Revamping Federated Learning Security from a Defender's Perspective: A Unified Defense with Homomorphic Encrypted Data Space

K Naveen Kumar
IIT Hyderabad, India
cs19m20p000001@iith.ac.in

Reshmi Mitra
SEMO, USA
rmitra@semo.edu

C Krishna Mohan
IIT Hyderabad, India
ckm@cse.iith.ac.in

Abstract

Federated Learning (FL) facilitates clients to collaborate on training a shared machine learning model without exposing individual private data. Nonetheless, FL remains susceptible to utility and privacy attacks, notably evasion data poisoning and model inversion attacks, compromising the system's efficiency and data privacy. Existing FL defenses are often specialized to a particular single attack, lacking generality and a comprehensive defender's perspective. To address these challenges, we introduce ***F****ederated* ***C****ryptography* ***D****efense (FCD), a unified single framework aligning with the defender's perspective. FCD employs row-wise transposition cipher based data encryption with a secret key to counter both evasion black-box data poisoning and model inversion attacks. The crux of FCD lies in transferring the entire learning process into an encrypted data space and using a novel distillation loss guided by the Kullback-Leibler (KL) divergence. This measure compares the probability distributions of the local pretrained teacher model's predictions on normal data and the local student model's predictions on the same data in FCD's encrypted form. By working within this encrypted space, FCD eliminates the need for decryption at the server, resulting in reduced computational complexity. We demonstrate the practical feasibility of FCD and apply it to defend against evasion utility attack on benchmark datasets (GTSRB, KBTS, CIFAR10, and EMNIST). We further extend FCD for defending against model inversion attack in split FL on the CIFAR100 dataset. Our experiments across the diverse attack and FL settings demonstrate practical feasibility and robustness against utility evasion (impact > 30) and privacy attacks (MSE > 73) compared to the second best method.*

1. Introduction

In recent years, the demand for collaborative learning algorithms that meet a comprehensive set of criteria has increased [24]. The criteria encompass key aspects such as data confidentiality, privacy, enhanced utility, robustness, and fairness [45]. Federated Learning (FL) [10, 12, 55] is a pioneering approach to collaborative machine learning (ML) that fulfils these vital requirements. The applicability of FL is evident in its wide-ranging applications, including mobile user personalization (Gboard) [1], healthcare [25], finance sector [17], among others.

Threats to FL. The decentralized nature of FL makes it highly susceptible to adversarial threats, primarily categorized as utility-centric and privacy-centric [48]. Utility-centric threats involve malicious attempts to poison data or models, thereby compromising accuracy [11, 40, 59]. Privacy-centric threats such as model inversion attacks (MIA) pertain to participant data reconstruction, risking the leakage of sensitive information and eroding trust in the FL system [20, 37, 51]. These threats are further classified based on their origin, distinguishing between causative (during training) and evasion (during testing) attacks [48]. Notably, evasion attacks manipulate the model's predictions by modifying the input test data during inference on deployed models, posing a severe threat to FL's utility compared to causative attacks [24, 48]. Effectively countering these threats is pivotal for enhancing FL's security and preserving the previously mentioned factors. Therefore, this paper focuses on mitigating the untargeted black-box evasion data poisoning attacks that are commonly encountered in real-world deployments [48]. Our overall aim is to implement a unified defense framework that can safeguard FL against both evasion utility and MIA privacy attacks during inference. This choice is motivated by a research gap to defend evasion attacks [24] [21, 50] and the limited work that tackles both attacks simultaneously.

Limitations of existing FL defense strategies. Table 1 illustrates the gaps within current FL defense approaches and underscores the vast scope of our work (please refer to the Supplementary material for a comprehensive related work discussion). Notably, existing research in FL adversarial defenses has predominantly concentrated on a particular single type, such as utility or privacy-centred attacks. They often neglect the defender's perspective, especially address-

Table 1. Comparison of existing defenses and their applicability for utility-centric attacks (U-CA) and privacy-centric attacks (P-CA) in FL. EA: evasion attack, DIm: defense impact, DB: defense budget, DV: defense visibility, src: source of the defense, S: server, and C: client. ● denotes strongly yes, ○ denotes strongly no.

FL defense category (src)	**FL defense methods**	**Can defend**			**U-P tradeoff analysis**	**Defender's perspective analysis**		
		U-CA	*EA*	P-CA		DIm	DB	DV
Adversarial training **(C)**	FAT [61], RS [8], FedDynAT [39], GALP [13]	◔	●	○	○	●	○	○
Byzantine robust aggregation techniques **(S)**	Krum [3], ShieldFL [30], FLTrust [5]	●	○	○	○	●	○	○
Data/ update analysis **(S and/or C)**	DeepSight [38], FL-Defender [18], SparseFed [36]	●	○	○	○	●	○	○
Secure multi-party computation **(C)**	AMPC [58], Byrd *et al.* [4]	○	○	●	◔	●	○	○
Trusted execution environments **(C)**	Flatee[35], Chen *et al.* [6], PPFL [34]	◔	◔	●	◔	●	○	○
Differential privacy **(C)**	NbAFL [52], Hu *et al.* [16], 2DP-FL [54]	○	○	●	◑	●	○	○
Homomorphic encryption **(C)**	DCAE [62], PEFL [29], Batchcrypt[57]	○	○	●	◑	●	○	○
	FCD (ours) src: S and C	●	●	●	●	●	●	●

ing both utility and privacy threats in a unified manner. This perspective is essential as it offers insights into the defense's robustness, adaptability, budget, multipurpose efficiency, visibility, and overall impact on the FL system. Further, within the domain of FL, adversaries can exhibit multifaceted intentions, spanning both utility and privacy threats. Deploying distinct defense mechanisms for each individual attack escalates the defender's budget, increases the overall computational complexity, and reduces the adaptability to practical FL trustworthy systems. Also, defenses that are easily accessible and transparent to attackers are susceptible to circumvention. While differential privacy (DP) [16, 52, 54], secure multi-party computation (MPC) [4, 58], and trusted execution environments (TEE) [6, 34, 35] have been proposed as potential solutions to strengthen privacy in FL, none of the existing solutions offers a satisfactory unification of privacy and utility defense under different attacks, coupled with reasonable efficiency [45].

In this work, we propose an approach called **F**ederated **C**ryptography **D**efense (FCD) to counter both utility and privacy evasion attacks through the strategic implementation of fully homomorphic encryption (FHE) in the data space via row-based cryptography transformation. Unlike existing defense methods (as shown in Table 1), which typically focus on defending against either utility or privacy threats, our defense framework aligns with the defender's perspective by working in the data space rather than gradient space. It stands out as a practical, cost-effective, and multipurpose defense mechanism. One key advantage is the preservation of data confidentiality, achieved through transferring the entire process on encrypted data using a shared secret key. Further, our approach enables server-side testing on encrypted data without the need for decryption, thus significantly reducing computational complexity. On the other hand, clients engage in training on encrypted data, where we introduce a novel distillation loss guided by the Kullback-Leibler (KL) divergence. This novel loss function minimizes the distillation loss between the pretrained teacher model and the local student model responses. In addition, the KL divergence loss, combined with the conventional cross-entropy loss, serves as a total loss for updating the local model. Furthermore, we have developed two distinct threat models, TM1 and TM2, addressing utility and privacy vulnerabilities in FL and expanding on the attacker's potential impact on the FL system. We present the first proposal for this type of unified defense in FL to the best of our knowledge. Our contributions are:

1. **Novel defense:** We introduce a unified defense framework called FCD, which uses row-based transposition cipher to defend against both evasion adversarial and MIA attacks at the server aligning to the defender's perspective.
2. **Loss function:** To enable effective learning, we propose a novel distillation loss guided by KL divergence between teacher and student model predictions on the client's side.
3. **Threat models:** We introduce two realistic threat models (TM1 and TM2) concerning to utility and privacy attacks.
4. **Theoretical analysis:** We present a possible theoretical analysis of our method *w.r.t.* convergence and resilience to attacks.
5. **Effective tradeoff analysis:** We perform extensive experiments with five different datasets and models, namely, GTSRB (CNN), KBTS (CNN), CIFAR10 (ResNet18), CIFAR100 (VGG11), and EMNIST (LeNet), alternating between threat models. We also investigate the performance under homogeneous and heterogeneous (non-IID) data shard settings.

2. Preliminaries

Definition 2.1 *(**FL data setup.**) We consider an FL system with a central server and n clients. Each client $\mathcal{C}_k$, where $k \in [1, n]$, possesses its private local dataset $\mathcal{D}_k$, referred to as a shard. In this context, we define $\mathcal{D}_k = \{\mathcal{X}_{i,k}, \mathcal{Y}_{i,k}\}_{i=1}^{\mathcal{N}_k} \subseteq \mathbb{R}^d \times \mathbb{R}$. Without loss of generality, we assume that $\|\mathcal{X}_{i,k}\|_2 = 1$ holds for all $k \in [1, n]$, where $i \in [1, \mathcal{N}_k]$ and the final component of each point is fixed at $1/2$, denoted as $\mathcal{X}^d = 1/2$ (l_2 norm data normalization). The server uses synchronous federated weighted averaging*

(FedAvg) [32] for aggregation, represented as:

$$\theta_g^{t+1} = \theta_g^t + \sum_{k=1}^{n} \lambda_k \nabla \theta_k^t \tag{1}$$

where $\lambda_k = \frac{\mathcal{N}_k}{\sum \mathcal{N}_k}$*, and* $\sum_k \lambda_k = 1$.

We investigate two FL data shard settings: *(i) Homogeneous*, where each client's dataset size is identical, i.e., $|\mathcal{D}_1| = |\mathcal{D}_2| = ..|\mathcal{D}_n| = \frac{|\mathcal{D}|}{n}$, and *(ii) Heterogeneous*, involving non-independent and non-identically (non-IID) distributed data achieved by partitioning the dataset using a Dirichlet distribution [33] with parameter $\beta = 1$ among clients. Additional information on standard FL, evasion attacks in FL, homomorphic encryption, and transposition cipher is provided in the supplementary material.

Threat model. We introduce two distinct threat models, TM1 and TM2, formulated to reflect real-world FL production deployment settings, as illustrated in Figure 1. Our threat models address an honest-but-curious (HbC) adversary at the central server, as inspired by related work [41, 48]. In TM1 **(evasion utility attack)**, the adversary conducts an indiscriminate evasion attack by manipulating test data at the central server during inference, aiming to misclassify a substantial portion of the inputs [24, 48]. In TM2 **(privacy attack)**, the adversary's goal shifts to performing a model inversion attack (MIA) [15] with the aim of reconstructing private data. We provide our practical assumptions and more details about the threat models in the Supplementary material.

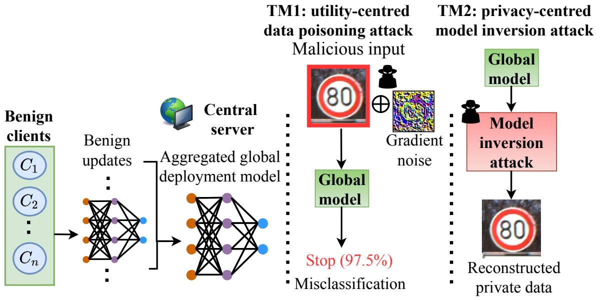

Figure 1. Overview of two different threat models (TM) with potential vulnerabilities and attacks during inference.

3. Proposed Framework

Problem statement. In each communication round t, suppose that the server receives model updates from m clients, and there exists a test dataset $\mathcal{D}_{test} = \{\mathcal{X}_i, \mathcal{Y}_i\}_{i=1}^{\mathcal{N}_{te}}$ with a size of $\mathcal{N}_{te}$. The adversary introduces a $\mu\rho$-bounded poisoned test dataset in TM1, denoted as $\tilde{\mathcal{D}}_{\text{test}} = \{\tilde{\mathcal{X}}_i, \mathcal{Y}_i\}_{i=1}^{\hat{\mathcal{N}}_{te}}$, with the dataset size $\hat{\mathcal{N}}_{te}$ varying based on the attack volume ρ as:

Definition 3.1 *(**μρ-bounded adversary.**) Let* $\mathcal{F}$ *be the model's function class. An adversarial perturbation is characterized by the mapping* $\mathcal{A} := \mathcal{F} \times \mathcal{X} \times \mathbb{R} \rightarrow \tilde{\mathcal{X}}$*. For* $\mu > 0$*, we define the* l_2 *norm ball as* $\mathcal{B}_2(\mathcal{X}, \mu) := \{\tilde{\mathcal{X}} \in \mathbb{R}^d : \|\tilde{\mathcal{X}} - \mathcal{X}\| \leq \mu\} \bigcap \mathcal{X}$*. We classify the adversary as* $\mu\rho$*-bounded if it adheres to the condition* $\mathcal{A}(\mathcal{F}, \mathcal{X}_i, \mathcal{Y}_i)_{i=1}^{\rho} \in \mathcal{B}(\mathcal{X}_i, \mu)_{i=1}^{\rho}$*. Furthermore, for a given* $\mu > 0, \rho > 0$*, we denote the worst-case adversarial perturbation as* $\mathcal{A}^+ := \arg\max_{\tilde{\mathcal{X}} \in \mathcal{B}(\mathcal{X}, \mu)} \mathcal{L}(f_\theta(\tilde{\mathcal{X}}_i, \mathcal{Y}_i)_{i=1}^{\rho})$*, where* $\mathcal{L}$ *represents the loss function of the global model.*

Next, we formulate the objective of the adversary in MIA attack (TM2) using the below definition.

Definition 3.2 *(**ϵ-distorted.**) Let* $\mathcal{X}^*$ *represent the data reconstructed by the adversary using the MIA attack. The objective of the adversary is to minimize the reconstruction error* P *in terms of as the mean squared error (MSE) between the reconstructed data* $\mathcal{X}^*$ *and the original input data* $\mathcal{X}$ *as*

$$P := \|\mathcal{X}^* - \mathcal{X}\|_2 \equiv \frac{1}{\mathcal{N}_{te}} \sum_{i=1}^{\mathcal{N}_{te}} (\mathcal{X}_i^* - \mathcal{X}_i)^2 \leq \epsilon, \textit{ for some } \epsilon \geq 0,$$

where ϵ *is the distortion of the reconstructed image.*

Futher, the global model, $\mathcal{G}_{\theta_g}$ attains a global test accuracy of $\mathcal{A}_g$ before the attack and $\mathcal{A}_g^*$ after the attack (Definition 3.1). Now, our FCD integrated clients have two primary objectives. First, they aim to minimize the impact of the attack, which is quantified by $U = (\mathcal{A}_g - \mathcal{A}_g^*)$, in order to preserve the global test accuracy. Second, they seek to enhance the privacy of local data $\mathcal{D}_k = \{\mathcal{X}_{i,k}, \mathcal{Y}_{i,k}\}_{i=1}^{\mathcal{N}_k}$ by maximizing the reconstruction error P. Here, U and P define the utility and privacy gains in the context of FL, where lower values of U and higher values of P are considered desirable.

FCD framework description. The FCD framework introduces modifications to standard FL, which can be categorized into server-side and client-side changes. Algorithm 1 and Figure 2 present our FCD integrated FL system.

Server side. To initiate the FL process, the central server initializes a key denoted as $\mathcal{K}$. This key consists of unique random integers in the range from 0 to $h-1$, where h represents the height of the input image data. The key $\mathcal{K}$ is defined as $\mathcal{K} = [\kappa_0, \kappa_1, \ldots, \kappa_i, \kappa_j, \ldots, \kappa_{h-1}]$, and it follows a property that if $i, j \in 0, \ldots, h-1$ with $i \neq j$, then $\kappa_i \neq \kappa_j$. This property guarantees that no row in the input image can be missed, ensuring the completeness of the data. During each communication round t, the central server follows the standard FL procedure by sending the current version of the global model parameters θ_g^t to all clients. Synchronous federated weighted averaging-based aggregation is employed at the server, as outlined in [32]. Further, the central server performs the FCD transformation on the evasion-attacked test data to obtain $\mathcal{E}(\hat{\mathcal{D}}_{test})$. Subsequently, the server tests the current aggregated global model $\mathcal{G}^t_\theta$ on the FCD transformed test data, i.e., $\mathcal{E}(\hat{\mathcal{D}}_{test})$, as opposed to standard FL systems that test on the original poisoned test data $\hat{\mathcal{D}}_{test}$.

Client side. Let $\mathcal{E}(\mathcal{X}_k)$ represent the FCD-encrypted form of input samples from client $\mathcal{C}_k$ based on the shared secret key $\mathcal{K}$. We denote $\mathcal{Q}_k$ as the prediction probabilities when the local model $f_{\theta,k}$ is applied to the transformed data $\mathcal{E}(\mathcal{X}_k)$, which can be expressed as $\mathcal{Q}_k = \sigma(f_{\theta,k}(\mathcal{E}(\mathcal{X}_k)))$, where $\sigma(.)$ denotes the softmax function. Consequently, the cross-entropy loss $\mathcal{L}_{CE_k}$ of client $\mathcal{C}_k$, trained on FCD-encrypted data (*in contrast to general FL systems that calculates $\mathcal{L}_{CE_k}$ on normal data $\mathcal{X}_k$*), can be calculated as follows:

$$\mathcal{L}_{CE_k} = \sum_{i=1}^{\mathcal{N}_k} \mathcal{L}_{CE}(\mathcal{Q}_{i,k}, \mathcal{Y}_{i,k}) = -\sum_{i=1}^{\mathcal{N}_k} \mathcal{Y}_{i,k}\log\mathcal{Q}_{i,k}. \tag{2}$$

To enhance the learning process and facilitate knowledge transfer, we employ the knowledge distillation (KD) approach [60], guided by the Kullback-Leibler (KL) divergence. In this approach, the student model $f_{\theta,k}$ is trained to mimic the predictions of the teacher model. In our FCD framework, there is no distinct teacher model, instead, it is the same local model $f_{\theta,k}$ initially trained on normal data at the beginning of the FL process. We introduce $\mathcal{P}_k$ to denote the prediction probabilities when the local teacher model $f_{\theta,k}$ is applied to normal data $\mathcal{X}_k$, represented as $\mathcal{P}_k = \sigma(f_{\theta,k}(\mathcal{X}_k))$. Subsequently, we calculate the distillation loss guided by KL divergence ($\mathcal{L}_{KLD_k}$) of client k between the prediction probabilities of the teacher and the student local model, which has been trained on normal and FCD-encrypted data, respectively, as follows:

$$\mathcal{L}_{KLD_k}(\mathcal{Q}_k\|\mathcal{P}_k) = \sum_{i=1}^{\mathcal{N}_k}\sum_{r=1}^{\mathcal{R}} \mathcal{Q}^r_{i,k}\log\frac{\mathcal{Q}^r_{i,k}}{\mathcal{P}^r_{i,k}}, \tag{3}$$

where $\mathcal{R}$ is the number of classes. Hence, the complete loss for client $\mathcal{C}_k$, unlike traditional FL systems that solely consider $\mathcal{L}_{CE_k}$, is expressed as:

$$\mathcal{L}_k = \mathcal{L}_{CE_k} + \alpha_k\mathcal{L}_{KLD_k}, \tag{4}$$

where α_k serves as the weighting factor, regulating the balance between the cross-entropy loss and the distillation loss for client $\mathcal{C}_k$. Finally, each client minimizes the total loss $\mathcal{L}_k$ over its *FCD-transformed* data $\mathcal{E}(\mathcal{X}_k)$ across E local iterations. The local model $f_{\theta,k}$ undergoes parameter updates through a backward pass, adjusting as $\theta_k^{\text{new}} = \theta_k^{\text{old}} - \eta_k\nabla_{\theta_k^{\text{old}}}\mathcal{L}_k$, where η_k signifies the learning rate of client $\mathcal{C}_k$. Following each training phase, the client(s) transmit their local updates, denoted as $\nabla\theta^t_k = \theta^t_k - \theta^t_g$, back to the server.

Algorithm 1 Standard FL with **our FCD framework**

```
Input: Global model $\mathcal{G}_{\theta,t}$, local data $\mathcal{D}_k = (\mathcal{X}_k, \mathcal{Y}_k)$
Output: Global test accuracy $\mathcal{A}_g$
Client execution ($\theta^t_g$, $\mathcal{K}$):
for each client $k = 1$ to $n$ do
    Train teacher model $f_{\theta,k}$ on normal data $\mathcal{D}_k$
    $\{\mathcal{P}_{i,k}\}_{i=(1,\mathcal{N}_k)} \leftarrow \{\sigma(f_{\theta,k}(\mathcal{X}_{i,k}))\}_{i=(1,\mathcal{N}_k)}$
$\mathcal{E}(\mathcal{X}_k) \leftarrow$ FCD $(\mathcal{X}_k, \mathcal{K})$
for each client $k = 1$ to $n$ do
    Initialize the local model $\theta^t_k \leftarrow \theta^t_g$
    for $b = 1$ to batches $\in \mathcal{E}(\mathcal{X}_k)$ do
        $\{\mathcal{Q}_{b,k}\} \leftarrow \{\sigma(f_{\theta,0}(\mathcal{E}(\mathcal{X}_k[b])))\}$
        $\mathcal{L}_{CE_k}(\mathcal{Q}_{b,k}, \mathcal{Y}_{b,k}) \leftarrow$ Eq. 2, Cross-entropy loss
        $\mathcal{L}_{KLD_k}(\mathcal{Q}_{b,k}\|\mathcal{P}_{b,k}) \leftarrow$ Eq. 3, Distillation loss
        $\mathcal{L}_k \leftarrow \mathcal{L}_{CE_k} + \alpha_k\mathcal{L}_{KLD_k}$   ▷ Total loss (Eq. 4)
        $\theta^t_k \leftarrow \theta^t_k - \eta\nabla_{\theta^t_k}\mathcal{L}_k$
    $\nabla\theta^t_k \leftarrow \theta^t_k - \theta^t_g$
    return $\nabla\theta^t_k$
Server execution ($\nabla\theta^t_k$):
Share $\theta^t_g$, $\mathcal{K}$ to all the clients
Receive model updates from selected clients $\leftarrow \nabla\theta^t_k$
Perform model aggregation using FedAvg (Eq. 1)
Update the global model parameter: $\theta^{t+1}_g$
$\hat{\mathcal{X}}_{test} \leftarrow$ Poisoned test data   ▷ TM1, TM2
$\mathcal{E}(\hat{\mathcal{X}}_{test}) \leftarrow$ FCD $(\hat{\mathcal{X}}_{test}, \mathcal{K})$
Compute $\mathcal{A}_g \leftarrow$ Test $(\mathcal{G}_{\theta_g^{t+1}}, \mathcal{E}(\hat{\mathcal{X}}_{test}))$
return $\mathcal{A}_g$
```

FCD cryptographic encryption. The missing element in the framework described above is the FCD encryption (a detailed algorithm and a visual representation are in the Supplementary material). We introduce a homomorphic encryption technique for FL utilizing a symmetric transposition cipher [43], with a secret key $\mathcal{K}$ shared among all clients by the server. To ensure data privacy and minimize the potential for brute-force attacks, we define the key's dimension as $\mathcal{K} \in \mathbb{R}^h$, where h corresponds to the image height. The FCD encryption process comprises two steps: (*1*) transposing the input image x to obtain $x' = x^T$, where rows are transformed into columns, and (*2*) reordering rows based on the sequence specified in the secret key $\mathcal{K}$. Essentially, the positions of pixel values in all colour channels (R, G, B) are altered according to $\mathcal{K}$, i.e., $\mathcal{R}_\mathcal{K}(x') = x'[:, \mathcal{K}, :]$, where $\mathcal{R}_\mathcal{K}(.)$ denotes the row-wise shuffling of input data based on the secret key. Further, we introduce the concept of ζ-separability for the FCD encrypted data space. This concept relates to the distance between FCD-encrypted data space and adversarial perturbations. It specifies that the FCD data space maintains a separation of at least ζ from attacked data, ensuring robustness against $\mu\rho$-bounded adversarial perturbation (as defined in Definition 3.1). This

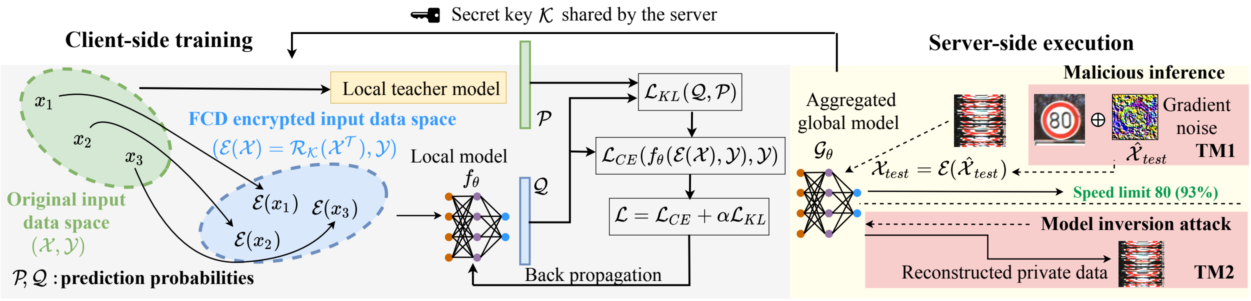
Secret key $\mathcal{K}$ shared by the server
Client-side training
Server-side execution
x_1
x_2
x_3
Local teacher model
$\mathcal{P}$
$\mathcal{L}_{KL}(\mathcal{Q}, \mathcal{P})$
FCD encrypted input data space
$(\mathcal{E}(\mathcal{X}) = \mathcal{R}_{\mathcal{K}}(\mathcal{X}^{\mathcal{T}}), \mathcal{Y})$
Local model
f_θ
$\mathcal{Q}$
$\mathcal{L}_{CE}(f_\theta(\mathcal{E}(\mathcal{X}), \mathcal{Y}), \mathcal{Y})$
$\mathcal{L} = \mathcal{L}_{CE} + \alpha \mathcal{L}_{KL}$
Original input data space
$(\mathcal{X}, \mathcal{Y})$
$\mathcal{E}(x_1)$
$\mathcal{E}(x_3)$
$\mathcal{E}(x_2)$
$\mathcal{P}, \mathcal{Q}$: prediction probabilities
Back propagation
Aggregated global model
$\mathcal{G}_\theta$
Malicious inference
80
⊕
Gradient noise
$\hat{\mathcal{X}}_{test}$
TM1
$\mathcal{X}_{test} = \mathcal{E}(\hat{\mathcal{X}}_{test})$
Speed limit 80 (93%)
Model inversion attack
Reconstructed private data
TM2

Figure 2. Overview of the FCD-integrated FL system, with a focus on one client's training and server-side execution. We compute the cross-entropy loss ($\mathcal{L}_{CE}$) for the local model on encrypted data and calculate the distillation loss ($\mathcal{L}_{KL}$) using KL divergence between the pretrained teacher model and the local student model for normal and encrypted data, respectively. These losses collectively update the local model f_θ. Similarly, test data is transformed into the encrypted space to counter potential attacks.

property is crucial to ensure that the FCD-encrypted data space effectively mitigates utility evasion attacks during inference (TM1).

Definition 3.3 *(ζ-**separability.**)*

$$Let\ \mathcal{E}(\mathcal{D}) := \bigcup_{k \in [1,n], i \in [1, \mathcal{N}_k]} \{(\mathcal{E}(\mathcal{X}_{k,i}), \mathcal{Y}_{k,i})\} \subseteq \mathbb{R}^d \times \mathbb{R}$$

represent the FCD encrypted local training data space. Similarly, $\mathcal{E}(\mathcal{D}_{test})$ denotes the FCD encrypted test data space such that for both training and test splits, $\mathcal{E}(\mathcal{X}) = \mathcal{R}_{\mathcal{K}}(\mathcal{X}^{\mathcal{T}})$. We say that our FCD encrypted data space is ζ-separable w.r.t. $\mu\rho$-bounded adversarial data perturbation $\tilde{\mathcal{X}}$ (as given in Definition 3.1) for some $\zeta > 0$ iff $\|\mathcal{E}(\mathcal{X}_i) - \tilde{\mathcal{X}}_j\| \geq \zeta$ holds for any $\mathcal{E}(\mathcal{X}_i) \in \mathcal{E}(\mathcal{D})$, $\mathcal{E}(\mathcal{X}_i) \in \mathcal{E}(\mathcal{D}_{test})$, and $\tilde{\mathcal{X}}_j \in \mathcal{A}_{\mu\rho}(\mathcal{X}_{test})$.

Rationale for FCD Design. The transposition of data before row-wise transformations enhances information security robustness by obscuring critical transformation details, providing defense against black-box evasion attacks. The use of a row-based transposition cipher, compatible with row-major programming languages (Python, C, and C++) [46], ensures efficiency in memory access, parallel processing, and cryptographic operations, aligning with both security and computational considerations in our FCD integrated FL system. Please refer to the Supplementary material for a more detailed rationale.

FCD dual property benefits aligned with defender's perspective. *(i) Low visibility:* The FCD transformation, incorporating both transposition and a secret key, ensures low visibility by creating unique gradients for each key, offering robust protection against various attacks. *(ii) Low budget:* FCD's vectorized operations and constrained key dimension facilitate efficient implementation, making it suitable for large-scale FL systems with minimal resource requirements and addressing concerns related to processing time and aligning *w.r.t.* defender's perspective. Please refer to the Supplementary material for more details about the FCD dual property.

Lemma 3.4 *The expected time complexity of our FCD encryption function $\mathcal{E}(\mathcal{X})$ is linear, specifically $\mathcal{O}(nh)$, where n represents the number of samples, and h denotes the image height. This is a significant enhancement compared to the $\mathcal{O}(n_w \mathcal{N}^2 \log \mathcal{N})$ time complexity required for encryption and decryption of model parameters, where n_W represents the number of model parameters, and $\mathcal{N}$ is the bit length of the key [19]. (Please refer Supplementary material for the proof.)*

3.1. Convergence and Feasibility Analysis of FCD

We start by establishing the convergence of our FCD integrated FL global model. Subsequently, we demonstrate the robustness of our FCD against gradient noise-based $\mu\rho$-bounded data poisoning attacks (as defined in Definition 3.1). Finally, we establish the resilience of our FCD against MIAs with ϵ-distorted characteristics (as described in Definition 3.2) at the server. We provide the respective proofs in the Supplementary material.

Theorem 3.5 ***FCD convergence.*** *Under the regularity conditions of L-smoothness, τ-strong convexity, and a decaying learning rate, our FCD integrated FL with clients trained on encrypted data $\mathcal{E}(X)$, obtained using FCD$(\mathcal{X}, \mathcal{K})$, the global model converges to*

$$\mathbb{E}[\mathcal{G}_{\theta_g}] - \mathcal{G}^* \leq \frac{2L}{\tau(\gamma + T)} \left(\frac{B + C}{\tau} + 2L\|\theta_g^0 - \theta_g^*\|^2 \right) + D.$$

Here, the variables have the following meanings: $B = \Gamma + (E-1)^2$, where Γ represents the measure of non-IID data distribution. C signifies the client selection for aggregation, with $C = 0$ when all n client updates are considered. T denotes the number of global communication *rounds, and $\mathcal{G}^*$ represents the optimal global model [28]. The positive constant $D \leq \psi$ quantifies how distillation enhances the convergence rate. It accelerates convergence by transferring learnable knowledge from the local teacher model trained on $\mathcal{X}$ to a student model trained on $\mathcal{E}(\mathcal{X})$. The constant ψ quantifies how distillation accelerates convergence in our specific setup.*

Corollary 3.5.1 *(Robustness to $\mu\rho$-bounded attacks.) Let $\mathcal{A} := \mathcal{G}_{\theta_g} \times \mathcal{X}_{test} \times \mathbb{R} \rightarrow \tilde{\mathcal{X}}_{test}$ be the induced adversarial perturbations on the test data using the black-box global model $\mathcal{G}_{\theta_g}$. This perturbation adheres to the constraint $\|\tilde{\mathcal{X}}_{test} - \mathcal{X}_{test}\| \leq \mu$ such that $\mathcal{A}(\mathcal{G}_{\theta_g}, \mathcal{X}_{test,i}, \mathcal{Y}_i)_{i=1}^{\rho} \in \mathcal{B}(\mathcal{X}_{test,i}, \mu)_{i=1}^{\rho}$ (according to Definition 3.1). Furthermore, $\mathcal{G}_{\theta_g}$ is robust to $\tilde{\mathcal{X}}_{test}$ through the mechanism of our FCD integrated FL system. It allows $\mathcal{G}_{\theta_g}$ to test on encrypted data based on $\mathcal{K}$, i.e., $\mathcal{E}(\tilde{\mathcal{X}}_{test})$ instead of $\tilde{\mathcal{X}}_{test}$. This robustness is achieved because the adversary's objective is to generate perturbations in the original test data space that maximize the global loss (Definition 3.1), as denoted by $\mathcal{A}^+ := \arg\max_{\tilde{\mathcal{X}}_{test} \in \mathcal{B}(\mathcal{X}_{test}, \mu)} \mathcal{L}(\mathcal{G}_{\theta_g}(\tilde{\mathcal{X}}_{test,i}, \mathcal{Y}_i)_{i=1}^{\rho})$. However, this adversarial strategy is effective when training and testing occur in the normal data space $\mathcal{X}$ at the client side, matching the distribution of $\mathcal{X}_{test}$. In our FCD integrated FL system, training and testing occur in the encrypted data space. Notably, $\mathcal{L}(\mathcal{G}_{\theta_g}(\tilde{\mathcal{X}}_{test,i}, \mathcal{Y}_i)_{i=1}^{\rho}) \neq \mathcal{L}(\mathcal{G}_{\theta_g}(\mathcal{E}(\tilde{\mathcal{X}}_{test,i}), \mathcal{Y}_i)_{i=1}^{\rho})$ as $\mathcal{E}(\tilde{\mathcal{X}}_{test}) \triangleq \mathcal{R}_{\mathcal{K}}(\tilde{\mathcal{X}}_{test}^{\mathcal{T}}) \neq \tilde{\mathcal{X}}_{test}$ and $\|\mathcal{E}(\mathcal{X}_i) - \tilde{\mathcal{X}}_j\| \geq \zeta$ holds for any $\mathcal{E}(\mathcal{X}_i) \in \mathcal{E}(\mathcal{D})$, $\mathcal{E}(\mathcal{X}_i) \in \mathcal{E}(\mathcal{D}_{test})$, and $\tilde{\mathcal{X}}_j \in \mathcal{A}_{\mu\rho}(\mathcal{X}_{test})$ (Definition 3.3). Hence, the learning process occurs in different data spaces unknown to the adversary. Consequently, the gradients of the loss function trained on normal data are notably distinct from those trained on FCD encrypted data, effectively mitigating $\mu\rho$-bounded adversarial data.*

Theorem 3.6 *(Resilience to ϵ-distorted MIA attacks.) Let $\mathcal{X}^*$ represent the data reconstructed by the adversary using MIA attack as per the conditions specified in Definition 3.2. We demonstrate that training on FCD-encrypted data space, denoted as $\mathcal{E}(\mathcal{X})$, imparts resilience to ϵ-distorted MIA attacks. Specifically, our result establishes that:*

$$\Big| \|\mathcal{X}^*\| - \|\mathcal{E}(\mathcal{X})\| \Big| \leq \epsilon + \delta, \text{ for some } \epsilon \geq 0 \text{ and } \delta \geq 0.$$

4. Experiments

Datasets, implementation details, and metrics. Table 2 presents comprehensive details about the datasets, FL setup, attack percentage, and metrics. For TM1, we employed the black-box and active data poisoning technique called MSimBA [23]. Each experiment was conducted thrice, and results were averaged with standard deviations presented. Please refer to the Supplementary material for more details.
Baselines. For TM1, we have chosen FAT [61] and Randomized Smoothing (RS) [8] based on their relevance and applicability in evasion attacks within FL, as outlined in Table 1. Additionally, we explore the effectiveness under the following configurations: (a) With/without attack and defense and (b) two FCD setups under different client counts and attack percentages (A_p). These setups include *FCD-homoFL* (homogeneous FL) and *FCD-hetFL* (heterogeneous FL). In TM2, we follow the baselines as per [27].

4.1. FCD Result Discussion

Benign setting. Table 3 showcases the performance of our FCD defense, along with other baselines and configurations, across four datasets under a **no attack scenario**. Notably, FCD demonstrates comparable or higher accuracy than vanilla FL, particularly excelling in datasets like GTSRB and KBTS. This improvement is attributed to FCD's ability to capture diverse transformation patterns with a secret key, facilitating reduced training loss and improved model convergence. Furthermore, with its symmetric cryptographic transformation, FCD outperforms FAT and RS for GTSRB and KBTS. However, in CIFAR10 and EMNIST, FCD's performance, while slightly lower than base FL, surpasses FAT and RS. This trade-off between utility and privacy is expected, as FCD provides both aspects with a modest decrease in base accuracy.
Attack setting. We analyse FCD effectiveness under two threat models separately.
(i) TM1: Table 4 illustrates the impact on utility (U) for FCD compared to other baselines under M-SimBA evasion attack on GTSRB and KBTS datasets. The results indicate a significant utility degradation with higher attack percentages and more clients, emphasizing the impact of substantial test data modification by M-SimBA. FCD consistently demonstrates superior performance and robustness across all attack percentages. In homogeneous FL settings, FCD shows substantial improvement, maintaining a consistently lower attack impact on utility compared to other defenses. Particularly at $A_p = 30\%$, the attack impact on utility is 2-6 units lower for FCD as the number of clients increases, compared to other baselines. Similar trends are observed for $A_p = 50\%, 100\%$. In heterogeneous FL settings, FCD outperforms other defenses, showcasing robustness even when clients are trained on varying amounts of non-IID data. For the KBTS dataset, FCD exhibits substantial improvement and lower attack impact across all attack percentages and FL settings compared to other defenses. The limited training data in the KBTS dataset results in the RS technique showing limited robustness, especially for $A_p = 30\%$ in homogeneous and heterogeneous FL, followed by FAT. Ta-

Table 2. Comprehensive experimental details: datasets, models, FL setup, attack configuration, and metrics.

Threat model	**Dataset**	**Total clients**, n	**Client updates used per round**, m	**Attack percentage (%)**, $A_p = \frac{\rho}{\mathcal{N}_te} \times 100$	**Model**	**Global epochs**	**Local epochs**	α	**Metrics**
TM1	GTSRB [44]	3, 5, 10, 15, 25	3, 5, 10, 15, 25	30, 50, and 100	Custom CNN	200	5	0.1, 0.2, 0.5 (default), 1, 1.5, 2	$U = \mathcal{A}_g - \mathcal{A}_g^*$
	KBTS [31]								
	CIFAR10 [22]	100	40 and 70		ResNet18 [14]	500	10		
	EMNIST [7]	10000	100 and 500		LeNet5 [26]				
TM2	CIFAR100 [22]	Exact setup used in [27], n = m = 2		-	VGG11 [42]	200	1		MSE

Table 3. Comparison of evasion defenses for homogeneous **(Hom)** and heterogeneous **(Het)** FL settings on four datasets, in terms of best global test accuracy ($\mathcal{A}_g\%$) ↑ **under no attack**. All values are percentages. **ND** denotes no defense FL system.

	GTSRB [44]		**KBTS [31]**		**CIFAR10 [22]**		**EMNIST [7]**	
Method	**Hom**	**Het**	**Hom**	**Het**	**Hom**	**Het**	**Hom**	**Het**
ND	$96.35_{\pm 0.02}$	$94.98_{\pm 0.06}$	$98.30_{\pm 0.80}$	$97.10_{\pm 0.62}$	$82.14_{\pm 1.15}$	$81.54_{\pm 1.64}$	$89.26_{\pm 2.12}$	$88.34_{\pm 1.34}$
FAT [61]	$97.38_{\pm 0.51}$	$96.22_{\pm 0.84}$	$97.37_{\pm 0.46}$	$96.75_{\pm 0.10}$	$78.12_{\pm 1.90}$	$75.43_{\pm 2.81}$	$84.12_{\pm 1.55}$	$83.28_{\pm 1.48}$
RS [8]	$97.63_{\pm 0.34}$	$96.39_{\pm 0.59}$	$97.09_{\pm 0.37}$	$96.26_{\pm 0.26}$	$76.45_{\pm 0.63}$	$75.18_{\pm 1.57}$	$84.65_{\pm 1.27}$	$82.95_{\pm 2.18}$
FCD (ours)	$97.72_{\pm 0.43}$	$96.68_{\pm 0.18}$	$97.43_{\pm 0.12}$	$96.80_{\pm 0.46}$	$77.28_{\pm 1.77}$	$76.16_{\pm 0.19}$	$85.56_{\pm 2.12}$	$83.16_{\pm 2.08}$

ble 5 presents results for CIFAR10 and EMNIST datasets, highlighting FCD's consistent robustness against evasion attacks. However, a slightly higher attack impact on utility is attributed to the server's random selection of fewer client updates for aggregation, leading to a less accurate global model. Despite this randomness favoring the attacker, FCD maintains superior performance compared to other defenses. Additionally, Figure 3 provides qualitative results of the FCD method under M-SimBA attack, complementing the quantitative analysis.

(ii) TM2: Table 6 presents the MSE for FCD compared to other baselines under MIA attacks in split FL [27]. FCD exhibits a very high reconstruction error for MIA due to its ability to operate under encrypted data space. This is because the FCD integrated FL system trains the local model on encrypted data space, resulting in the reconstructed images by the adversary also being in encrypted form, leading to a high MSE, as shown in Figure 4. *In summary, FCD, as a unified defense, demonstrates robustness against evasion utility attacks and resilience to MIA privacy attacks under TM1 and TM2, respectively.*

5. Limitations and Potential Solutions

(i) **Adaptive key attack**: An adversary can introduce noise to transform images and attempt inverse transformations with an assumed key to generate adversarial samples. To counter such attacks, employing *multi-key encryption* standards, such as $\mathcal{E}(\ldots\mathcal{E}(\mathcal{E}(\mathcal{X},\mathcal{K}_1),\mathcal{K}_2),\ldots\mathcal{K}_l)$, proves effective. The adaptive nature of the attack necessitates the adversary to guess keys randomly or heuristically, given the lack of direct access to the key(s). Successfully misleading the model becomes challenging unless the estimated key is sufficiently close to the correct key. Due to the high dimensionality of $\mathcal{K} \in \mathbb{R}^h$, finding a key proximate to $\mathcal{K}$ is a non-trivial task. Moreover, conducting a computational search for all keys in the multi-key encryption becomes a computationally demanding problem.

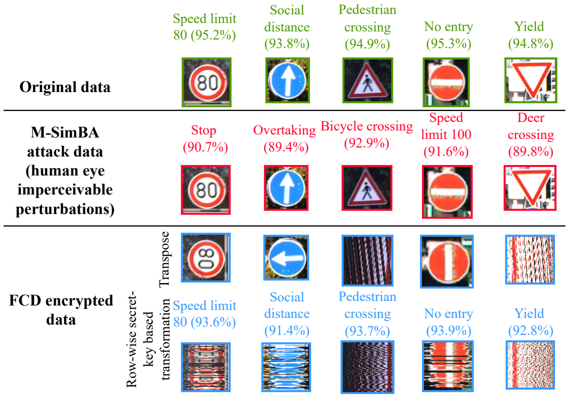

Figure 3. Visualization of FCD transformed M-SimBA adversarial samples on the GTSRB dataset in TM1 with original class labels, adversarial attack samples and labels, and FCD-transformed images and labels, with corresponding confidence scores provided beneath each label.

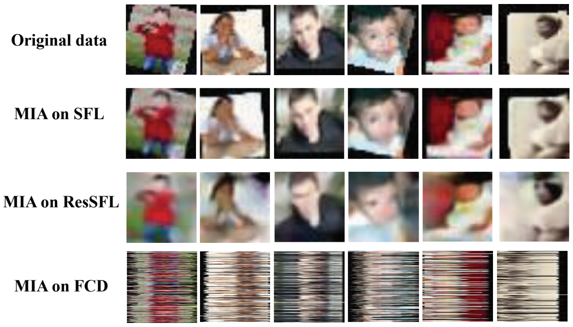

Figure 4. Visualization of FCD's successful defense against MIA on the CIFAR100 dataset in TM2. As FL is trained on FCD-transformed images, the regenerated images from MIA closely resemble the encrypted data space, preserving the original data form.

(iii) **Adaptive gradient attack**: Estimation over transformations proves beneficial for calculating gradients in adversarial defenses [2]. Instead of taking a single step in the direction of gradients $\nabla_x f(x)$, the adversary can aggregate over multiple steps, given by $\sum_{i=1}^{s} \nabla_x f(x)$, where s is the number of keys used by the attacker to gener-

Table 4. Comparison of evasion defenses in terms of impact on utility $(U) \downarrow$ **under M-SimBA attack** across different FL configurations for two datasets. **ND** denotes an FL system without defense. **Bold** indicate best results.

Dataset	Setting	$A_p \rightarrow$	30%				50%				100%			
		$n = m$	**ND**	**FAT [61]**	**RS [8]**	**FCD (ours)**	**ND**	**FAT [61]**	**RS [8]**	**FCD (ours)**	**ND**	**FAT [61]**	**RS [8]**	**FCD (ours)**
GTSRB [44]	**Hom**	3	49.55±1.60	0.99±0.57	0.94±0.71	**0.52±0.73**	50.05±1.34	1.35±1.11	0.43±1.90	**0.27±1.12**	53.25±1.10	1.92±1.16	1.74±1.23	**0.85±0.99**
		5	58.15±1.69	1.17±0.56	1.09±0.29	**0.71±1.32**	58.25±1.41	1.43±1.47	1.83±1.50	**1.01±0.87**	64.25±1.83	3.01±0.34	2.32±1.61	**1.27±1.59**
		10	71.85±0.90	2.47±0.62	1.86±0.51	**0.91±1.04**	73.45±0.41	4.64±0.43	4.3±1.85	**3.12±0.84**	79.35±1.80	7.88±1.13	5.93±1.31	**4.41±1.72**
		15	70.35±1.23	8.84±1.88	3.77±1.12	**2.84±0.70**	72.65±1.84	9.95±1.18	5.94±1.87	**4.13±1.38**	75.25±1.17	11.81±1.25	7.83±1.03	**5.2±1.25**
		25	79.55±1.13	13.03±1.51	8.93±1.53	**6.62±0.68**	79.95±1.05	13.47±0.97	14.48±1.62	**7.92±1.39**	81.55±0.67	17.12±0.94	11.3±0.27	**9.02±0.61**
	Het	3	56.18±1.93	0.34±1.57	0.31±0.11	**0.23±0.19**	56.58±1.15	0.86±0.66	0.15±1.05	**0.10±0.77**	59.48±1.53	1.03±1.36	0.96±1.70	**0.22±0.75**
		5	62.68±1.09	2.41±1.81	3.59±1.75	**1.78±1.09**	64.28±1.81	4.39±1.42	2.38±0.58	**1.41±0.26**	64.68±0.59	3.38±1.64	4.59±1.86	**2.9±0.97**
		10	64.28±1.46	7.74±1.27	6.79±0.29	**5.82±0.93**	69.28±1.54	3.76±1.65	7.29±1.52	**1.81±1.43**	73.78±1.62	9.72±1.15	7.39±0.98	**5.52±0.36**
		15	73.78±1.12	7.68±0.95	7.44±0.69	**0.96±1.53**	74.58±0.76	8.62±0.19	10.21±0.93	**6.33±0.73**	75.38±1.18	11.92±1.57	13.15±1.72	**9.28±1.01**
		25	74.28±0.45	14.33±0.40	17.79±1.73	**8.08±0.11**	76.98±0.35	15.11±0.92	23.61±0.15	**9.28±1.79**	77.08±1.32	18.86±1.93	24.86±1.30	**11.58±0.17**
KBTS [31]	**Hom**	3	15.5±0.29	2.74±0.22	6.42±1.64	**0.13±1.45**	17.3±0.71	4.73±1.37	7.75±1.49	**0.13±1.06**	17.7±0.36	5.64±1.40	8.41±1.51	**1.73±0.39**
		5	20±1.39	4.03±0.72	8.63±1.91	**0.63±0.52**	20.5±1.29	8.96±0.70	10.37±0.66	**0.73±1.66**	24.1±1.96	12.24±1.87	16.35±1.16	**4.23±0.61**
		10	20.4±0.88	16.65±1.95	36.85±0.20	**0.63±1.10**	25.5±1.55	17.43±1.72	38.7±1.25	**1.33±1.27**	27.7±0.64	17.57±0.13	46.6±1.68	**6.93±0.28**
		15	31.7±0.81	21.41±0.95	51.56±0.63	**1.03±1.21**	32.5±0.42	30.21±1.22	66.5±0.43	**1.67±1.04**	41.4±0.91	31.94±1.41	77.45±1.26	**8.53±1.01**
		25	46.9±1.04	34.17±1.07	85.94±1.60	**1.33±1.65**	40.7±1.88	34.41±1.23	86.11±2.01	**2.03±1.42**	48.4±1.94	35.38±1.35	87.29±1.27	**11.83±0.15**
	Het	3	15.2±1.59	1.56±0.51	9.47±1.23	**0.35±1.18**	18.1±1.89	2.05±1.58	11.3±1.02	**0.5±0.62**	20.8±1.86	6.48±1.08	17.38±1.81	**1.7±0.89**
		5	22.2±0.39	5.49±1.10	13.13±1.74	**0.53±1.02**	23.4±0.13	7.02±1.73	19.64±1.09	**2.4±1.24**	26±1.45	10.87±1.71	22.77±1.26	**5.24±0.51**
		10	27.3±1.68	14.59±0.39	39.77±0.20	**1.32±1.48**	29.6±0.58	14.79±1.99	40.29±0.32	**5.12±1.95**	32.2±1.14	18.15±0.32	41.76±1.23	**6.7±0.30**
		15	32.7±0.38	17.09±0.66	57.1±1.56	**1.7±0.21**	36.6±0.98	21.19±1.93	58.34±0.36	**6.6±1.53**	42.6±0.23	23.59±0.57	60.82±1.37	**9.5±0.12**
		25	43.6±0.49	31.7±1.39	73.31±1.40	**1.1±1.12**	45.1±1.32	32.66±1.63	76.1±1.53	**8.2±1.20**	46.2±0.51	34.11±1.20	78.19±0.51	**12.2±0.70**

Table 5. Comparison of evasion defenses in terms of impact on utility $(U) \downarrow$ **under M-SimBA attack** across different FL configurations, n, m for two datasets. **ND** denotes an FL system without defense. **Bold** indicate best results.

Dataset	Setting	$A_p \rightarrow$		30%				50%				100%			
		n	m	**ND**	**FAT [61]**	**RS [8]**	**FCD (ours)**	**ND**	**FAT [61]**	**RS [8]**	**FCD (ours)**	**ND**	**FAT [61]**	**RS [8]**	**FCD (ours)**
CIFAR10 [22]	**Hom**	100	40	27.37±1.34	16.05±0.85	18.23±0.27	**10.81±0.49**	32.82±0.42	20.2±1.24	42.89±0.37	**11.95±0.69**	58.53±0.99	46.18±0.92	65.26±1.13	**13.04±0.44**
			70	30.54±0.46	18.93±1.50	22.37±1.42	**12.94±0.60**	45.85±0.24	24.67±1.06	45.41±0.77	**13.04±1.42**	66.07±0.94	48.1±0.98	65.44±0.40	**14.02±1.16**
	Het	100	40	25.39±1.46	10.47±1.44	20.22±0.82	**9.82±1.11**	35.95±0.39	21.11±0.38	39.87±1.32	**10.8±0.95**	57.1±0.71	37.49±0.17	61.45±0.35	**13.84±0.11**
			70	32.97±0.54	11.12±0.31	24.14±1.09	**11.83±1.26**	41.96±0.50	22.63±1.01	41.88±1.18	**12.92±0.97**	61.68±0.79	40.12±0.91	63.92±1.30	**14.04±0.86**
EMNIST [7]	**Hom**	10000	100	22.1±0.42	9.47±0.19	27.73±0.18	**10.44±0.23**	39.76±0.64	19.7±0.18	38.42±1.19	**12.47±0.76**	58.17±1.07	36.65±1.41	56.86±0.58	**12.96±0.51**
			500	34.91±0.45	17.4±0.46	36.2±0.96	**12.34±0.51**	41.8±0.68	27.77±0.20	47.43±1.15	**15.28±0.73**	73.71±1.13	47.49±0.80	72.99±0.57	**15.7±1.39**
	Het	10000	100	20.32±1.03	10.04±0.47	25.98±0.54	**9.49±0.84**	40.32±0.75	17.8±1.10	38.67±0.22	**11.10±0.84**	61.85±0.47	40.02±0.52	61.56±0.49	**12.06±1.45**
			500	34.63±0.73	14.19±1.34	34.37±0.46	**11.88±0.28**	47.12±0.46	30.04±0.28	49.55±1.29	**13.71±0.70**	70.14±1.48	45.5±0.58	71.23±0.85	**15.93±0.36**

Table 6. Comparison of MSE (↑) of FCD with other methods **under TM2 MI attack** for the CIFAR100 dataset.

Defense method	**MSE**
Laplacian [47]	0.011
Dropout [15]	0.009
TopkPrune [56]	0.005
AdvNoise [53]	0.018
DistCorr [49]	0.019
Bottleneck Layers [9]	0.02
ResSFL [27]	0.05
FCD (ours)	**73.23**

ate adversarial samples. The introduced multi-key encryption method $\mathcal{E}(\ldots\mathcal{E}(\mathcal{E}(\mathcal{X},\mathcal{K}_1),\mathcal{K}_2),\ldots\mathcal{K}_l)$, with a sufficiently large $l >> s$ and each key having dimension $\mathbb{R}^h$, makes it computationally challenging to search for nearly identical keys. Additionally, the order of $\mathcal{E}$ encryption in multi-key encryption is unique, i.e., $\mathcal{E}(\mathcal{E}(\mathcal{X},\mathcal{K}_i),\mathcal{K}_j) \neq \mathcal{E}(\mathcal{E}(\mathcal{X},\mathcal{K}_j),\mathcal{K}_i)$ for any given keys $\mathcal{K}_i, \mathcal{K}_j$. Thus, the proposed solutions offer simplicity and robustness from a two-level perspective (multi-key, order of encryption). We plan to explore such attacks and solutions in our future work.

6. Conclusion

In this paper, we introduced FCD, a unified defense method in FL that simultaneously protects against utility evasion attacks and privacy model inversion attacks (MIA). The crux of FCD is that it respects the defender's perspective, carefully managing the trade-off between utility and privacy by transforming the entire learning process into a homomorphic encrypted data space. We formulated two threat models, TM1 & TM2, to address utility and privacy attacks and provided theoretical analysis on convergence along with robustness evaluations for both utility and privacy. Our extensive evaluations across various attack scenarios demonstrate that FCD maintains a lower attack impact on utility and achieves higher Mean Squared Error (MSE) on reconstructed data for privacy attacks, aligning closely with the ideal defender's perspective under different attack settings. However, our approach shows promising results but can be further improved to minimize attack impact under random client selection. Future efforts will also involve investigating multi-key multi-level encryption standards to enhance FCD's resilience against adaptive attackers.

References

[1] Federated learning: Collaborative machine learning without centralized training data. 2017. 1

[2] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International Conference on Machine Learning*, pages 274–283. ICML, 2018. 7

[3] Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. *Advances in Neural Information Processing Systems*, 30, 2017. 2

[4] David Byrd and Antigoni Polychroniadou. Differentially private secure multi-party computation for federated learning in financial applications. In *Proceedings of the First ACM International Conference on AI in Finance*, pages 1–9, 2020. 2

[5] Xiaoyu Cao, Minghong Fang, Jia Liu, and Neil Zhenqiang Gong. Fltrust: Byzantine-robust federated learning via trust bootstrapping. *arXiv preprint arXiv:2012.13995*, 2020. 2

[6] Yu Chen, Fang Luo, Tong Li, Tao Xiang, Zheli Liu, and Jin Li. A training-integrity privacy-preserving federated learning scheme with trusted execution environment. *Information Sciences*, 522:69–79, 2020. 2

[7] Gregory Cohen, Saeed Afshar, Jonathan Tapson, and Andre Van Schaik. Emnist: Extending mnist to handwritten letters. In *2017 international joint conference on neural networks (IJCNN)*, pages 2921–2926. IEEE, 2017. 7, 8

[8] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *international conference on machine learning*, pages 1310–1320. PMLR, 2019. 2, 6, 7, 8

[9] Amir Erfan Eshratifar, Amirhossein Esmaili, and Massoud Pedram. Bottlenet: A deep learning architecture for intelligent mobile cloud computing services. In *2019 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED)*, pages 1–6. IEEE, 2019. 8

[10] Chen Fang, Yuanbo Guo, Yongjin Hu, Bowen Ma, Li Feng, and Anqi Yin. Privacy-preserving and communication-efficient federated learning in internet of things. *Computers & Security*, 103:102199, 2021. 1

[11] Minghong Fang, Xiaoyu Cao, Jinyuan Jia, and Neil Zhenqiang Gong. Local model poisoning attacks to byzantine-robust federated learning. In *Proceedings of the 29th USENIX Conference on Security Symposium*, pages 1623–1640, 2020. 1

[12] Xiaojie Guo, Zheli Liu, Jin Li, Jiqiang Gao, Boyu Hou, Changyu Dong, and Thar Baker. V eri fl: Communication-efficient and fast verifiable aggregation for federated learning. *IEEE Transactions on Information Forensics and Security*, 16:1736–1751, 2020. 1

[13] Ehsan Hallaji, Roozbeh Razavi-Far, Mehrdad Saif, and Enrique Herrera-Viedma. Label noise analysis meets adversarial training: A defense against label poisoning in federated learning. *Knowledge-Based Systems*, page 110384, 2023. 2

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. 7

[15] Zecheng He, Tianwei Zhang, and Ruby B Lee. Attacking and protecting data privacy in edge–cloud collaborative inference systems. *IEEE Internet of Things Journal*, 8(12):9706–9716, 2020. 3, 8

[16] Rui Hu, Yuanxiong Guo, Hongning Li, Qingqi Pei, and Yanmin Gong. Personalized federated learning with differential privacy. *IEEE Internet of Things Journal*, 7(10):9530–9539, 2020. 2

[17] Ahmed Imteaj and M Hadi Amini. Leveraging asynchronous federated learning to predict customers financial distress. *Intelligent Systems with Applications*, 14:200064, 2022. 1

[18] Najeeb Moharram Jebreel and Josep Domingo-Ferrer. Fl-defender: Combating targeted attacks in federated learning. *Knowledge-Based Systems*, 260:110178, 2023. 2

[19] Zoe L Jiang, Hui Guo, Yijian Pan, Yang Liu, Xuan Wang, and Jun Zhang. Secure neural network in federated learning with model aggregation under multiple keys. In *2021 8th IEEE International Conference on Cyber Security and Cloud Computing (CSCloud)/2021 7th IEEE International Conference on Edge Computing and Scalable Cloud (EdgeCom)*, pages 47–52. IEEE, 2021. 5

[20] Xiao Jin, Pin-Yu Chen, Chia-Yi Hsu, Chia-Mu Yu, and Tianyi Chen. Cafe: Catastrophic data leakage in vertical federated learning. *Advances in Neural Information Processing Systems*, 34:994–1006, 2021. 1

[21] Taejin Kim, Shubhranshu Singh, Nikhil Madaan, and Carlee Joe-Wong. pfeddef: Characterizing evasion attack transferability in federated learning. *Software Impacts*, page 100469, 2023. 1

[22] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 7, 8

[23] K Naveen Kumar, C Vishnu, Reshmi Mitra, and C Krishna Mohan. Black-box adversarial attacks in autonomous vehicle technology. In *2020 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, pages 1–7. IEEE, 2020. 6

[24] K Naveen Kumar, C Krishna Mohan, and Linga Reddy Cenkeramaddi. The impact of adversarial attacks on federated learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 1, 3

[25] Yogesh Kumar and Ruchi Singla. Federated learning systems for healthcare: perspective and recent progress. *Federated Learning Systems: Towards Next-Generation AI*, pages 141–156, 2021. 1

[26] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 7

[27] Jingtao Li, Adnan Siraj Rakin, Xing Chen, Zhezhi He, Deliang Fan, and Chaitali Chakrabarti. Ressfl: A resistance transfer framework for defending model inversion attack in split federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10194–10202, 2022. 6, 7, 8

[28] Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, 2020. 6

[29] Xiaoyuan Liu, Hongwei Li, Guowen Xu, Zongqi Chen, Xiaoming Huang, and Rongxing Lu. Privacy-enhanced federated learning against poisoning adversaries. *IEEE Transactions on Information Forensics and Security*, 16:4574–4588, 2021. 2

[30] Zhuoran Ma, Jianfeng Ma, Yinbin Miao, Yingjiu Li, and Robert H Deng. Shieldfl: Mitigating model poisoning attacks in privacy-preserving federated learning. *IEEE Transactions on Information Forensics and Security*, 17:1639–1654, 2022. 2

[31] Markus Mathias, Radu Timofte, Rodrigo Benenson, and Luc Van Gool. Traffic sign recognition — how far are we from the solution? In *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2013. 7, 8

[32] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pages 1273–1282. PMLR, 2017. 3

[33] Thomas Minka. Estimating a dirichlet distribution, 2000. 3

[34] Fan Mo, Hamed Haddadi, Kleomenis Katevas, Eduard Marin, Diego Perino, and Nicolas Kourtellis. Ppfl: privacy-preserving federated learning with trusted execution environments. In *Proceedings of the 19th annual international conference on mobile systems, applications, and services*, pages 94–108, 2021. 2

[35] Arup Mondal, Yash More, Ruthu Hulikal Rooparaghunath, and Debayan Gupta. Poster: Flatee: Federated learning across trusted execution environments. In *2021 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 707–709. IEEE, 2021. 2

[36] Ashwinee Panda, Saeed Mahloujifar, Arjun Nitin Bhagoji, Supriyo Chakraborty, and Prateek Mittal. Sparsefed: Mitigating model poisoning attacks in federated learning with sparsification. In *International Conference on Artificial Intelligence and Statistics*, pages 7587–7624. PMLR, 2022. 2

[37] Hanchi Ren, Jingjing Deng, and Xianghua Xie. Grnn: generative regression neural network—a data leakage attack for federated learning. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 13(4):1–24, 2022. 1

[38] Phillip Rieger, Thien Duc Nguyen, Markus Miettinen, and Ahmad-Reza Sadeghi. Deepsight: Mitigating backdoor attacks in federated learning through deep model inspection. *arXiv preprint arXiv:2201.00763*, 2022. 2

[39] Devansh Shah, Parijat Dube, Supriyo Chakraborty, and Ashish Verma. Adversarial training in communication constrained federated learning. *arXiv preprint arXiv:2103.01319*, 2021. 2

[40] Virat Shejwalkar, Amir Houmansadr, Peter Kairouz, and Daniel Ramage. Back to the drawing board: A critical evaluation of poisoning attacks on production federated learning. In *2022 IEEE Symposium on Security and Privacy (SP)*, 2022. 1

[41] Young Ah Shin, Geontae Noh, Ik Rae Jeong, and Ji Young Chun. Securing a local training dataset size in federated learning. *IEEE Access*, 10:104135–104143, 2022. 3

[42] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. 7

[43] Massoud Sokouti, Babak Sokouti, and Saeid Pashazadeh. An approach in improving transposition cipher system. *Indian Journal of Science and Technology*, 2(8):9–15, 2009. 4

[44] Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. The German Traffic Sign Recognition Benchmark: A multi-class classification competition. In *IEEE International Joint Conference on Neural Networks*, pages 1453–1460, 2011. 7, 8

[45] Nurbek Tastan and Karthik Nandakumar. Capride learning: Confidential and private decentralized learning based on encryption-friendly distillation loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8084–8092, 2023. 1, 2

[46] Jeyarajan Thiyagalingam, Olav Beckmann, and Paul HJ Kelly. An exhaustive evaluation of row-major, column-major and morton layouts for large two-dimensional arrays. In *Performance Engineering: 19th Annual UK Performance Engineering Workshop*, pages 340–351. University of Warwick Coventry, UK, 2003. 5

[47] Tom Titcombe, Adam J Hall, Pavlos Papadopoulos, and Daniele Romanini. Practical defences against model inversion attacks for split neural networks. *arXiv preprint arXiv:2104.05743*, 2021. 8

[48] Dmitrii Usynin, Alexander Ziller, Marcus Makowski, Rickmer Braren, Daniel Rueckert, Ben Glocker, Georgios Kaissis, and Jonathan Passerat-Palmbach. Adversarial interference and its mitigations in privacy-preserving collaborative machine learning. *Nature Machine Intelligence*, 3(9): 749–758, 2021. 1, 3

[49] Praneeth Vepakomma, Abhishek Singh, Otkrist Gupta, and Ramesh Raskar. Nopeek: Information leakage reduction to share activations in distributed deep learning. In *2020 International Conference on Data Mining Workshops (ICDMW)*, pages 933–942. IEEE, 2020. 8

[50] Su Wang, Rajeev Sahay, and Christopher G Brinton. How potent are evasion attacks for poisoning federated learning-based signal classifiers? *arXiv preprint arXiv:2301.08866*, 2023. 1

[51] Zhibo Wang, Mengkai Song, Zhifei Zhang, Yang Song, Qian Wang, and Hairong Qi. Beyond inferring class representatives: User-level privacy leakage from federated learning. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*, pages 2512–2520. IEEE, 2019. 1

[52] Kang Wei, Jun Li, Ming Ding, Chuan Ma, Howard H Yang, Farhad Farokhi, Shi Jin, Tony QS Quek, and H Vincent Poor. Federated learning with differential privacy: Algorithms and performance analysis. *IEEE Transactions on Information Forensics and Security*, 15:3454–3469, 2020. 2

[53] Jing Wen, Siu-Ming Yiu, and Lucas CK Hui. Defending against model inversion attack by adversarial examples. In *2021 IEEE International Conference on Cyber Security and Resilience (CSR)*, pages 551–556. IEEE, 2021. 8

[54] Zuobin Xiong, Zhipeng Cai, Daniel Takabi, and Wei Li. Privacy threat and defense for federated learning with non-iid

data in aiot. *IEEE Transactions on Industrial Informatics*, 18(2):1310–1321, 2021. 2

[55] Guowen Xu, Hongwei Li, Sen Liu, Kan Yang, and Xiaodong Lin. Verifynet: Secure and verifiable federated learning. *IEEE Transactions on Information Forensics and Security*, 15:911–926, 2019. 1

[56] Ziqi Yang, Jiyi Zhang, Ee-Chien Chang, and Zhenkai Liang. Neural network inversion in adversarial setting via background knowledge alignment. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pages 225–240, 2019. 8

[57] Chengliang Zhang, Suyi Li, Junzhe Xia, Wei Wang, Feng Yan, and Yang Liu. Batchcrypt: Efficient homomorphic encryption for cross-silo federated learning. In *Proceedings of the 2020 USENIX Annual Technical Conference (USENIX ATC 2020)*, 2020. 2

[58] Chi Zhang, Sotthiwat Ekanut, Liangli Zhen, and Zengxiang Li. Augmented multi-party computation against gradient leakage in federated learning. *IEEE Transactions on Big Data*, 2022. 2

[59] Jiale Zhang, Bing Chen, Xiang Cheng, Huynh Thi Thanh Binh, and Shui Yu. Poisongan: Generative poisoning attacks against federated learning in edge computing systems. *IEEE Internet of Things Journal*, 8(5):3310–3322, 2020. 1

[60] Linfeng Zhang, Chenglong Bao, and Kaisheng Ma. Self-distillation: Towards efficient and compact neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4388–4403, 2021. 4

[61] Giulio Zizzo, Ambrish Rawat, Mathieu Sinn, and Beat Buesser. Fat: Federated adversarial training. *arXiv preprint arXiv:2012.01791*, 2020. 2, 6, 7, 8

[62] Tianyuan Zou, Yang Liu, Yan Kang, Wenhan Liu, Yuanqin He, Zhihao Yi, Qiang Yang, and Ya-Qin Zhang. Defending batch-level label inference and replacement attacks in vertical federated learning. *IEEE Transactions on Big Data*, 2022. 2