figure 1: Left graphs represents the accuracy of the student model for different epsilon value with a constant number of teacher model=5 and the right graph represents the accuracy of the student model when the number of teacher model is varied with a constant epsilon value of 0.5

In summary, the experimental result shows that data privacy and data utility, two opposite constraints, both can be enhanced by increasing the number of teacher models.

7 Conclusions

This privacy-preserving machine learning framework can effectively prevent two deadly attacks which are membership inference attacks and model inversion attacks. Moreover, this project shows that this framework can effectively protect the privacy of sensitive healthcare data. In addition, this approach can be used to annotate an unlabeled public dataset.

This project can be extended further. In this project, only one healthcare dataset has been used. However, in practice, there can be different datasets. So, there is a scope to analyze how the diversity of the dataset affects the accuracy of the public model. Moreover, using multiple different datasets will generalize the claims of this framework. A limitation of this project is that the proposed framework can only annotate a publicly available dataset, not a private dataset. This limitation can be solved by applying another layer of differential privacy on the student dataset and that is not done in this project. This can be a significant improvement over the existing framework proposed in this project.

References

[1] Wei, Kang, Jun Li, Ming Ding, Chuan Ma, Howard H. Yang, Farhad Farokhi, Shi Jin, Tony QS Quek, and H. Vincent Poor. "Federated learning with differential privacy: Algorithms and performance analysis." IEEE Transactions on Information Forensics and Security 15 (2020): 3454-3469.

[2] Fredrikson, Matt, Somesh Jha, and Thomas Ristenpart. "Model inversion attacks that exploit confidence information and basic countermeasures." In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, pp. 1322-1333. 2015.

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

[4] Papernot, Nicolas, Martín Abadi, Ulfar Erlingsson, Ian Goodfellow, and Kunal Talwar. "Semi-supervised knowledge transfer for deep learning from private training data." arXiv preprint arXiv:1610.05755 (2016).

[5] Papernot, Nicolas, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Úlfar Erlingsson. "Scalable private learning with pate." arXiv preprint arXiv:1802.08908 (2018).

[6] Zhao, Yang, Jun Zhao, Linshan Jiang, Rui Tan, Dusit Niyato, Zengxiang Li, Lingjuan Lyu, and Yingbo Liu. "Privacy-preserving blockchain-based federated learning for IoT devices." IEEE Internet of Things Journal (2020).

[7] Hu, Rui, Yuanxiong Guo, Hongning Li, Qingqi Pei, and Yanmin Gong. "Personalized federated learning with differential privacy." IEEE Internet of Things Journal 7, no. 10 (2020): 9530-9539.

[8] Hao, Meng, Hongwei Li, Xizhao Luo, Guowen Xu, Haomiao Yang, and Sen Liu. "Efficient and privacy-enhanced federated learning for industrial artificial intelligence." IEEE Transactions on Industrial Informatics 16, no. 10 (2019): 6532-6542.

[9] Lu, Yunlong, Xiaohong Huang, Yueyue Dai, Sabita Maharjan, and Yan Zhang. "Differentially private asynchronous federated learning for mobile edge computing in urban informatics." IEEE Transactions on Industrial Informatics 16, no. 3 (2019): 2134-2143.

[10] Schnell, Rainer, Tobias Bachteler, and Jörg Reiher. "Privacy-preserving record linkage using Bloom filters." BMC medical informatics and decision making 9, no. 1 (2009): 1-11.

[11] https://www.kaggle.com/fedesoriano/stroke-prediction-dataset . Accessed: May 8, 2021.

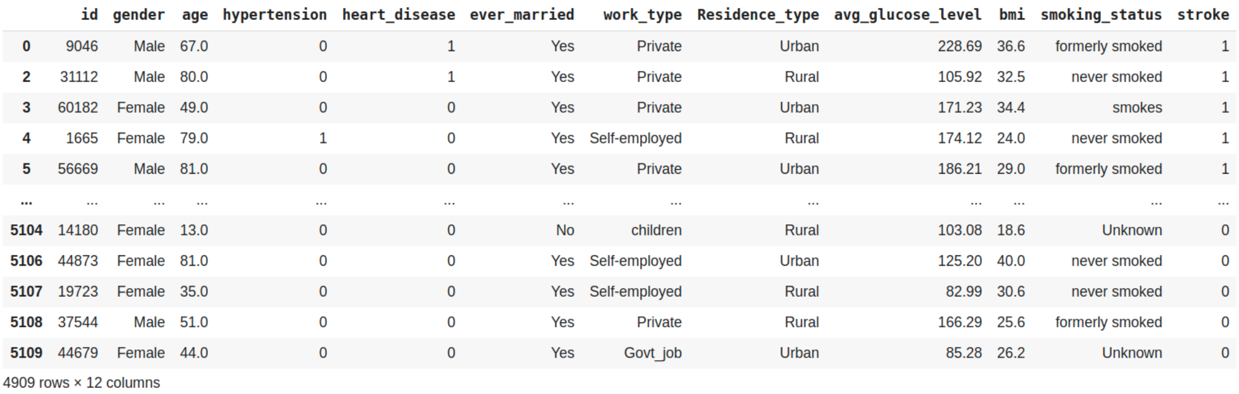
[12] en.wikipedia.org/wiki/Laplace_distribution . Accessed: May 8, 2021.

[13] numpy.org/doc/stable/reference/random/generated/numpy.random.laplace.html . Accessed: May 8, 2021.

Appendices

A Stroke prediction dataset

This dataset is collected from the Kaggle [11]. This is a stroke prediction binary dataset and one of the trending datasets on Kaggle. There is a total of 12 columns and around 5,000 rows. Some rows contain NULL values. All the attributes detailed information is given below.



	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
2	31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
3	60182	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
4	1665	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1
5	56669	Male	81.0	0	0	Yes	Private	Urban	186.21	29.0	formerly smoked	1
...	...	...	...	...	...	...	...	...	...	...	...	...
5104	14180	Female	13.0	0	0	No	children	Rural	103.08	18.6	Unknown	0
5106	44873	Female	81.0	0	0	Yes	Self-employed	Urban	125.20	40.0	never smoked	0
5107	19723	Female	35.0	0	0	Yes	Self-employed	Rural	82.99	30.6	never smoked	0
5108	37544	Male	51.0	0	0	Yes	Private	Rural	166.29	25.6	formerly smoked	0
5109	44679	Female	44.0	0	0	Yes	Govt_job	Urban	85.28	26.2	Unknown	0

4909 rows × 12 columns

Figure 2: Stroke Prediction Dataset

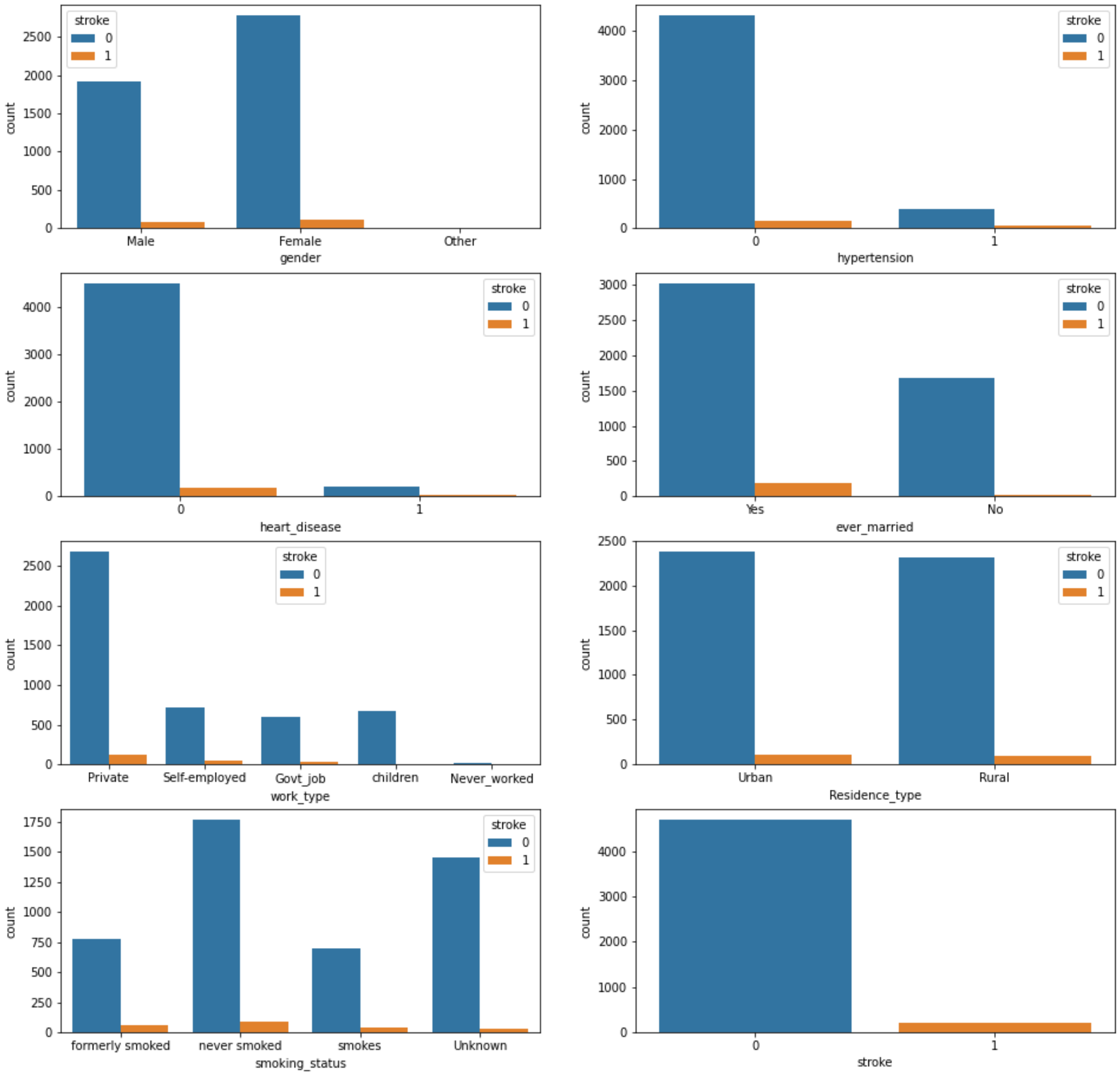


Figure 3: Dataset Feature Count

Attributes:
1) id: unique identifier
2) gender: "Male", "Female" or "Other"
3) age: age of the patient
4) hypertension: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension
5) heart_disease: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
6) ever_married: "No" or "Yes"
7) work_type: "children", "Govt_jov", "Never_worked", "Private" or "Self-employed"
8) Residence_type: "Rural" or "Urban"
9) avg_glucose_level: average glucose level in blood
10) bmi: body mass index
11) smoking_status: "formerly smoked", "never smoked", "smokes" or "Unknown"*
12) stroke: 1 if the patient had a stroke or 0 if not
*Note: "Unknown" in smoking_status means that the information is unavailable for this patient

In this experiment, those rows are dropped to make the dataset simpler for processing. The categorical attributes are then converted to numerical labels by the one-hot-encoding technique. The ID is just a numerical identifier and doesn't contribute to stroke prediction. So, ID is removed from the dataset. From figure 3, it is evident that this dataset is highly imbalanced as there are a few data points of stroke patients compared to benign patients.

B Result

Table 1: Classification Result for Different Values of Epsilon

Number of Teachers	**Teacher Accuracy (Max)**	**Epsilon (privacy budget) = 1/b**	**Number of Wrong Labels / Total Datapoints (after adding noise)**	**Student Accuracy**
5	0.984	0.005	58/295	0.797
5	0.984	0.05	55/295	0.815
5	0.984	0.5	17/295	0.937
5	0.984	5	0/295	1.00

Each of the trained teacher models predicts a label of a particular data-point in the student train set. So, each data point has n number of labels associated with it. At the time of prediction, a certain amount of noise is added to the predicted label. This is called differential privacy. In general, there are two kinds of noise are used which are Gaussian Distribution and Laplacian Distribution.

For many problems in healthcare, Laplace distribution models better than the Gaussian distribution. For this reason, in this project, the Laplace distribution has been selected. The Laplace distribution is controlled by two attributes σ and b where σ is the mean of the distribution and the b can be seen as the variance or expansion of the distribution center the mean.

Another important attribute of differential privacy is the estimation of Epsilon (ϵ). ϵ means the privacy budget. The higher privacy budget means more data leakage can be allowed. If the value of ϵ is zero (0), it means there is no data leakage and the data privacy is perfectly preserved. So, the goal is to keep the privacy budget as small as possible.

In this project, the ϵ is set to the inverse of b ($\epsilon = 1/b$). So, if the privacy budget decreases, the Laplace

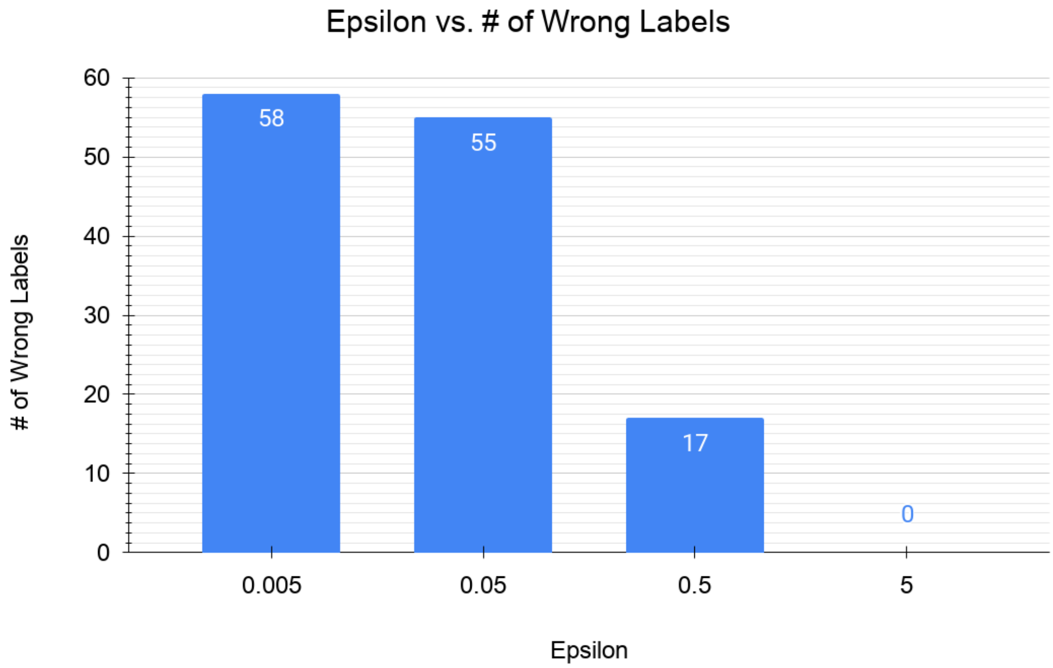


Figure 4: Value of epsilon (privacy budget vs. number of wrong labels in the public dataset

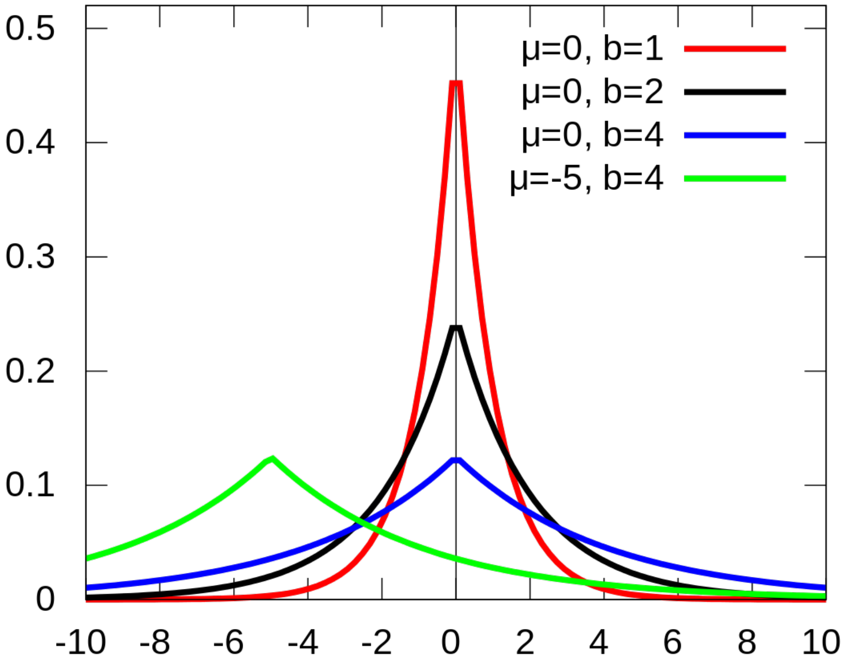


Figure 5: Laplace distribution for different b [12]

variance increase which means more randomness. This makes sense because if the privacy budget is low that means less privacy leakage is allowed, so randomization is added to the prediction.

Table 1 and figure 4 show the relationship between Epsilon and the accuracy of the student model. In the first row, the $\epsilon = 0.005$ which means the privacy budget is very low and a little data leakage is allowed. This means b = 1/0.005 = 200 which is very high. This introduces more randomness to the result. The 4th column shows that how many labels are miss-matched after adding noise. For this, row there are 58 false labels out of a total of 295 data points. This results in a low student model accuracy which is 79%. On the other hand, when the privacy budget is high (Epsilon = 5 and b = 0.2), there are no wrong labels in the student train dataset which results in a perfect accuracy of the student model. So, the relationship between the privacy budget and the model accuracy is proportional.

Table 2: Classification Result for Different Number of Teachers

Number of Teachers	**Teacher Accuracy (Max)**	**Epsilon (privacy budget) = 1/b**	**Number of Wrong Labels / Total Datapoints (after adding noise)**	**Student Accuracy**
2	0.984	0.5	62/295	0.794
3	0.984	0.5	43/295	0.856
4	0.984	0.5	21/295	0.922
5	0.984	0.5	17/295	0.937
6	0.984	0.5	16/295	0.946
7	0.984	0.5	19/295	0.939
8	0.984	0.5	15/295	0.951
10	0.984	0.5	10/295	0.969
20	0.984	0.5	2/295	0.992

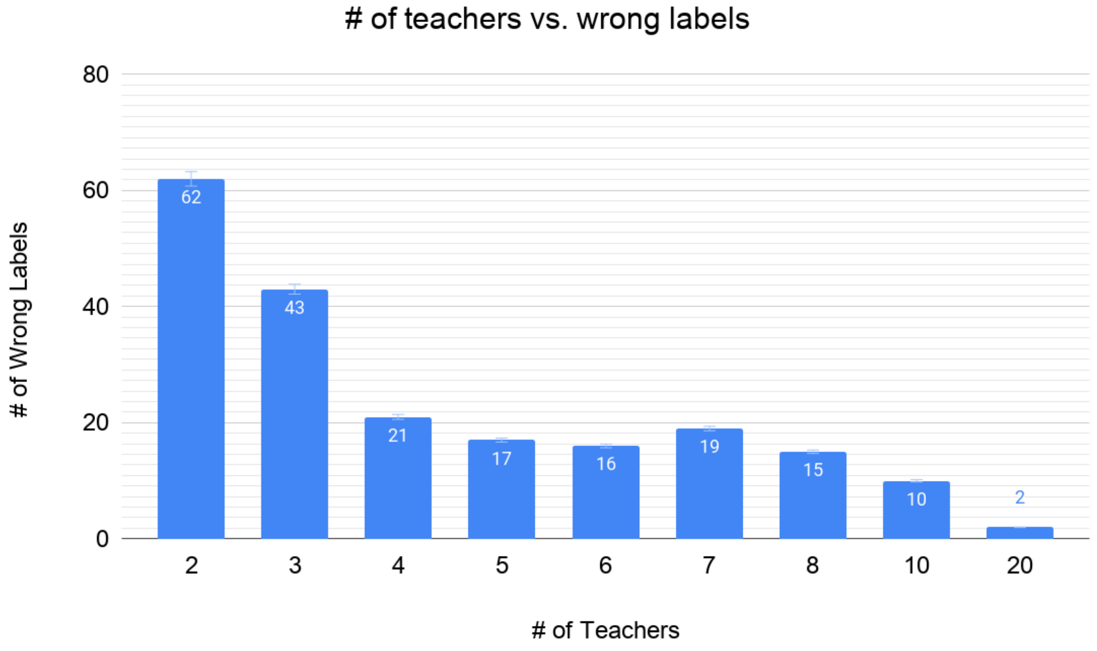


Figure 6: Number of teacher vs. number of wrong labels in the public dataset

Interestingly, the conflicting requirement between data utility and data privacy can be satisfied by increasing the number of teacher models. Table 2 and figure 6 represent the relationship between the number of teachers and the accuracy. For this experiment, the value of epsilon is fixed to 0.5. When there are only two teacher models the number of wrong labels is 62 out of 295 which is very high. And this leads to a poor model accuracy of 79%.

When the number of teacher models increases the number of wrong labels decreases. When the number of teachers is 10, there are only 10 wrong labels due to added noise. And this leads to a very high model accuracy of 96% compared to 78% which is achieved with 2 teacher models.

So, it can be concluded that both data utility and data privacy can be increased if the number of teacher models is sufficient.