future internet

IMPACT FACTOR 2.8

CITESCORE 7.1

Article

Differential Private Federated Learning in Geographically Distributed Public Administration Processes

Mirwais Ahmadzai and Giang Nguyen

Special Issue

Information and Future Internet Security, Trust and Privacy II

Edited by

Prof. Dr. Weizhi Meng and Dr. Christian D. Jensen

MDPI

https://doi.org/10.3390/fi16070220

future internet

MDPI

Article

Differential Private Federated Learning in Geographically Distributed Public Administration Processes

Mirwais Ahmadzai * and Giang Nguyen *

Faculty of Informatics and Information Technologies, Slovak University of Technology, Ilkovičova 2, 84216 Bratislava, Slovakia

* Correspondence: mirwais.ahmadzai@stuba.sk (M.A.); giang.nguyen@stuba.sk (G.N.)

Abstract: Public administration frequently deals with geographically scattered personal data between multiple government locations and organizations. As digital technologies advance, public administration is increasingly relying on collaborative intelligence while protecting individual privacy. In this context, federated learning has become known as a potential technique to train machine learning models on private and distributed data while maintaining data privacy. This work looks at the trade-off between privacy assurances and vulnerability to membership inference attacks in differential private federated learning in the context of public administration applications. Real-world data from collaborating organizations, concretely, the payroll data from the Ministry of Education and the public opinion survey data from Asia Foundation in Afghanistan, were used to evaluate the effectiveness of noise injection, a typical defense strategy against membership inference attacks, at different noise levels. The investigation focused on the impact of noise on model performance and selected privacy metrics applicable to public administration data. The findings highlight the importance of a balanced compromise between data privacy and model utility because excessive noise can reduce the accuracy of the model. They also highlight the need for careful consideration of noise levels in differential private federated learning for public administration tasks to provide a well-calibrated balance between data privacy and model utility, contributing toward transparent government practices.

Keywords: public administration; federated learning; differential privacy; membership inference attacks

Citation: Ahmadzai, M.; Nguyen, G. Differential Private Federated Learning in Geographically Distributed Public Administration Processes. *Future Internet* **2024**, *16*, 220. https://doi.org/10.3390/fi16070220

Academic Editors: Weizhi Meng and Christian D. Jensen

Received: 7 April 2024
Revised: 12 June 2024
Accepted: 21 June 2024
Published: 23 June 2024

1. Introduction

Public administrations (PAs) play an important role in modern societies, processing huge amounts of personal data related to governance, public services, and citizen interactions. The growing reliance on digital technologies has dramatically increased the volume and complexity of data created and managed by public authorities [1]. These data sets encapsulate essential information for policy formulation, resource allocation, and overall decision-making processes, shaping the direction of public policies and services and highlighting the critical role of data in effective governance [2]. Furthermore, as digital technologies advance, PAs are increasingly relying on collaborative initiatives to harness collective intelligence while protecting individual privacy [3,4]. The integration of modern technologies into PA practices reflects and highlights data-driven governance solutions, as well as the need to find a balance between innovation and privacy [5].

In this context of PA requirements, federated learning (FL) is a potential solution to collaborative model training that does not share raw data, allowing organizations to secure sensitive information while benefiting from collective intelligence. FL presents a viable solution to address privacy concerns in the context of data sharing between organizations [6]. While ensuring privacy in FL for PA is critical, especially when working with real-world sensitive data sets such as public opinions and financial records, differential privacy (DP) emerges as an essential and advanced privacy-preserving mechanism to achieve this goal.

DP purposefully injects controlled noise into model updates, creating a layer of ambiguity that discourages attempts to identify individual contributors. This proactive technique greatly improves the entire privacy protection process in FL systems [7,8].

The combination of FL and DP, called differential private federated learning (DPFL), is an effective way to train machine learning (ML) models that use geographically distributed data (such as PA data) while maintaining data privacy. DPFL is a promising approach, but accurately determining the amount of privacy achieved remains a challenge. This shortcoming prevents businesses from successfully evaluating and ensuring the efficacy of privacy-preserving measures in FL implementations.

This study investigates and quantifies privacy in the context of FL, using data provided by collaborating entities to assess the level of privacy protection obtained in FL systems. An effort is made to provide valuable information on the DPFL process for organizations to support decision making. The insight is also obtained in light of membership inference attacks (MIAs), which assess the effectiveness of DPFL from various points of view for PA. The main contributions of our work are as follows.

- Insights for PA process modeling: This study investigates the possibilities of DPFL as a procedure for PA that protects citizens' privacy while enabling data-driven governance. DPFL enables PA to use collaborative data analysis to improve service delivery, make informed decisions, and increase efficiency.
- Reduced MIA vulnerability: This study investigates how the use of DPFL with noise injection considerably reduces the sensitivity of participant data to MIA in PA contexts. The constant decrease in MIA success rates with increasing noise levels demonstrates the potential of DPFL to improve data privacy.
- Real-world evaluation and competitive performance: The proposed DPFL technique undergoes evaluations in two real-world PA scenarios that use data from two public sectors. The results show that the strategy outperforms traditional ML techniques in both settings while maintaining the anonymity of the participants through DP.

The rest of this paper is structured as follows. Section 2 provides a background of the work, including an overview of DPFL in the PA context. Our proposed approaches for the modeling of PA processes by DPFL, as well as the FL architecture for differential privacy with quantification of privacy, are presented in Section 3. Section 4 presents experimental evaluations of the proposed approach carried out on real-world PA data. Future research directions are noted on the basis of identified limitations. Section 5 concludes with a look at the implications and potential applications of privacy quantification in the context of personal data exchange between organizations that enable effective collaboration while protecting data privacy.

2. Related Work

Public administration (PA) focuses on enforcing government regulations and providing public services, while public management emphasizes efficiency, effectiveness, and results-oriented approaches to governance and service delivery [9]. Three main obstacles to PA were identified, including subjectivity to value, which hinders the development of universally applicable principles due to various social objectives, disregard for human complexity, and various social settings [10]. These findings highlight the complexities of PA and the challenges of developing a unified and universally applicable set of rules while respecting privacy rights.

2.1. Federated Learning

Federated learning (FL) has recently received much interest due to its ability to protect privacy in diverse data ecosystems. There are also many challenges that must be addressed, particularly the challenge of improving privacy protection in non-independently and identically distributed (non-IID) data. Non-IID data refer to instances in which data distributions between participating locations or organizations differ, which means that is heterogeneous and biased input data, creating significant problems for FL algorithms [11]. The work

in [12] introduces a Paillier federated multi-layer perceptron (PFMLP) framework, which combines encryption and federated learning to provide privacy-preserving ML. Another study in [13] identifies numerous important issues, including privacy, communication, and distributed optimization in FL. The work in [14] demonstrates that adversarial inference attacks can be used to infer the properties of training data sets in collaborative learning situations. The work in [15] suggests a secure and efficient technique for collecting user data. These studies demonstrate that FL is a potential strategy for privacy-preserving ML (PPML), although several issues remain, such as robustness to adversarial attacks, especially in non-IID data, and scalability to large-scale data sets.

2.2. Machine Learning Privacy Attacks

Table 1 covers the different types of privacy attacks that can be performed against ML models, as well as the references that address these attacks.

Table 1. Machine learning privacy attacks.

No.	Study	Attacks	Category
1	[16] [17] [18] [19] [20] [21] [22]	Membership Inference Measuring Membership Privacy MIA LOGAN Data Provenance Privacy Risk in ML Fredrikson et al.	MIA: These methods are used by attackers to determine whether a data point was used to train an ML model. The attackers do not have direct access to ML model parameters but observe its output, and their intention is to access sensitive information of individuals.
2	[23] [24] [25] [26]	MIA w/ Confidence Values Evaluating model inversion attacks while protecting privacy Updates Leak Collaborative Inference MIA	Model Inversion Attacks: These methods are used first to understand the structure of the model and then to reconstruct the original data using optimization techniques on input data to produce the same output.
3	[27] [28] [29]	The Secret Sharer Property Inference on FCNNs Hacking Smart Machines	Property Inference Attacks: These attacks uncover sensitive properties of a model. They are not related to training tasks.
4	[30] [31]	Cache Telepathy Stealing hyperparameters	Inference attacks of parameters: steals model parameters.
5	[32]	Stealing ML Models	Hyperparameter Inference Attacks: steals the hyperparameters used to train the model.

The work [16], in particular, studies ways to reduce MIAs, with an emphasis on detecting records in the set of model training. The study by [17] investigates the evaluation of MIAs in ML models, which helps to improve the understanding of privacy issues [33]. Another study by [18] emphasizes the vulnerability of ML models to MIAs, and attempts to determine whether a data record was used to train a target model. The work in [19] explores MIAs against generative models, where the adversary aims to identify whether a specific data point was used to train the model. Another work by [20] addresses the effect of data provenance on membership inference, addressing privacy concerns related to the origin and history of data. Another study by [21] investigates larger privacy problems in ML models, with a particular emphasis on tackling MIAs. The groundwork for understanding MIAs and their consequences for privacy in ML is investigated by [22].

The first row in Table 1 investigates several studies on membership inference attacks, and is aimed at determining whether specific data points were used in model training.

The second row in Table 1 investigates various aspects of model inversion attacks, including their impact on privacy, methods to measure model inversion risks, and strategies to mitigate these attacks. The work in [23] focuses on the impact of confidence values on increasing the level of complexity of these attacks. The work in [24] develops an evaluation approach for DP learning against model inversion attacks in the context of neural network models. The work in [25] investigates updates leak attacks, highlighting the possibility of

information leakage during model updates and the implications for privacy. Collaborative inference attacks and proposed solutions to mitigate membership inference vulnerabilities in collaborative ML systems are addressed by [26].

The third row in Table 1 investigates several aspects of property inference attacks, such as their influence on privacy, methods to measure property inference risks, and mitigation solutions. Specifically, the work in [27] investigates property inference attacks, with an emphasis on the confidentiality of shared information in ML models. The work of [28] investigates property inference attacks on fully connected neural networks, emphasizing model property extraction. Emphasizing their potential exploitation and compromise of smart machine security, property inference attacks are investigated by [29].

The fourth row in Table 1 investigates various aspects of parameter inference attacks. The work in [30] investigates parameter inference attacks, focusing on cache telepathy as a method of retrieving critical model parameters. The work in [31] investigates attempts to steal hyperparameters, identifying flaws in model hyperparameter protection. The fifth row in Table 1 deals with hyperparameter inference attacks. The work in [32] focuses on attacks that steal whole ML models, including hyperparameters and parameters, exposing comprehensive model-stealing vulnerabilities.

2.3. Quantification of Privacy Loss

Privacy quantification is used to measure the level of privacy protection of a system or algorithm [34]. Privacy quantification involves a wide range of mathematical concepts and metrics that serve as the foundation for measuring and evaluating privacy. Mutual information, entropy, and (ε, δ)-DP are examples of these concepts [35–37]. Other entropy concepts, such as Rényi's min-entropy, are used in privacy quantification to represent different types of attacks and information leakage scenarios [38].

The most well-known formal definition of differential privacy (DP) is the Cynthia Dwork's formula [39], which is as in Equation (1).

A randomized algorithm $\mathcal{M}$ with domain $\mathbb{N}^{|\mathcal{X}|}$ is (ε, δ)-differentially private if $\forall S \subseteq \text{Range}(\mathcal{M})$ and $\forall x, y \in \mathbb{N}^{|\mathcal{X}|}$ such that $\|x - y\|_1 \leq 1$:

$$\Pr[\mathcal{M}(x) \in S] \leq \exp(\varepsilon) \Pr[\mathcal{M}(y) \in S] + \delta \tag{1}$$

where

$\mathcal{M}$ is a random algorithm (or also called a query mechanism);

S is the set of possible outcomes of $\mathcal{M}$;

Epsilon (ϵ) is called the privacy budget and presents the maximum distance between $\mathcal{M}(x)$ and $\mathcal{M}(y)$;

Delta (δ) is the probability that information is accidentally leaked.

If $\delta = 0$, we say that $\mathcal{M}$ is ε-differentially private or, in short, $(\varepsilon, 0)$-DP or ε-DP. Otherwise, we say that it is (ε, δ)-differentially private or, in short, (ε, δ)-DP.

$(\varepsilon, 0)$-DP controls the amount of privacy protection provided, while (ε, δ)-DP adds the second layer of privacy protection, which represents the maximum probability of privacy violation. The idea is that including or excluding an individual's data should not have a major impact on the outcome of a query or expose specific information about that individual. (ε, δ)-DP offers a quantitative measure of privacy guarantees, and denotes the level of privacy protection. (ε, δ)-DP guarantees that the information disclosed about individuals remains within acceptable limits by imposing specific constraints on data release techniques, such as adding properly adjusted noise to the query results.

Various studies have been conducted on the use of DP to protect privacy in FL systems. However, practical evaluation in the real world is still limited. The work in this study addresses the limitations of previous research by conducting a thorough experimental evaluation of DPFL on two real-world data sets from two different organizations in Afghanistan. We study the robustness of DPFL to adversarial attacks and evaluate its performance and implications for the privacy protection process.

3. Differential Private Federated Learning in Distributed Public Administration

In practice, the PA process is built as distributed. Its modeling is called distributed administration process modeling (DAPM). The distributed approach aims to overcome the limits of the centralized one by distributing ownership and responsibility for design and execution among stakeholders in the ecosystem. This approach aims to promote agility, transparency, and public participation by using collaborative platforms and distributed technologies to model and execute distributed processes. DAPM promotes greater openness, accountability, and security by allowing local governments, communities, and maybe even people to participate in processes relevant to their own needs. However, thorough consideration of difficulties such as governance frameworks, data privacy, data standardization, and interoperability is critical to successful large-scale adoption [40].

3.1. Differential Private Federated Learning Architecture

This section describes our DPFL design strategy and its implementation approach. The DPFL workflow is explained in Figure 1. Each participating region serves as a client, training a local model with a unique data set. This initial training period is followed by the following unique phases:

- Phase 1: FL Training (Baseline Model): Local models are trained on regional data to build a baseline for future improvement.
- Phase 2: Noise addition with DP: To preserve data privacy, properly calibrated noise is introduced to local model updates before they are transmitted to the central server.
- Phase 3: MIA Evaluation: The baseline and noisy models are compared to the simulated MIA attempts. This step assesses how effectively models avoid exposing whether individual data items contributed to their training.
- Phase 4: Secure Model Aggregation: All *noisy* model updates are then transmitted to a centralized server. These updates are combined to produce a global model. This model is ultimately improved by repeated iterations of local training and global aggregation while preserving data privacy.

$$f(w) = \frac{1}{K} \sum_{k=1}^{K} f_k(w_k) \tag{2}$$

where:

$f(w)$ is the objective function of FL, also know as the loss function;
K is the number of total clients in the FL system;
w_k indicates the weight of the model on each client;
f_k is the local objective function of the client.

Equation (2) represents FL's objective function [41]. It is used to collect local updates from individual clients during the training process. The intention is to minimize the sum of each client's local objective functions, resulting in global optimization while maintaining the data as distributed and private [42].

Figure 1 also shows the workflow of inference attacks in FL that extract information about member data from leaked data representations or from the global model [43].

Our work focuses on membership inference attacks (MIAs). The methodical execution of our research design is described in Figure 2. Pre-processing steps are performed to improve the quality of the data. After that, the setup of the federated environment is developed, and the data are distributed to the chosen clients, such as different regions. We develop the model's architecture and choose the training settings as the setup comes together, putting a particular focus on incorporating DP protections. A central server manages the secure transmission and aggregation of model changes throughout the training loop. In addition to increasing the accuracy of the model, these developments also check the level of privacy assurance shown by measures such as Renyi differential privacy (RDP) and the privacy parameters epsilon (ϵ) and delta (δ) (Section 3.2).

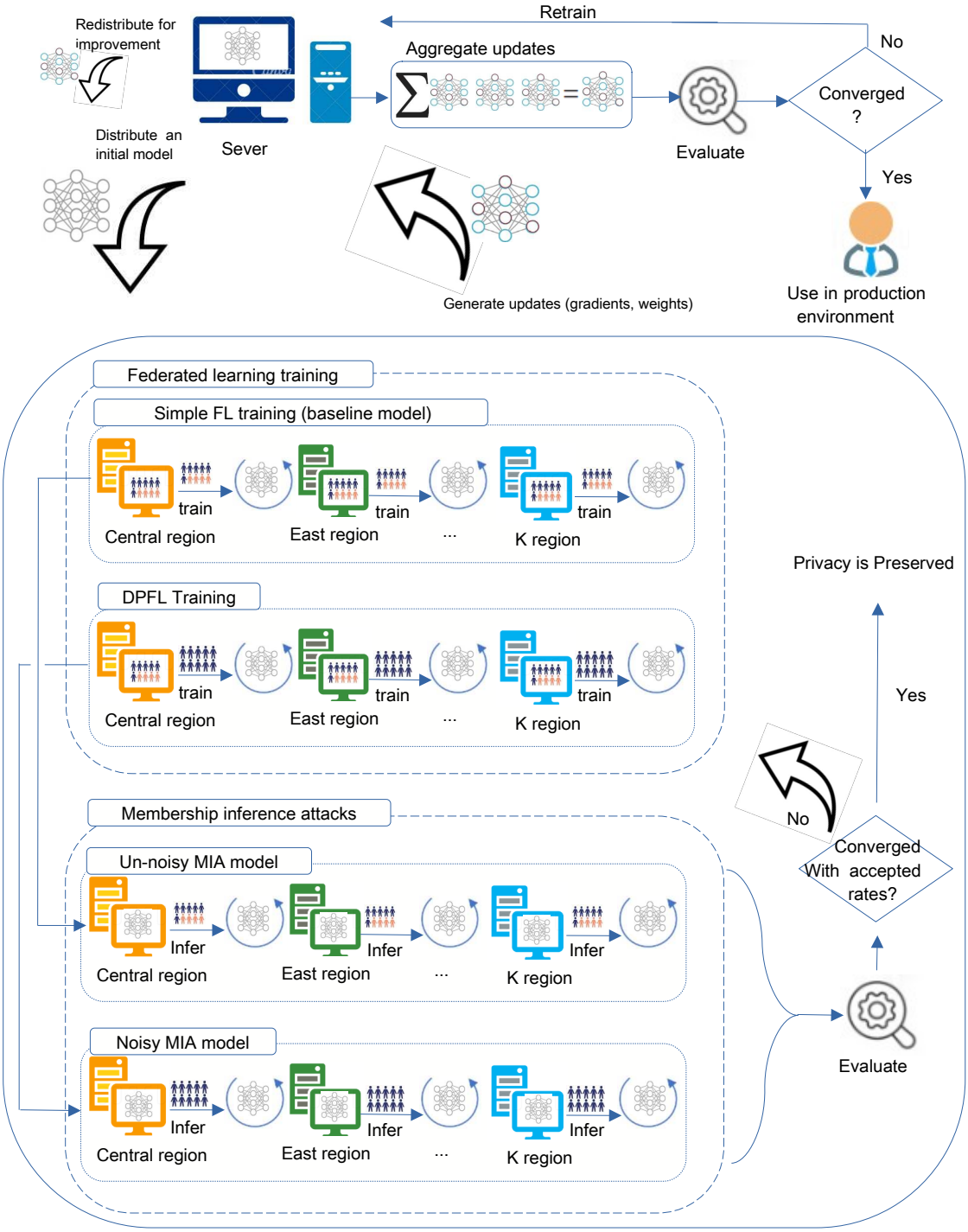

Figure 1. Differential private federated learning model workflow for distributed public administration process; the workflow also illustrates threats of the membership inference attack model.

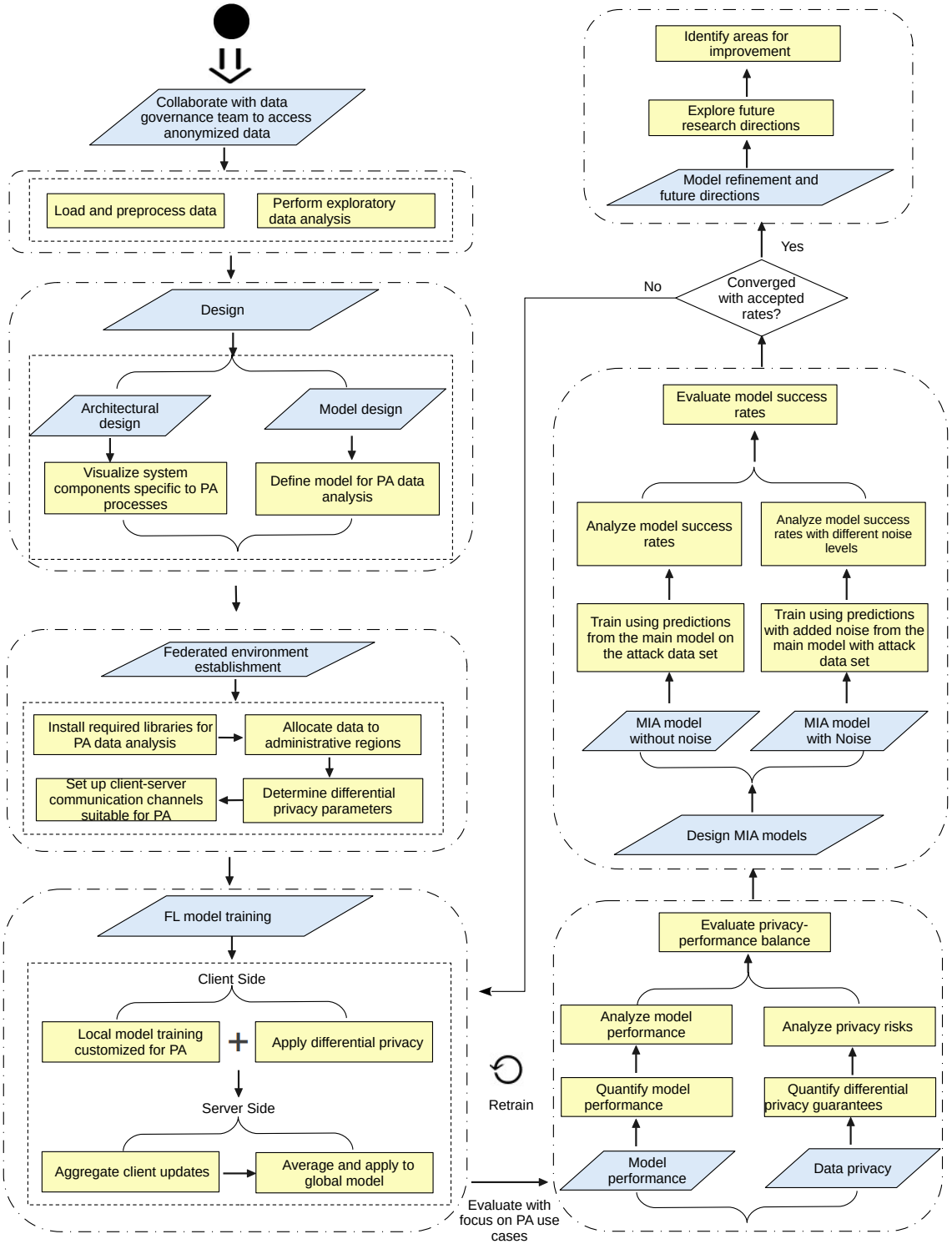

Figure 2. Federated learning workflow in distributed public administration.

In MIAs, an additional layer is added during the evaluation process. This involves assessing the vulnerability of the core model to privacy breaches. Two MIA models are examined: (1) MIAs without noise; (2) MIAs with added noise to simulate DP conditions.

These models are trained on the attack data set with predictions from the primary model. Evaluation criteria, such as accuracy and the F1 score, are used to assess performance (success rates) in the presence of various levels of noise. This research sheds light on

potential privacy attacks, directing changes to the primary model privacy procedures that improve overall resilience and security in real-world applications. The final phase of the diagram focuses on model refinement, which ensures continuous excellence in real-time application performance and privacy needs.

3.2. Privacy Quantification

Differential privacy concepts used in our proposed system enable a DPFL as follows.

The Gaussian mechanism [39] adds noise drawn from a Gaussian distribution whose variance is calibrated according to the parameters of sensitivity and privacy. For any ($\delta \in (0,1)$ and $\epsilon \in (0,1)$, the Gaussian mechanism $\mathcal{M}_{Gaussian}$ is defined in (Equation (3)) as follows:

$$\mathcal{M}_{Gaussian}(x, f, \epsilon, \delta) = f(x) + \mathcal{N}(\mu = 0, \sigma^2) \tag{3}$$

where

$$\sigma^2 = \frac{2\, log(1.25/\delta)}{\epsilon^2}(\Delta f)^2 \tag{4}$$

$\mathcal{N}$ stands for the normal distribution, μ is the mean, σ denotes the standard deviation of the distribution, and the logarithm is natural.

$\mathcal{M}_{Gaussian}$ only satisfies (ε, δ)-DP with $\epsilon < 1$. This statement is proven by the closeness of the $\mathcal{M}_{Gaussian}(x, f, \epsilon, \delta)$ and $\mathcal{M}_{Gaussian}(y, f, \epsilon, \delta)$ distributions with the probability of at least $(1 - \delta)$ in the appendix of the work [39].

The privacy parameter or privacy budget epsilon (ϵ) determines the amount of noise added. The parameter delta (δ) is the privacy parameter that controls the likelihood of information leakage [44].

We also denote that $\mathcal{N}(\mu = 0, \sigma^2)$ as $\mathcal{N}(0, \sigma^2)$, which is Gaussian noise added to the query response, such as in Equation (5).

$$Q(D) = f(D) + \mathcal{N}(0, \sigma^2) \tag{5}$$

The most challenging problem of the DP mechanism is that the privacy leakage increases due to composition. Determining a tighter bound of the privacy leakage due to composition allows for learning more features from a data set while protecting individual sensitive information [45]. This leads to the Rényi divergence definition (Equation (6)) of order $\alpha > 1$.

$$D_\alpha(d||d') = \frac{1}{\alpha - 1}\, log\, E_{x \sim d'}\left(\frac{d(x)}{d'(x)}\right)^\alpha \tag{6}$$

where $d(x)$ is the probability of seeing data point x in data set d and $d'(x)$ is the privacy-preserving probability of x in data set d'. The logarithm is natural and $x \sim d'$ means that x follows the distribution of d'.

Then, (α, ϵ)-RDP [45] is defined as in Equation (7) as a generalization of the notion of differential privacy based on the concept of the Rényi divergence definition (Equation (6)). It provides a quantitatively accurate way, with a tighter bound, to track cumulative privacy leakage under composition.

$$(\alpha, \epsilon)\text{-RDP} = D_\alpha(\mathcal{M}(D) \| \mathcal{M}(D')) \leq \epsilon \tag{7}$$

where $\mathcal{M}$ is a mechanism and D and D' are two neighboring data sets.

The statement that RDP provides guarantees for the composition of many steps of a private process is presented in [46]: a composition of a number of mechanisms m_i with each (α, ϵ_i)-RDP satisfies $(\alpha, \sum_i \epsilon_i)$-RDP and it is a tighter bound in comparison with the Gaussian mechanism for composition.

According to [47] and based on [46,48], the following result allows for converting from (α, ϵ)-RDP to (ϵ', δ)-DP: for any α, ϵ, and $\epsilon' > \epsilon$, (α, ϵ)-RDP implies (ϵ, δ)-DP, where

$$\delta = exp(-(\alpha - 1)(\epsilon' - \epsilon)) \tag{8}$$

Since this result holds for all orders of α, to obtain the best guarantees, the moment accountant needs to optimize over continuous $1 < \alpha < 32$ [46]. It is also shown that the use of only a restricted set of discrete α values is sufficient to preserve the tightness of privacy analysis. Practically, these bounds can be obtained by calculating RDP guarantees for various orders of α and converting them to (ϵ, δ)-DP guarantees. The best order that gives the lowest ϵ is chosen as in TensorFlow Privacy accountant implementation [49].

Our proposed approach described in the following parts is focused on evaluating the integration of DPFL and understanding the privacy protection level of the resulting model. The RDP of an FL system is determined by the privacy budget epsilon (ϵ), the privacy parameter delta (δ), and the number of clients (K) as a composition (see Equation (2)).

3.3. Federated Learning with Privacy Quantification

Algorithm 1 trains a global model using distributed private data from multiple locations (regions). Start with the global model parameters, local payroll or opinion data for each location, and DP settings. The global model is updated iteratively through communication between clients and a central server. Each location uses stochastic gradient descent to train a local model, which is then perturbed with Gaussian noise for privacy before being sent to the server. The central server collects these changes, adjusts the global model, evaluates its performance, and calculates privacy loss using the Gaussian RDP accountant. This cycle continues until convergence, which results in a globally trained model with DP guarantees.

Algorithm 1 DPFL with Quantifiable Privacy for Distributed Public Administration

Require:
Global model parameters θ
Local data $D_1, D_2, \ldots, D_K$
Differential privacy parameters: epsilon (ϵ), delta (δ)
Central server
Ensure: Differential private global model parameters θ^*
Initialize $\theta^* = \theta$
for $t = 1$ **to** T **do**
  for each client $k \in \{1, 2, \ldots, K\}$ **do**
    Train local model θ_k on D_k using global parameters θ and SGD optimizer
    Add Gaussian noise to local model updates: $\Delta\theta_k = \Delta\theta_k + \epsilon\mathcal{N}(0, \sigma^2)$ where σ is calculated according to Equation (4) with SGD as Δf
    Send local model updates $\Delta\theta_k$ for all k to the central server.
  end for
  At the central server:
  Combine model updates $\Delta\theta_k$ from all clients $k \in \{1, 2, ..., K\}$ to build the aggregated update $\Delta\Theta$.
  $\Delta\theta = \frac{1}{K}\sum_{k=1}^{K}\Delta\theta_k$
  Update the global model parameters $\theta^* = \theta^* + \Delta\theta$
  Calculate model performance metrics (as in Table 2) on the testing set
  The RDP accountant is used to convert the accumulated RDP privacy loss into (ϵ, δ)-DP guarantees at the end of the training.
end for
return DPFL global model parameters (θ^*)

Algorithm 2 addresses inference attacks in FL. Using a test data set called $\mathcal{T}$, the program compares the predictions of an FL model, f, with the actual labels in $\mathcal{T}$. Successful predictions and actual label matches imply that the corresponding data were used throughout the model training phase. The algorithm explores the vulnerabilities associated with data leaks, indicated by $\mathcal{D}$. The algorithm differentiates between data from the training set that have been leaked and data from outside the training set (MIA).

Although the technique offers a potentially valuable method for verifying who utilized whose data in the model's training, it also raises privacy concerns. If an attacker discovers which employee's data were used in model training, that person could be subject to vulnerabilities. In FL, there are several ways to prevent MIAs. Using DP measures to prevent MIAs is one method. DP is a technique that adds noise to data representations so that it is difficult to determine whether a given data point was used in model training [50,51].

Algorithm 2 MIA and Usage of Participant Data in Distributed Public Administration

Require:
Test set $\mathcal{T}$ with labeled data y_t,
indicating whether the participant was used to train the FL model.
Model of FL f for Payroll/Opinion data.
Representations of data leakage $\mathcal{D}$.
Ensure: Inference of the usage of sensitive information
(whether $\mathcal{T}$ data points were used in training f)
Initialize inference vector $\mathcal{I} = \mathbf{0}$ ▷ 0 indicates no participation
Initialize the inferred sensitive information $S = \emptyset$
for $t \in \mathcal{T}$ **do**
    Make a prediction for data point t using model f: $\hat{y} = f(t)$.
    if $\hat{y} \equiv y_t$ **then**
        $\mathcal{I}[t] = 1$ ▷ 1 indicates participation
    else
        $\mathcal{I}[t] = 0$
    end if
end for
if Membership inference attack **then**
    Train the model θ to distinguish between training and non-training
    data points in data set $\mathcal{D}$.
    Predict the membership status of each data point and update S accordingly.
end if
return the Inference vector $\mathcal{I}$ as well as inferred sensitive information S.

Table 2. Experiment setting.

Item	Description
Data Sets	Opinion data [52], payroll data [53].
Data Split	Training set: normal data set (70%), Testing set: normal data set (30%) Shadow data set: 50%, Shadow Train: 70%, Shadow Test: 30%
Federated Learning	
Framework	TensorFlow Federated (TFF) [54], Version 0.64.0
Model architecture	FFNN with 2 dense layers
Optimizer	SGD with default parameters
Model Aggregation	Federated averaging
Rounds	30
Clients per Round	Seven PA regions in the country
Evaluation Metrics	Accuracy, F1 score, RDP, epsilon (ϵ), delta (δ)
Differential Privacy	
Library	TensorFlow Privacy (TF-Privacy) [49], Version 0.8.12
Privacy Guarantee	Renyi-DP (RDP)
Privacy Budget	epsilon (ϵ) = 0.1 (fixed)
Privacy Parameter	delta (δ) = 10^{-5} (fixed)
Noise Mechanism	Gaussian mechanism
Noise Multipliers	2, 4, 6, 8, 10
Learning Rate	0.01

Table 2. *Cont.*

Item	Description
Membership Inference Attack (MIA)	
Attack Model	FFNN with 2 dense layers
Noise Factor	0.1
Shadow Data set	Used to create MIA models, split into shadow train and test
Training Set	Used for training models and evaluating the defense.
Testing Set	Used to evaluate attack success.
Metrics	Accuracy, F1 score

4. Experiments and Evaluation

4.1. Experiment Setting

Table 2 summarizes our study on DPFL to train ML models cooperatively while maintaining user privacy using two real-world data sets:

- Opinion Data: A survey of the Afghan People Opinion (2018) [52] was carried out by the Asia Foundation, which is an international development organization that has worked extensively in Afghanistan and focuses on issues such as leadership, justice administration, security, and economic growth. The survey aimed to determine how people felt about various aspects of the country's progress and governance.
- Payroll Data: The data set is provided by the Afghan Ministry of Education and provides the individual information of each employee in the educational institution payroll system of the provinces [53]. It includes unique sensitive identification, such as names, district and school, specific fields of study, and more. In particular, attributes also include personal and professional details such as gender, marital status, contract type, position, grade, and step, as well as financial information such as bank account numbers and salaries. The most important attribute in our analysis is *Attrition*, which indicates whether an employee has left the institution.

A separate FFNN with two dense layers as the attack model is also used to evaluate the resistance of our models against potential MIAs to validate the approach. The incorporation of a noise factor as a regularization parameter is used to improve the resistance against the mentioned attacks. Although the shadow data set was divided into shadow train and shadow test sets, the normal data set was used for both model training and defense mechanism evaluation. The latter set was then used to build shadow models, which were used in the attack simulation. The effectiveness of the attack and defense mechanisms was assessed using accuracy and F1 score as metrics.

Section 3.2 presents the key concepts of our methods and experiments to evaluate the trade-off between privacy and performance in DPFL. A Gaussian mechanism was used to inform noise calibration and assess the influence on accuracy and privacy. RDP was used for privacy quantification. This direct conformity with established principles ensures that our conclusions are methodologically solid and relevant.

4.2. Differential Private Federated Learning on the Public Administration Opinion Data

The opinion data [52] with all ethnic groups of Afghan residents were represented in the data collection, which was carried out in 34 provinces divided into seven different regions. Regional and gender categories were used to group the survey respondents. The gender split was nearly equal between the male and female respondents. About 75% of the respondents came from rural areas, and the remaining 25% came from urban areas. It is significant to note that some survey participants did not respond or were unable to do so for security, cultural sensitivity, or other reasons.

Results and Analysis

This experiment validates our approach described in Section 3 on the people's opinion data sets (Section 4.1). Figure 3 explores the relationship between noise levels, training

rounds, and the privacy parameter epsilon (ϵ). Observation demonstrates a stronger correlation between increased noise and lower levels of epsilon (ϵ), which implies a greater protection of privacy. This allows stakeholders to make informed decisions about their privacy preferences when analyzing public opinion data using FL. The image also introduces RDP, another metric for evaluating privacy protection, and shows that increasing noise levels correspond to a lower epsilon (ϵ) and a higher RDP, indicating better privacy protection. Figure 3 also provides insights to help practitioners navigate the privacy–performance trade-off and effectively use FL models when dealing with sensitive public opinion data.

Various levels of noise (multipliers ranging from 2 to 10) were added in the FL setting. The best choices for epsilon (ϵ) and RDP were evaluated as a privacy quantification measure to ensure effective privacy protection. For MIAs, since model performance is measured by how effectively an attack can distinguish between data points that were in the training data set from those that were not, the evolution of model performance during the iterative training process and the robustness of the algorithm to MIAs are presented in Figure 4.

The F1 score, which considers recall and precision in recognizing insights in public opinion, shows decreasing values with increasing noise. This highlights the importance of adjusting privacy settings so that they correspond to the allowable ranges of false positives and negatives in various analysis scenarios. Consider researching public trust in particular government agencies. It may be critical to reduce false negatives (missing instances of trust), even at the expense of some privacy protection due to lower noise (Table 3).

Table 3. DPFL model performance (accuracy, F1 score) and privacy quantification (RDP, epsilon (ϵ)) results (noise multipliers: 2–10).

Metric	Start/End Value	Opinion Data					Payroll Data				
		2.0	**4.0**	**6.0**	**8.0**	**10.0**	**2.0**	**4.0**	**6.0**	**8.0**	**10.0**
Accuracy	Start	0.95	0.94	0.94	0.94	0.96	0.98	0.89	0.90	0.89	0.89
	End	0.95	0.95	0.95	0.96	0.96	0.89	0.90	0.89	0.89	0.91
F1 Score	Start	0.94	0.93	0.93	0.93	0.92	0.99	0.93	0.89	0.86	0.85
	End	0.94	0.94	0.93	0.93	0.92	0.93	0.89	0.86	0.85	0.85
epsilon (ϵ)	Start	0.01	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00
	End	0.23	0.08	0.04	0.02	0.01	0.14	0.04	0.02	0.01	0.004
RDP	Start	45.29	62.86	63.00	100.14	118.71	52.57	63.00	90.86	128.00	128.00
	End	14.43	25.00	33.29	40.14	46.29	19.14	32.43	42.43	50.43	56.43

The accuracy of the model in evaluating public opinion decreases steadily as we add more noise, which is an essential barrier to personal privacy (Table 3). This emphasizes how important it is to precisely adjust the noise level to strike the right balance between protecting privacy and providing perceptive analysis. Reducing noise may be essential if our top priority is to comprehend complex public opinion regarding particular government policies, even at the expense of a minor privacy trade-off. In contrast, a certain amount of privacy compromise might be justified for more in-depth evaluations of public pleasure.

Figure 4 shows a clear consequence of the decrease in F1 scores with increasing noise levels. These findings imply that, in the context of DPFL, increasing noise levels can be an effective strategy for lowering MIA risks, hence strengthening the model's privacy guarantees.

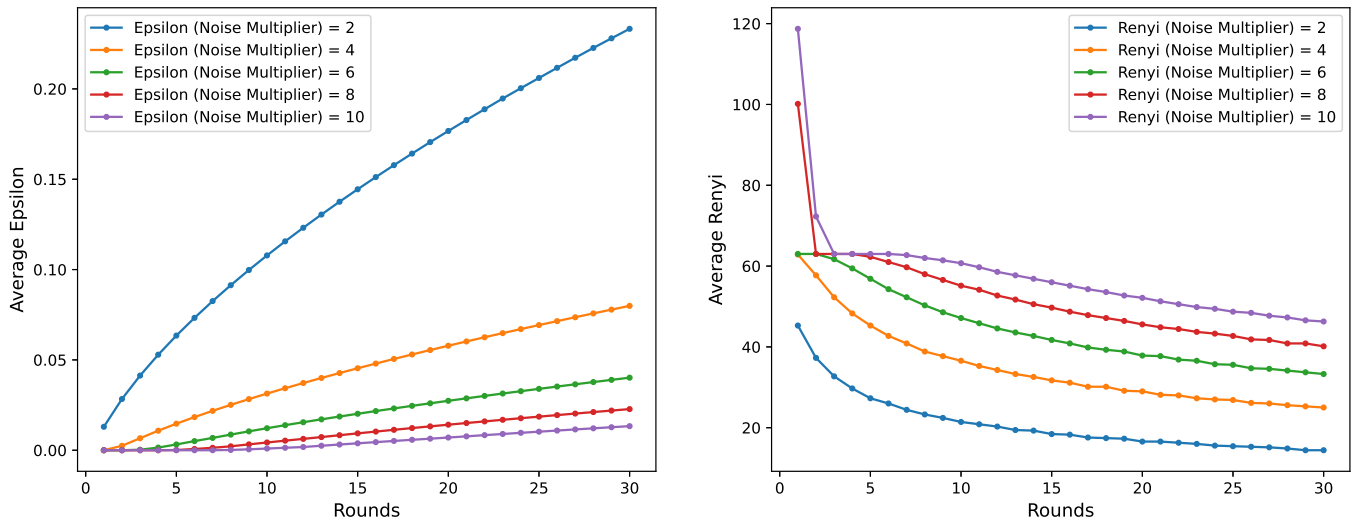

Figure 3. Opinion data: epsilon (ϵ) and Renyi differential privacy in relation with round number and noise multipliers.

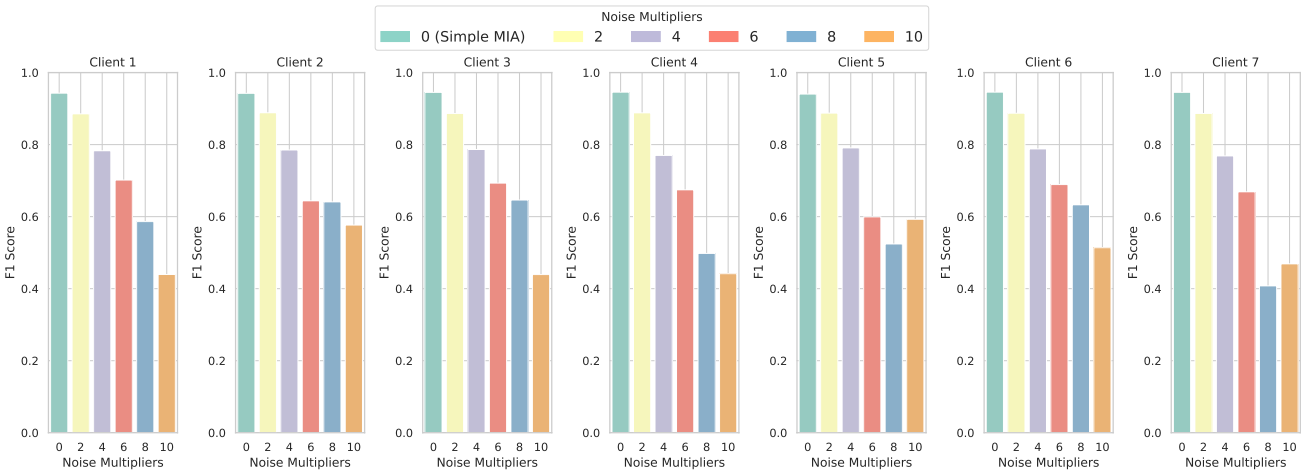

Figure 4. Opinion data: inference attack F1 score success rates of un-noisy and noisy model. Lower rates indicate stronger privacy against attacks. Colors/bars represent different noise levels.

However, introducing noise into noise multiplier 2 or greater results in a consistent decrease in accuracy and F1 scores for all clients, indicating stronger resilience against MIAs (Table 4). For example, in the opinion data set, in noise multiplier 2, the accuracy is reduced to approximately 91% and the F1 score falls to approximately 89%. The results show a progressive improvement in privacy protection for noise multipliers 4, 6, 8, and 10.

Table 4. Opinion data: un-noisy and noisy MIA models' performance (success rates).

Noise Multiplier	Client 1		Client 2		Client 3		Client 4		Client 5		Client 6		Client 7	
	Acc	F1 Score	Acc	F1 Score	Acc	F1 Score	Acc	F1 Score	Acc	F1 Score	Acc	F1 Score	Acc	F1 Score
0 (Un-noisy Model)	0.95	0.94	0.95	0.94	0.95	0.94	0.95	0.95	0.95	0.94	0.96	0.95	0.95	0.95
2	0.91	0.89	0.89	0.89	0.89	0.89	0.90	0.89	0.91	0.89	0.90	0.89	0.89	0.89
4	0.79	0.79	0.79	0.79	0.79	0.78	0.79	0.78	0.78	0.79	0.78	0.79	0.79	0.79
6	0.67	0.72	0.69	0.69	0.69	0.69	0.72	0.61	0.70	0.67	0.72	0.63	0.73	0.59
8	0.62	0.51	0.62	0.50	0.63	0.43	0.61	0.53	0.62	0.49	0.59	0.59	0.59	0.58
10	0.54	0.23	0.47	0.56	0.45	0.60	0.51	0.45	0.49	0.48	0.46	0.58	0.46	0.59

4.3. Differential Private Federated Learning on the Public Administration Payroll Data

Data for this experiment are from a consolidated payroll database system within the Afghan Ministry of Education [53]. This system has a hierarchical structure, with information collected from schools to provincial directorates and then to the central ministry.

This analysis uses the DPFL approach to investigate employee attrition in the education sector of the country. Payroll data comprise 34.7 MB of data and 394194 records, which provide a complete picture of employee data. Using FL approaches, the study aims to identify important factors (features) that influence employee decisions to leave their jobs. This approach enables collaborative analysis across organizations while protecting individual employee privacy, which is critical when working with sensitive data. The model uses binary classification algorithms on anonymized data spread across various organizations to reveal significant insights to establish successful retention strategies and improve staff well-being in the education sector.

Results and Analysis

The purpose of this experiment is to explore the effects of noise introduction on payroll data analysis in an FL context with payroll data. It evaluates the relationship between privacy and performance, as well as the effectiveness of applying DP to prevent MIAs.

Adding noise to payroll data for privacy protection has an influence on the model's ability to accurately interpret payroll features Table 3). This underscores the need to carefully balance privacy and utility according to the specific objectives of the analysis. However, unlike public opinion data, payroll data analysis frequently includes the detection of anomalies and outliers in addition to uncovering general insights. Table 3 also shows how adding noise affects the F1 score, which combines precision and recall. In the context of payroll data, eliminating false negatives (missing anomalies) can be crucial, even if it means sacrificing privacy protections. This is because unnoticed deviations can result in financial loss or security vulnerabilities.

Furthermore, Figure 5 explores the relationship between noise levels, training rounds, and the privacy parameter epsilon (ϵ). It reveals a stronger association between increased noise and lower levels of epsilon (ϵ), indicating stronger privacy guarantees. This empowers organizations to make informed decisions about their privacy goals when analyzing sensitive payroll data. The figure also introduces RDP, another crucial metric for quantifying privacy protection. Similarly to opinion data, increasing noise levels correspond to a lower epsilon (ϵ) and a higher RDP, signifying improved privacy protection. Organizations prioritizing privacy could choose a scenario with a lower epsilon (ϵ) and a higher RDP, but this could require accepting a compromise in the clarity or granularity of the insights obtained. This underscores the critical need to navigate the privacy–performance trade-off when working with sensitive data sets, such as payroll data.

The accuracy of the attack on Client 1 drops from 0.94 to 0.33 as the noise multiplier increases, and this applies to the other clients as well. This shows that the addition of noise restricts attacker attempts to identify participants by effectively reducing the model's ability to consistently determine whether a given data point belonged to the training set Table 5.

A consistent correlation can be observed in Figure 6 for every client, where the F1 score decreases as the noise multiplier increases. This implies that introducing noise effectively prevents adversaries from recognizing training data points, even while doing so lowers the model's overall performance, as measured by the F1 score, which strikes a compromise between precision and recall. This experiment also highlights the trade-off between privacy and performance that comes with employing noise to mitigate MIAs.

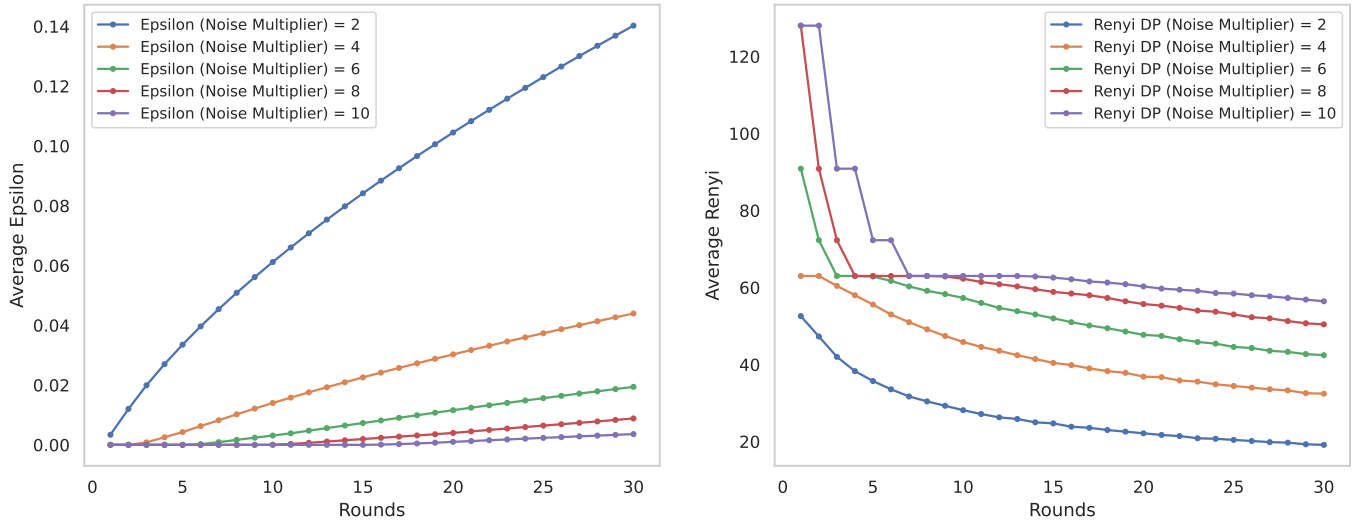

Figure 5. Payroll data: epsilon (ϵ) and Renyi differential privacy in relation with round number and noise multipliers.

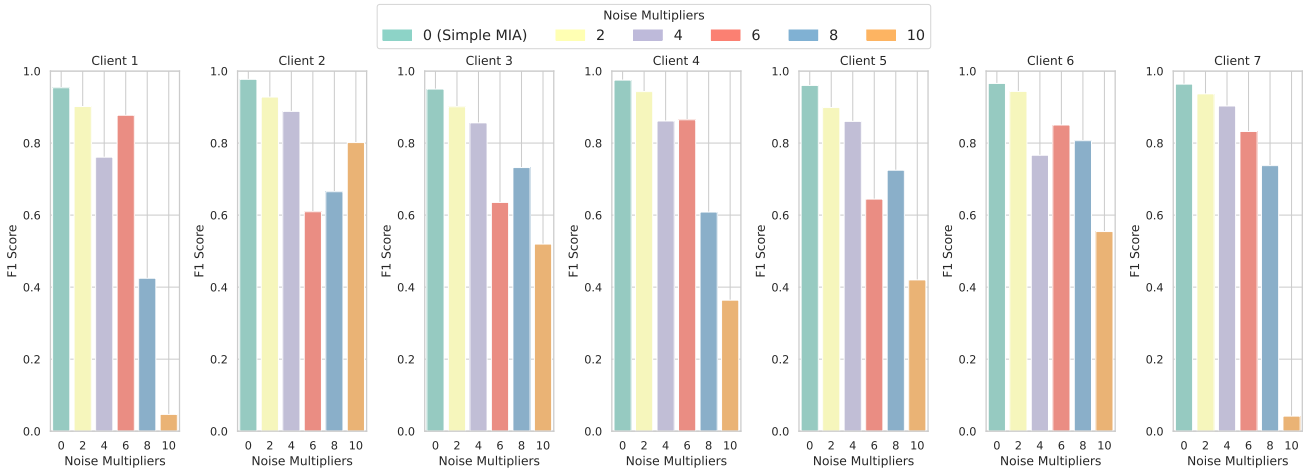

Figure 6. Payroll data: inference attack F1 score success rates of un-noisy and noisy model. Lower rates indicate stronger privacy against attacks. Colors/bars represent different noise levels.

4.4. Key Finding

In general, our findings help to build robust and privacy-preserving approaches for FL by shedding light on the effectiveness of noise in minimizing MIAs while acknowledging the inherent trade-off with model performance. This understanding enables practitioners to make informed decisions that emphasize data protection and meaningful insight, even in difficult situations, such as the analysis of sensitive payroll data.

Table 3 provides important information on how noise multipliers affect the performance of a DPFL model and privacy guarantees. The performance of both data sets was comparatively consistent at all noise levels. This implies that the model's capacity to learn well was not considerably impacted by the addition of noise for privacy protection.

- Opinion Data: At the end of the training, the accuracy and F1 score showed a slight improvement for several noise levels. This analysis indicates that the model's ability to detect minor details in the data can improve at certain noise levels. It could be very instructive to take a closer look at these noise levels and the kinds of particular insights that they might provide.

- Payroll Data: There were clear relationships in the F1 score for the payroll data. It dropped quickly at first, but as training continued, it started to gradually increase and slow down. This relationship raises two important ideas, such as the following:
  - Initial Difficulties: The model's initial inability to detect nuances in the payroll data may have been impeded by higher noise levels created by smaller multipliers, which is what caused the F1 score to drop quickly.
  - Improvement Potentials: The modest increase and gradual decrease observed in the F1 score show that the model is still learning despite noise. It is possible that the F1 score will also show a significant increase with more training iterations (beyond 30 rounds) while still offering respectable privacy guarantees. It is recommended to investigate how prolonged training affects the F1 score.

Table 3 also provides information on the relationship between noise levels and privacy guarantees in DPFL. We use RDP as a metric to quantify these guarantees, with higher values representing stronger privacy protection. Evaluating the RDP values between training rounds (1 to 30) for each noise level indicates a consistent behavior. RDP values frequently drop from the beginning (Start) to the end (End) of the training. This represents a gradual weakening of privacy assurances as the model learns from the injected noise. This observation is consistent with the inherent privacy–utility trade-off in DPFL, in which improved utility (model performance) often comes at the cost of decreased privacy.

Two important measures that quantify the trade-off between privacy and performance in DPFL for both opinion and payroll data are epsilon (ϵ) and RDP, as can be seen in Table 3. Epsilon (ϵ) shows a continuous sequence with smaller values, denoting higher privacy assurances. For both data sets, it begins at a low value and increases over training rounds. This is consistent with the intrinsic trade-off in DPFL: larger noise multipliers, which introduce less noise into the training set, provide better privacy at first (lower epsilon (ϵ)), but they may also result in more privacy loss over time (higher epsilon (ϵ)). On the other hand, weaker initial privacy (higher epsilon (ϵ)) is produced by smaller noise multipliers (which introduce more noise), but they may also result in a less gradual loss of privacy.

- Opinion Data: Epsilon (ϵ) often increased gradually during the training period, indicating the inherent trade-off between privacy and performance. However, the particular rate of increase could change depending on the chosen noise multiplier. Analyzing the connection between the rate of epsilon (ϵ) increase and noise multipliers can provide insight into how best to ensure privacy for these data.
- Payroll Data: During training, the epsilon (ϵ) often grew, much like the opinion data. However, more research is necessary because some initial epsilon (ϵ) values of 0.0000 are present. This could be the result of certain implementation specifics or the DP mechanism's restrictions. For an appropriate assessment of the real privacy guarantees in this circumstance, it is essential to understand the origin of such values.

However, when we compare RDP values across different noise multipliers, we obtain a more detailed picture. The starting RDP values increase as we proceed from left to right columns (lower to higher noise multipliers). This suggests that higher noise levels (right columns) often provide better initial privacy than lower noise levels (left columns). This is intuitive, as larger noise levels add more randomness into the training data, making it more difficult to differentiate between different data points, thus improving privacy protection from the start.

Tables 4 and 5 provide insight into the sensitivity of participant data to MIAs using a DPFL model trained on opinion and payroll data, respectively. Each value in the table reflects the MIA success rates (out of a total number of attempts) by which an attacker was able to accurately determine if a certain client's data were used in model training. Ideally, these rates should be as close to 0% as possible. A lower success rate for the MIA indicates a more effective defense against MIAs, which protects client privacy by making it impossible for attackers to identify training participants. In the absence of noise (noise multiplier 0), both data sets are susceptible to MIAs, with excellent accuracy and F1 scores

for all clients. In the public opinion data set, the accuracy is around 95%, and the F1 score is around 94%. Similarly, the accuracy of the payroll data is around 95%, and the F1 score is around 97%. These findings highlight the fundamental vulnerability of FL models to MIAs when privacy-preserving protections are not included.

Introducing noise into noise multiplier 2 or greater results in a consistent decrease in accuracy and F1 scores for all clients, indicating stronger resilience against MIAs. Similar results are found in the Opinion data (Table 4) and in the Payroll data (Table 5) where noise levels contribute to a decrease in accuracy and F1 scores. These findings underscore the critical importance of noise in reducing the risks of MIAs, as well as the importance of a sophisticated strategy to balance privacy and usefulness in FL systems.

Table 5. Payroll data: un-noisy and noisy MIA models' performance (success rates).

Noise Multiplier	Client 1		Client 2		Client 3		Client 4		Client 5		Client 6		Client 7	
	Acc	F1 Score	Acc	F1 Score	Acc	F1 Score	Acc	F1 Score	Acc	F1 Score	Acc	F1 Score	Acc	F1 Score
0 (Un-noisy Model)	0.91	0.97	0.94	0.95	0.95	0.96	0.97	0.98	0.95	0.96	0.94	0.95	0.94	0.95
2	0.91	0.93	0.91	0.94	0.92	0.94	0.90	0.92	0.87	0.90	0.92	0.94	0.89	0.91
4	0.81	0.86	0.80	0.84	0.85	0.89	0.84	0.89	0.79	0.83	0.77	0.81	0.83	0.88
6	0.63	0.64	0.64	0.66	0.68	0.73	0.75	0.82	0.67	0.71	0.68	0.72	0.64	0.66
8	0.60	0.67	0.51	0.49	0.69	0.79	0.50	0.45	0.54	0.55	0.48	0.39	0.49	0.44
10	0.39	0.30	0.44	0.44	0.45	0.46	0.45	0.47	0.34	0.08	0.37	0.21	0.66	0.79

The results of Table 3 demonstrate our key contribution: empirical validation on real-world data. We carefully selected two unique data sets: the public *Opinion Data*, which represents data from a medium-sized organization, and *Payroll Data*, which represents data from a large-sized organization. This representation of real-world data ensures that our models are applicable in a variety of real-world contexts, where data scale and privacy issues frequently diverge significantly. The results of Table 3 also implicitly support our contribution, which is the robustness of our models against adversarial attacks. Even at high noise levels, as the table illustrates, our models successfully achieve a compromise between privacy and performance in both data sets and they are also more resistant to adversarial attacks.

Incorporating privacy-related metrics into our analysis, such as epsilon (ϵ) and *RDP*, offers detailed insights into how our strategies affect privacy at different noise levels. The findings clarify the complex relationship between data value and privacy protection, highlighting the relevance of our research from a practical point of view.

Regarding limitations, this work examined the relationship between privacy and vulnerability to MIAs in DPFL for the modeling of PA processes. The work provides the effectiveness of noise injection against MIAs, and more steps are needed toward practical realization. This includes delving deeper into the privacy–performance trade-off across various settings and techniques, looking into additional defense mechanisms beyond noise injection to address broader privacy concerns such as secure communication and secure computation, and evaluating the finding applicability to different data and domains.

5. Conclusions

This study investigates the potential of DPFL for public administration services, focusing on both theoretical and practical aspects. DPFL provides a potential approach to data-driven governance, allowing PAs to make use of collaborative data analysis while protecting citizen privacy and encouraging informed decision-making processes. Its main contributions are in establishing the effectiveness of DPFL in reducing the vulnerability of participant data to membership inference attacks via noise injection. It helps to reinforce the practical implications of DPFL in public administration. Beyond privacy-enhancing characteristics, our work provides practical benefits to public administration, including potentially improved service delivery through better and deeper data-driven insights. It provides a better understanding of the technology behind the decision-making support for

stakeholders, where privacy protection is based on collaborative analysis and modeling, and the efficiency of the public administration process is increased through the use of distributed data resources while maintaining participant anonymity. In this way, citizen privacy is at the top priority.

Author Contributions: Conceptualization, G.N. and M.A.; data curation, M.A.; formal analysis, M.A.; funding acquisition, G.N.; investigation, G.N. and M.A.; methodology, G.N.; project administration, M.A.; resources, G.N.; software, M.A.; supervision, G.N.; validation, M.A.; visualization, M.A.; writing—original draft preparation, M.A. and G.N.; writing—review and editing, M.A. and G.N. All authors have read and agreed to the published version of the manuscript.

Funding: This publication has been written thanks to the support of the Operational Programme Integrated Infrastructure for the project: International Center of Excellence for Research on Intelligent and Secure Information and Communication Technologies and Systems Phase II (ITMS code: 313021W404), co-funded by the European Regional Development Fund (ERDF).

Data Availability Statement: Data and code are available at https://github.com/privacy-assurance/pa-dpfl (accessed on 30 April 2024). Anonymized and preprocessed data are available upon request for the private data set under a non-disclosure agreement (NDA).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Djeffal, C. Artificial intelligence and public governance: Normative guidelines for artificial intelligence in government and public administration. In *Regulating Artificial Intelligence*; Springer: Cham, Switzerland, 2020; pp. 277–293. [CrossRef]
2. Henman, P. Improving public services using artificial intelligence: Possibilities, pitfalls, governance. *Asia Pac. J. Public Adm.* **2020**, *42*, 209–221. [CrossRef]
3. Wirtz, B.W.; Weyerer, J.C.; Sturm, B.J. The dark sides of artificial intelligence: An integrated AI governance framework for public administration. *Int. J. Public Adm.* **2020**, *43*, 818–829. [CrossRef]
4. da Costa Alexandre, A.; Pereira, L.M. Ethics and development of advanced technology systems in public administration. In *Ethics and Responsible Research and Innovation in Practice: The ETHNA System Project*; Springer: Berlin/Heidelberg, Germany, 2023; pp. 224–247. [CrossRef]
5. Pandya, S.; Srivastava, G.; Jhaveri, R.; Babu, M.R.; Bhattacharya, S.; Maddikunta, P.K.R.; Mastorakis, S.; Piran, M.J.; Gadekallu, T.R. Federated learning for smart cities: A comprehensive survey. *Sustain. Energy Technol. Assessments* **2023**, *55*, 102987. [CrossRef]
6. Li, T.; Sahu, A.K.; Talwalkar, A.; Smith, V. Federated learning: Challenges, methods, and future directions. *IEEE Signal Process. Mag.* **2020**, *37*, 50–60. [CrossRef]
7. Xie, R.; Li, C.; Zhou, X.; Chen, H.; Dong, Z. Differentially Private Federated Learning for Multitask Objective Recognition. *IEEE Trans. Ind. Inform.* **2024**, *20*, 7269–7281. [CrossRef]
8. Zhou, C.; Yi, S.; Degang, W. Federated learning with Gaussian differential privacy. In Proceedings of the 2020 International Conference on Robotics, Intelligent Control and Artificial Intelligence, Shanghai, China, 17–19 October 2020; pp. 296–301. [CrossRef]
9. Lapuente, V.; Van de Walle, S. The effects of new public management on the quality of public services. *Governance* **2020**, *33*, 461–475. [CrossRef]
10. Csontos, B.; Heckl, I. Accessibility, usability, and security evaluation of Hungarian government websites. *Univers. Access Inf. Soc.* **2021**, *20*, 139–156. [CrossRef]
11. McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; y Arcas, B.A. Communication-efficient learning of deep networks from decentralized data. In Proceedings of the Artificial Intelligence and Statistics. PMLR, Fort Lauderdale, FL, USA, 20–22 April 2017; pp. 1273–1282.
12. Fang, H.; Qian, Q. Privacy preserving machine learning with homomorphic encryption and federated learning. *Future Internet* **2021**, *13*, 94. [CrossRef]
13. Kairouz, P.; McMahan, H.B.; Avent, B.; Bellet, A.; Bennis, M.; Bhagoji, A.N.; Bonawitz, K.; Charles, Z.; Cormode, G.; Cummings, R.; et al. Advances and open problems in federated learning. *Found. Trends® Mach. Learn.* **2021**, *14*, 1–210. [CrossRef]
14. Melis, L.; Song, C.; De Cristofaro, E.; Shmatikov, V. Exploiting unintended feature leakage in collaborative learning. In Proceedings of the 2019 IEEE Symposium on Security and Privacy, San Francisco, CA, USA, 19–23 May 2019; pp. 691–706. [CrossRef]
15. Bonawitz, K.; Ivanov, V.; Kreuter, B.; Marcedone, A.; McMahan, H.B.; Patel, S.; Ramage, D.; Segal, A.; Seth, K. Practical secure aggregation for privacy-preserving machine learning. In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, Dallas, TX, USA, 30 October–3 November 2017; pp. 1175–1191. [CrossRef]
16. Shokri, R.; Stronati, M.; Song, C.; Shmatikov, V. Membership inference attacks against machine learning models. In Proceedings of the 2017 IEEE Symposium on Security and Privacy, San Jose, CA, USA, 22–26 May 2017; pp. 3–18. [CrossRef]

17. Saeidian, S.; Cervia, G.; Oechtering, T.J.; Skoglund, M. Quantifying membership privacy via information leakage. *IEEE Trans. Inf. Forensics Secur.* **2021**, *16*, 3096–3108. [CrossRef]
18. Hu, H.; Salcic, Z.; Sun, L.; Dobbie, G.; Yu, P.S.; Zhang, X. Membership inference attacks on machine learning: A survey. *ACM Comput. Surv.* **2022**, *54*, 1–37. [CrossRef]
19. Hayes, J.; Melis, L.; Danezis, G.; De Cristofaro, E. Logan: Membership inference attacks against generative models. *arXiv* **2017**, arXiv:1705.07663.
20. Song, C.; Shmatikov, V. Auditing data provenance in text-generation models. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA, 4–8 August 2019; pp. 196–206. [CrossRef]
21. Yeom, S.; Giacomelli, I.; Fredrikson, M.; Jha, S. Privacy risk in machine learning: Analyzing the connection to overfitting. In Proceedings of the 2018 IEEE 31st Computer Security Foundations Symposium (CSF), Oxford, UK, 9–12 July 2018; pp. 268–282. [CrossRef]
22. Fredrikson, M.; Lantz, E.; Jha, S.; Lin, S.; Page, D.; Ristenpart, T. Privacy in pharmacogenetics: An End-to-End case study of personalized warfarin dosing. In Proceedings of the 23rd Usenix Security Symposium, San Diego, CA, USA, 20–22 August 2014; pp. 17–32.
23. Fredrikson, M.; Jha, S.; Ristenpart, T. Model inversion attacks that exploit confidence information and basic countermeasures. In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, Denver, CO, USA, 12–16 October 2015; pp. 1322–1333. [CrossRef]
24. Park, C.; Hong, D.; Seo, C. An attack-based evaluation method for differentially private learning against model inversion attack. *IEEE Access* **2019**, *7*, 124988–124999. [CrossRef]
25. Salem, A.; Bhattacharya, A.; Backes, M.; Fritz, M.; Zhang, Y. Updates-Leak: Data set inference and reconstruction attacks in online learning. In Proceedings of the 29th Usenix Security Symposium, Boston, MA, USA, 12–14 August 2020; pp. 1291–1308.
26. He, Z.; Zhang, T.; Lee, R.B. Model inversion attacks against collaborative inference. In Proceedings of the 35th Annual Computer Security Applications Conference, San Juan, PR, USA, 9–13 December 2019; pp. 148–162. [CrossRef]
27. Carlini, N.; Liu, C.; Erlingsson, Ú.; Kos, J.; Song, D. The secret sharer: Evaluating and testing unintended memorization in neural networks. In Proceedings of the 28th Usenix Security Symposium, Santa Clara, CA, USA, 14–16 August 2019; pp. 267–284. https://www.usenix.org/system/files/sec19-carlini.pdf
28. Ganju, K.; Wang, Q.; Yang, W.; Gunter, C.A.; Borisov, N. Property inference attacks on fully connected neural networks using permutation invariant representations. In Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, Toronto, ON, Canada, 15–19 October 2018; pp. 619–633. [CrossRef]
29. Ateniese, G.; Mancini, L.V.; Spognardi, A.; Villani, A.; Vitali, D.; Felici, G. Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers. *Int. J. Secur. Netw.* **2015**, *10*, 137–150. [CrossRef]
30. Yan, M.; Fletcher, C.W.; Torrellas, J. Cache telepathy: Leveraging shared resource attacks to learn DNN architectures. In Proceedings of the 29th Usenix Security Symposium, Boston, MA, USA, 12–14 August 2020; pp. 2003–2020.
31. Wang, B.; Gong, N.Z. Stealing hyperparameters in machine learning. In Proceedings of the 2018 IEEE Symposium on Security and Privacy, San Francisco, CA, USA, 20–24 May 2018; pp. 36–52. [CrossRef]
32. Tramèr, F.; Zhang, F.; Juels, A.; Reiter, M.K.; Ristenpart, T. Stealing machine learning models via prediction APIs. In Proceedings of the 25th Usenix Security Symposium, Austin, TX, USA, 10–12 August 2016; pp. 601–618.
33. Arca, S.; Hewett, R. Analytics on anonymity for privacy retention in smart health data. *Future Internet* **2021**, *13*, 274. [CrossRef]
34. Alvim, M.S.; Andrés, M.E.; Chatzikokolakis, K.; Degano, P.; Palamidessi, C. Differential privacy: On the trade-off between utility and information leakage. In *Formal Aspects of Security and Trust (FAST): 8th International Workshop, 2011*; Revised Selected Papers 8; Springer: Berlin/Heidelberg, Germany, 2012; pp. 39–54. [CrossRef]
35. Paninski, L. Estimation of entropy and mutual information. *Neural Comput.* **2003**, *15*, 1191–1253. [CrossRef]
36. Zhang, Z.; Lu, Z.; Tian, Y. Data Privacy Quantification and De-identification Model Based on Information Theory. In Proceedings of the 2019 International Conference on Networking and Network Applications, Daegu, Republic of Korea, 10–13 October 2019; pp. 213–222. [CrossRef]
37. Dwork, C. Differential privacy. In *International Colloquium on Automata, Languages, and Programming*; Springer: Berlin/Heidelberg, Germany, 2006; pp. 1–12. [CrossRef]
38. Cuff, P.; Yu, L. Differential privacy as a mutual information constraint. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, Austria, 24–28 October 2016; pp. 43–54. [CrossRef]
39. Dwork, C.; Roth, A. The algorithmic foundations of differential privacy. *Found. Trends® Theor. Comput. Sci.* **2014**, *9*, 211–407. [CrossRef]
40. Bennett, C.J.; Raab, C.D. Revisiting the governance of privacy: Contemporary policy instruments in global perspective. *Regul. Gov.* **2020**, *14*, 447–464. [CrossRef]
41. Xu, J.; Glicksberg, B.S.; Su, C.; Walker, P.; Bian, J.; Wang, F. Federated learning for healthcare informatics. *J. Healthc. Inform. Res.* **2021**, *5*, 1–19. [CrossRef]
42. Shi, Y.; Xu, X. Deep federated adaptation: An adaptative residential load forecasting approach with federated learning. *Sensors* **2022**, *22*, 3264. [CrossRef]

43. Nasr, M.; Shokri, R.; Houmansadr, A. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In Proceedings of the 2019 IEEE Symposium on Security and Privacy, San Francisco, CA, USA, 19–23 May 2019; pp. 739–753. [CrossRef]
44. Yu, R.; Yang, W.; Yang, C. Differentially Private XGBoost Algorithm for Traceability of Rice Varieties. *Appl. Sci.* **2022**, *12*, 11037. [CrossRef]
45. El Ouadrhiri, A.; Abdelhadi, A. Differential privacy for deep and federated learning: A survey. *IEEE Access* **2022**, *10*, 22359–22380. [CrossRef]
46. Mironov, I.; Talwar, K.; Zhang, L. Rényi Differential Privacy of the Sampled Gaussian Mechanism. *arXiv* **2019**, arXiv:1908.10530.
47. Ponomareva, N.; Hazimeh, H.; Kurakin, A.; Xu, Z.; Denison, C.; McMahan, H.B.; Vassilvitskii, S.; Chien, S.; Thakurta, A.G. How to dp-fy ml: A practical guide to machine learning with differential privacy. *J. Artif. Intell. Res.* **2023**, *77*, 1113–1201. [CrossRef]
48. Abadi, M.; Chu, A.; Goodfellow, I.; McMahan, H.B.; Mironov, I.; Talwar, K.; Zhang, L. Deep learning with differential privacy. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, New York, NY, USA, 24–28 October 2016; pp. 308–318. [CrossRef]
49. Google. TensorFlow Privacy | Responsible AI Toolkit. 2024. Available online: https://github.com/tensorflow/privacy (accessed on 12 June 2024).
50. Bernau, D.; Robl, J.; Grassal, P.W.; Schneider, S.; Kerschbaum, F. Comparing local and central differential privacy using membership inference attacks. In Proceedings of the IFIP Annual Conference on Data and Applications Security and Privacy, Calgary, AB, Canada, 19–20 July 2021; Springer: Cham, Switzerland, 2021; pp. 22–42. [CrossRef]
51. Ye, D.; Shen, S.; Zhu, T.; Liu, B.; Zhou, W. One parameter defense—Defending against data inference attacks via differential privacy. *IEEE Trans. Inf. Forensics Secur.* **2022**, *17*, 1466–1480. [CrossRef]
52. The Australian Data Archive (ADA), Asia Foundation Dataverse. Survey of the Afghan People (2004–2019), ADA Dataverse, V2. Available online: https://dataverse.ada.edu.au/dataset.xhtml?persistentId=doi:10.26193/VDDO0X (accessed on 30 April 2024).
53. MinistryEducation. Payroll Data Set, 2021. Private Personal Data in Afghanistan Education Sector, Available for the Research Purpose under Non-Disclosure Agreement (NDA). Available online: https://github.com/privacy-assurance/pa-dpfl (accessed on 30 April 2024).
54. Google. TensorFlow Federated (TFF): An Open-Source Framework for Machine Learning and Other Computations on Decentralized Data. 2024. Available online: https://www.tensorflow.org/federated (accessed on 12 June 2024).