End-to-End Privacy-Preserving Vertical Federated Learning using Private Cross-Organizational Data Collaboration

Anonymous Author(s)

ABSTRACT

As data utilization in organizations is advancing in various fields, insights that data brings will be more diverse when it is sourced through collaboration across different organizations than from a single organization. However, such data collaboration amongst organizations raises an issue of privacy protection. Federated learning, a method of building a machine learning (ML) model with distributed data across organizations, protects privacy by sharing only the model parameters and the information necessary for model update, without having to share the data each organization holds. On the other hand, it has been pointed out that data used for training may be leaked even from just the gradient necessary for model updates. To prevent such privacy leakage, local differential privacy can be applied where noise is added to the gradient to be shared in the model training in each organization. However, there is a problem with local differential privacy, where the amount of noise increases, leading to the degradation in model accuracy. In this paper, we propose a method of reducing the impact of noise compared to conventional federated learning by leveraging private cross-organizational data collaboration, called Private Cross-aggregation Technology (PCT). PCT combines Private Set Intersection Cardinality, Trusted Execution Environment and Differential Privacy, and outputs a cross-tabulation table that is private from input to output. Our method consists of two steps: (1) creating a private cross-tabulation table using PCT, and (2) training a ML using the private cross-tabulation table. In our implementation, we train a Naive Bayes classifier as an ML model. To confirm the effectiveness of the proposed method, we conducted an accuracy evaluation of the classification problem using DP-SGD, which is a safe learning method for deep learning used in federated learning, and the proposed method. We confirmed that the classification accuracy of the proposed method is higher in situations where the privacy budget is limited.

CCS CONCEPTS

• **Computing methodologies** → **Distributed algorithms**; **Machine learning**.

KEYWORDS

Vertical Federated Learning

FedKDD '24, August 25–29, 2024, Barcelona, Spain

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/12.345/

ACM Reference Format:
Anonymous Author(s). 2024. End-to-End Privacy-Preserving Vertical Federated Learning using Private Cross-Organizational Data Collaboration. In *FedKDD '24: International Joint Workshop on Federated Learning for Data Mining and Graph Analytics, August 25–29, 2024, Barcelona, Spain.* ACM, New York, NY, USA, 5 pages. https://doi.org/12.345/

1 INTRODUCTION

Data-driven decision-making and the utilization of artificial intelligence are advancing in various organizations. While data held by a single organization only reflects a limited aspect of what to be captured, e.g. user behavior, cross-organizational data collaboration will bring more diverse insights, enabling us to capture user behavior in a multifaceted manner.

A method of building a machine learning (ML) model in a situation where data is distributed across organizations is federated learning(FL) [17]. FL is categorized into horizontal FL and vertical FL based on the way the target data is partitioned. In horizontal FL, each organization has the same data items (features), and FL is used to increase the number of records in training data. On the other hand, in vertical FL, each organization has different features for records with the same identifier (e.g., user ID), and data are federated to increase the number of features. As mentioned above, we would like to capture user behavior from multiple perspectives, therefore in this paper, we focus on vertical FL setting. In FL, data is not shared between organizations, and only the parameters of the model and the information necessary for model updates are shared to protect privacy, which corresponds to the information of weights and gradients in the case of deep learning. On the other hand, it has been pointed out that the data used for training may leak from the gradient information alone [18]. Thus, a method of securely training the model by adding noise to the shared gradient on each client (i.e., applying local differential privacy) has been proposed [3]. However, since noise is added to the gradient on each client, there is a problem of degradation in model accuracy due to large amounts of noise.

In this paper, we propose an approach for vertical FL that reduces the impact of noise by leveraging private cross-organizational data collaboration, called Private Cross-aggregation Technology (PCT). PCT combines Private Set Intersection Cardinality (PSI-CA), Trusted Execution Environment (TEE), and Differential Privacy (DP), and outputs a cross-tabulation table that is private from input to output. Our proposed method consists of two steps: (1) securely creating cross-tabulation tables across organizations using PCT, and (2) training a ML using the private cross-tabulation tables. In our implementation, we train a Naive Bayes classifier as an ML model because of the following two reasons. First, in the creation of cross-tabulation tables, if the supervision label is set on the side of the table and the feature is set on the head of the table, the parameters of the Naive Bayes classifier are calculated using the values of each cell in the cross-tabulation table itself. Secondly, although

many existing studies on federated learning focus on deep learning models, it has been pointed out that deep learning is not always effective for tabular data [4]. As an initial consideration, we assume collaboration between two organizations, with Organization A holding features and Organization B holding supervision labels. To confirm the effectiveness of the proposed method, we compared the proposed method with DP-SGD [1] which is a representative method combining deep learning with DP. As a result, for small privacy budgets (e.g.,1 and 0.5), the proposed method had a higher classification accuracy. In addition, DP-SGD with small privacy budget (less than 0.5) cannot be trained due to the lack of privacy budget. These results confirmed the effectiveness of the proposed method.

The proposed method has three advantages over existing methods. The first is that there is no need to place a trustworthy external server. The second is that if there is no fraud by oneself, models can be trained securely. These two advantages are originated by PCT. The third is that the proposed method can reduce noise effects by applying central DP by first performing private data collaboration, while existing methods have realized secure model training by locally adding noise (i.e., local DP) based on differential privacy to the gradient needed for model update at each client.

The contributions of this paper are as follows.

- We propose an approach for vertical FL that combines Private Cross-aggregation Technology that allows for the creation of private cross-tabulation tables and an ML model that can be trained from a cross-tabulation table. Our approach has three advantages: (1) no need to place a trustworthy external server, (2) if there is no fraud committed by oneself, models can be trained securely, and (3) the impact of noise is reduced compared to a local-DP-based method.
- As an example, we demonstrate a case of Naive Bayes classifier trained from the private cross-tabulation tables. We leverage the fact that the parameters of the Naive Bayes classifier can be calculated using the values of each cell in the cross-tabulation table itself.
- We confirmed the effectiveness of the proposed method by comparing the proposed method with DP-SGD. The classification accuracy of the proposed method is higher than that of the baseline method when the privacy budget is small (such as 1 or 0.5) which indicates strong privacy protection.

2 PRELIMINARY AND RELATED WORK

2.1 Preliminary: Private Cross-aggregation Technology

In order to securely create statistical information without violating privacy, based on the data each organization holds across organizations, the following two requirements need to be met:

(1) The output data should be *statistical information* that properly protects privacy.
(2) Unless there is no fraud committed by oneself, information about their own data will not leak to others beyond the statistical information produced.

Private Cross-aggregation Technology (PCT) is a method for creating secure statistical information across organizations without violating privacy, based on the data held by each organization, by PSI-CA, TEE, and DP [5, 15]. PCT produces a cross-tabulation table that meets the differential privacy and consists of the following three processing steps:

(1) De-identification process that a hash function is applied to the data because no individuals could be re-identified before data linking.
(2) Secure aggregation process that aggregates anonymized hashed data using PSI-CA, a Homomorphic-Encryption-based method.
(3) Disclosure limitation process that adds noise of DP to aggregated data while being encrypted.

By such processing, PCT enables the creation of secure statistical information with privacy protection without revealing data to other organizations. Moreover, PCT ensures information about their own data will not be leaked to others, except the output statistical information, unless it is due to fraud committed by oneself. For technical details, please refer to [5, 15].

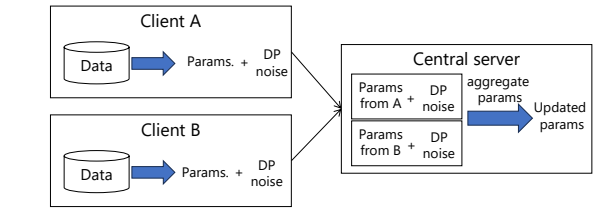
2.2 Related Work

Federated Learning (FL) is a distributed ML method proposed by Google [14]. FL is used to train deep learning models with data kept on each client device without aggregating data on a central server, targeting numerous mobile devices. The scope of FL has since been expanded to include data collaboration between organizations. The former problem setting is called cross-device FL, and the latter is called cross-silo FL [10]. In the cross-device setting, it assumes a very large number of client devices (e.g., millions or hundreds of millions), while in the cross-silo setting, it assumes data collaboration between at least two organizations [10]. In this paper, we assume the cross-silo setting.

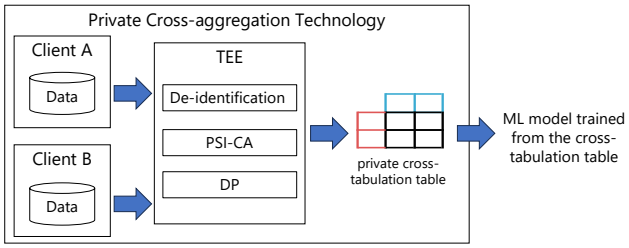
In FL, a method called local differential privacy (local DP), which applies noise based on the differential privacy to the learning data and the information (gradients) needed for model updates on each client, is used to securely train an ML model even in situations where the central server, which updates the model parameters, is not trustworthy. Although local DP is useful when there is a very large number of clients, as is assumed in cross-device setting, it has been pointed out that it is difficult to implement local DP while maintaining its utility when the number of clients is small [10]. In this regard, several studies have been conducted to reduce the impact of noise by combining Local DP and secure multi-party computation [9, 16]. In general, central DP, which applies DP after aggregating data in one place, has more utility than local DP. It is problematic, however, in that a trustworthy aggregation server is needed to apply central DP in existing federated learning [10].

As mentioned so far, while FL usually focuses on deep learning models, this paper focuses on Naive Bayes classifiers. There has also been studies targeting Naive Bayes classifiers in the framework of FL, such as [8, 13]. These studies calculate the co-occurrence frequency of supervision labels and features at each client in the framework of horizontal FL, as well as the noise based on DP in each client. Finally, the the co-occurrence frequencies are summed up at the aggregation server.

Our study differs from the existing studies because the order of data collaboration and model training is different from existing



(a) General flow of conventional federated learning



(b) Flow of the proposed approach

Figure 1: (a) General flow of existing FL method and (b) the flow of the proposed approach. The order of data collaboration and model training is different.

FL methods including deep learning. A general flow of existing FL methods and the flow of the proposed approach are shown in Figure 1. A method of federated learning that creates a cross-tabulation table through private data collaboration and builds a machine learning model from it in vertical federated learning has not been proposed so far.

3 PROPOSED METHOD

This section explains the proposed method of safely learning a machine learning model across organizations. To securely learn the Naive Bayes classifier in the framework of vertical federated learning, the process is carried out in the following two steps. In this paper, assuming collaboration between two organizations, Organization A holds features and Organization B holds teacher labels.

(1) Secure cross-tabulation table creation across organizations using secure cross-statistics technology. (3.1 section)
(2) Learning the Naive Bayes classifier from the cross-tabulation table. (3.2 section)

3.1 Private Cross-Tabulation Table Creation Across Organizations Using PCT

As mentioned in Section 2, PCT can output securely created cross-tabulation tables across organizations. In utilizing PCT for federated learning, it is noted that the input values are limited to discrete values because it outputs a cross-tabulation table. If the input data is continuous, it needs to be discretized based on domain knowledge. For example, when the data is age, it can be cut in 10-year bins, or if it is assumed that whether or not one is an adult is relevant to solve a problem, it can be represented as binary as whether one is over the age of adulthood (such as 18 years old) or older or not. The output of this step is a cross-tabulation table with the classification of supervision labels on the side and each feature on the head of the table, and noise based on DP is added to each cell of the cross-tabulation table. Let c denote the supervision label (class), w_i denote i-th feature, and w_i has $k^{(i)}$ types of categorical values. Then, the output cross-tabulation table is denoted as $T_{i,k^{(i)},c}$. Figure 2 shows an example of the cross-tabulation table output by PCT. For example, for the $i = 1$ (Age), assuming $k = 1$ for Adult and $k = 2$ for Child, and the supervision label is 1 ($c = 1$), the corresponding value of the cross-tabulation table is $T_{1,1,1} = 2$.

Org. A

ID	Feat. 1 (Age)	Feat. 2 (Sex)	Feat. 3 (Past Purchase)
1	Adult	Female	Yes
2	Adult	Male	No
3	Child	Male	Yes

Org. B

ID	Label
1	0
2	1
3	1

Create private cross-tabulation table using PCT

	Feat. 1		Feat. 2		Feat. 3	
	Adult	Child	Male	Female	Yes	No
Label 0	1	2	2	0	0	1
Label 1	2	0	2	1	2	2

Output: cross-tabulation table T (output of PCT which is applied DP)

Figure 2: Example of a cross-tabulation table output by Secure Cross-Statistics Technology

3.2 Training Naive Bayes Classifier from Cross-Tabulation Table

In this section, we first explain about the Naive Bayes classifier [12]. Afterwards, we explain how to train the Naive Bayes classifier from the cross-tabulation table output by PCT and predict for new data.

3.2.1 Naive Bayes Classifier [12]. We define a data set D as $D = \{(d^{(1)}, c^{(1)}), (d^{(2)}, c^{(2)}) \cdots, (d^{(N)}, c^{(N)})\}$. Here, $d^{(n)}$ is each instance, $c^{(n)}$ is the supervision label (class) for each instance, and the number of instances is $|D| = N$. Naive Bayes classifier predicts the class that has the highest posterior probability $P(c|d)$ when an instance $d^{(n)}$ is given. From Bayes' theorem,

$$P(c|d) = \frac{P(c)P(d|c)}{P(d)} \tag{1}$$

is obtained. The denominator $P(d)$ on the right-hand side does not depend on the class, thus the c_{max} that maximizes the numerator is the prediction result.

$$c_{max} = \underset{c}{\operatorname{argmax}}\, P(c)P(d|c) \tag{2}$$

$$= \underset{c}{\operatorname{argmax}}\, P(c)P(w_1, w_2, \cdots, w_m|c) \tag{3}$$

Here, it is assumed that the instance $d^{(n)}$ is composed of m types of features w. In Naive Bayes classifier, it is assumed that features occur independently of the class. Then, the following equation is obtained.

$$P(w_1, w_2, \cdots, w_m|c) = \prod_{i=1}^{m} P(w_i|c) \tag{4}$$

From equations (3) and (4),

$$c_{max} = \underset{c}{\operatorname{argmax}} P(c) \prod_{i=1}^{m} P(w_i|c) \quad (5)$$

In multi-variate Bernoulli model, the parameters of Naive Bayes model can be obtained by maximum likelihood estimation as follows.

$$P(c) = \frac{N_c}{\sum_c N_c} \quad (6)$$

$$P(w_i|c) = \frac{N_{w_i,c}}{N_c} \quad (7)$$

Here, N_c is the number of instances in class c, and $N_{w_i,c}$ is the number of instances that include feature w among the instances belonging to class c.

3.2.2 Training and Inference Using Cross-Tabulation Tables. As mentioned in the previous section, we have to obtain the two parameters, the number of instances in each class N_c and the number of instances which contain a feature in each class $N_{w_i,c}$. First, N_c can be calculated by marginalizing in the class direction with the feature in focus as follows.

$$P(c) = \frac{N_c}{\sum_c N_c} = \frac{\sum_k T_{*,k,c}}{\sum_c \sum_k T_{*,k,c}} \quad (8)$$

where $*$ indicates any of the feature values. Next, it is sufficient to refer to $T_{w_{(i,k)},c}$ of the cross-tabulation table to calculate $P(w_i|c)$ as follows.

$$P(w_i|c) = \frac{N_{w_i,c}}{N_c} = \frac{T_{i,k,c}}{\sum_k T_{i,k,c}}. \quad (9)$$

Therefore, we can train a Naive Bayes classifier from the cross-tabulation table output by PCT. When inferring, we create features in Organization A to calculate $P(c)P(d|c)$ for each class, and can predict by finding the class where the posterior probability is maximum.

4 EVALUATION EXPERIMENT

In this section, we explain the evaluation conducted to confirm the effectiveness of the proposed method.

4.1 Evaluation Task

We evaluate the classification performance by classifying annual income based on user attribute information, which is used in existing studies on FL [7, 13]. In this task, we use the Adult dataset of the US Census data that is publicly available in the UCI repository[1] to perform binary classification on whether a user's annual income exceeds $50,000 based on the user's age, gender, education level, etc. If the annual income exceeds $50,000, the label is 1, otherwise 0, and the ratio of labels is 0 : 1 = 76.4% : 23.6%.

4.2 Evaluation Settings

4.2.1 Details of Comparison Method and Implementation. Since most existing studies on FL target deep learning models, we use deep learning models as our baseline models. Additionally, because central DP generally has more utility than local DP, if it is more accurate than central DP, it is more accurate than LDP. Thus, we use DP-SGD [1] in central DP as the baseline method which is a method to apply DP to deep learning models. DP-SGD applies noise based on Gaussian mechanism to gradient in the training phase. For the model in DP-SGD, we use a multilayer perceptron (MLP) with three layers. For the features, the categorical features and numerical features were pre-processed using one-hot encoding and min-max normalization, respectively. In deep learning models, data is used not only for learning the model, but also for tuning hyper-parameters. Therefore, in our evaluation, we conducted experiments under the condition that half of the given privacy budget is used for hyper-parameter tuning and the remaining half is used for model training. The model training uses a method of adding noise to the gradient of the stochastic gradient descent (SGD) proposed in DP-SGD. Hyper-parameters include the learning rate, the dimension of the hidden layer of MLP, and the batch size. Table 1 shows the target hyper-parameters and target ranges for tuning. We used Optuna [2] to efficiently carry out parameter tuning. Because both the proposed method and DP-SGD have random variable in each method, we evaluated the performance three times by changing the random seed and calculated averages for each method.

Table 1: The target and ranges of hyper-parameters in DP-SGD

Parameter	Range
Learning rate	1e-3, 5e-3, 1e-2, 5e-2, 1e-1, 5e-1, 1
Dim. of the hidden layer	8, 16, 32, 64
Batch size	16, 32, 64, 128, 256

4.3 Evaluation Results and Discussion

We set the privacy budget ϵ to 0.01, 0.05, 0.1, 0.5, 1, 2, 5, 10, and applied noise to the data based on the Gaussian mechanism for the baseline method (DP-SGD) and the Laplace mechanism for the proposed method. The accuracy for each privacy budget obtained from the evaluation is shown in Figure 3 The dashed lines in the figure indicate the accuracy of each method with no addtiion of noise. Figure 3 shows that the proposed method has a higher accuracy when ϵ is 1 or lower, while the baseline method has a higher accuracy when ϵ is 2 or higher. In addition, our results show that while the proposed method is able to train the model at any privacy budget despite a decrease in accuracy due to the influence of noise, this is not the case with DP-SGD, which failed to train the model when ϵ was smaller than 0.5 (i.e., 0.01, 0.05, 0.1). These results confirm the effectiveness of the proposed method.

There is no clear consensus, either theoretically or practically, on the optimal value of the privacy budget, and it is generally set to around 0.1 [11]. For example, in [6], it is mentioned that $\epsilon \geq 3$ is weak privacy protection and $\epsilon \leq 0.1$ is strong privacy protection. It should be noted that there is little decrease in the accuracy up to $\epsilon = 0.05$ in the proposed method. In light of the descriptions in [6], it can be considered that the proposed method can maintain utility better than the baseline method under strong privacy protection. However, in this paper, our evaluation is limited to one dataset,

[1] https://archive.ics.uci.edu/dataset/2/adult

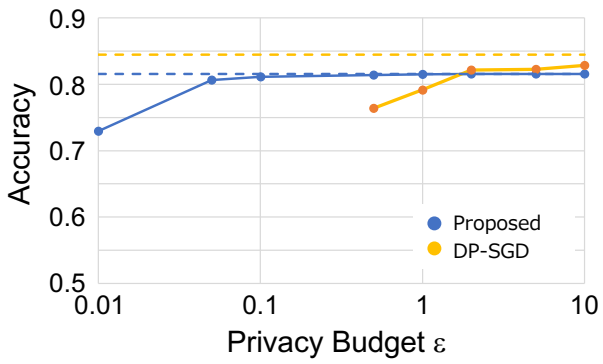


Figure 3: Evaluation Results

thus we need to perform further evaluation with various datasets in the future and verify the generality of the proposed method.

5 CONCLUSION

In this paper, we have proposed an approach for vertical FL that reduces the impact of noise by leveraging private cross-organizational data collaboration, called Private Cross-aggregation Technology (PCT). PCT integrates Private Set Intersection Cardinality (PSI-CA), Trusted Execution Environment (TEE), and Differential Privacy (DP) to generate a secure cross-tabulation table while preserving privacy from input to output. The method involves two main steps: securely creating cross-tabulation tables across organizations using PCT, and training a machine learning model (ML) using this private cross-tabulation data. We have demonstrated that the proposed approach achieves higher classification accuracy, especially for small private budgets which indicates strong privacy protection, in comparison with DP-SGD, a representative deep learning method incorporating DP. In the future, we would like to conduct further evaluation with more datasets and verify the generality of the usefulness of the proposed method. We believe that this study opens up a novel federated learning approach for vertical setting.

REFERENCES

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*. 308–318.

[2] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 2623–2631.

[3] Robin C Geyer, Tassilo Klein, and Moin Nabi. 2017. Differentially private federated learning: A client level perspective. *arXiv preprint arXiv:1712.07557* (2017).

[4] Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. 2022. Why do tree-based models still outperform deep learning on typical tabular data? *Advances in neural information processing systems* 35 (2022), 507–520.

[5] Keita Hasegawa, Takuya Chida, Keiichi Ochiai, Tomohiro Nakagawa, and Tetsuya Okuda. 2023. Guaranteeing Integrity in Private Cross-aggregation Technology. *NTT DOCOMO Technical Journal* 25, 1 (2023), 7.

[6] Xueyang Hu, Mingxuan Yuan, Jianguo Yao, Yu Deng, Lei Chen, Qiang Yang, Haibing Guan, and Jia Zeng. 2015. Differential privacy in telco big data platform. *Proceedings of the VLDB Endowment* 8, 12 (2015), 1692–1703.

[7] Yoshitaka Inoue, Hiroki Moriya, Qiong Zhang, and Kris Skrinak. 2023. SparseVFL: Communication-Efficient Vertical Federated Learning Based on Sparsification of Embeddings and Gradients. In *International Workshop on Federated Learning for Distributed Data Mining*.

[8] Tanzir Ul Islam, Noman Mohammed, and Dima Alhadidi. 2022. Private federated framework for health data. In *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 1919–1926.

[9] Kazuki Iwahana, Naoto Yanai, Jason Paul Cruz, and Toru Fujiwara. 2022. SPGC: integration of secure multiparty computation and differential privacy for gradient computation on collaborative learning. *Journal of Information Processing* 30 (2022), 209–225.

[10] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. 2021. Advances and open problems in federated learning. *Foundations and trends in machine learning* 14, 1–2 (2021), 1–210.

[11] Kazutoshi Kan. 2023. Seeking the ideal privacy protection: Strengths and limitations of differential privacy. *Monetary and Economic Studies* 41 (2023), 49–80.

[12] Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval*. Cambridge university press.

[13] Thomas Marchioro, Lodovico Giaretta, Evangelos Markatos, and Sarunas Girdzijauskas. 2022. Federated naive bayes under differential privacy. In *19th International Conference on Security and Cryptography (SECRYPT), JUL 11-13, 2022, Lisbon, Portugal*. Scitepress, 170–180.

[14] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*. PMLR, 1273–1282.

[15] Kazuma Nozawa, Keita Hasegawa, Keiichi Ochiai, Tomohiro Nakagawa, Kazuya Sasaki, Masayuki Terada, Masanobu Kii, Atsunori Ichikawa, and Toshiyuki Miyazawa. 2023. Technique for Achieving Privacy and Security in Cross-company Statistical Data Usage. *NTT DOCOMO Technical Journal* 25, 1 (2023), 7.

[16] Stacey Truex, Nathalie Baracaldo, Ali Anwar, Thomas Steinke, Heiko Ludwig, Rui Zhang, and Yi Zhou. 2019. A hybrid approach to privacy-preserving federated learning. In *Proceedings of the 12th ACM workshop on artificial intelligence and security*. 1–11.

[17] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. 2019. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)* 10, 2 (2019), 1–19.

[18] Ligeng Zhu, Zhijian Liu, and Song Han. 2019. Deep leakage from gradients. *Advances in neural information processing systems* 32 (2019).