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

K Naveen Kumar, Reshmi Mitra, and C Krishna Mohan

SOUTHEAST MISSOURI STATE UNIVERSITY · 1873® CVPR JUNE 17-21, 2024

Federated Learning (FL)

- **A distributed machine learning** framework with a central server and n clients
- Central server makes use of computational resources of clients, **without the need to collect their data**
- **Data remain on the clients** and only model parameters are exchanged ensuring privacy

Threats models (TM) in our FL setup (Fig. 1)

- **Source:** Honest-but-curious (HbC) central server
- **Knowledge:** Black-box
- **TM1 (utility-centric):** Untargeted evasion data poisoning attack
- **TM2 (privacy-centric):** Model inversion attack (MIA)
- **Capabilities:** No access to clients' data, model updates, aggregation algorithm, and shared-secret key

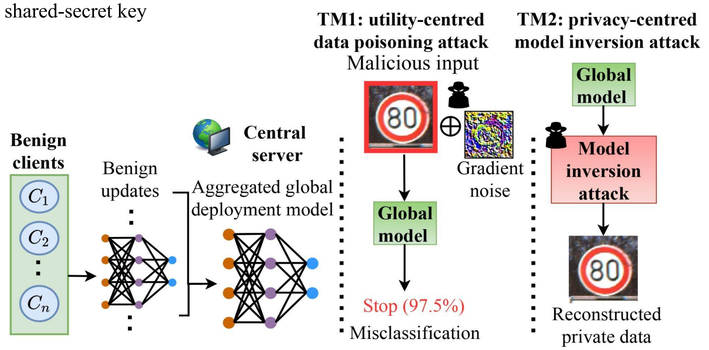

Fig 1. Overview of two different threat models (TM) with potential vulnerabilities and attacks during inference.

Federated cryptography defense (FCD)

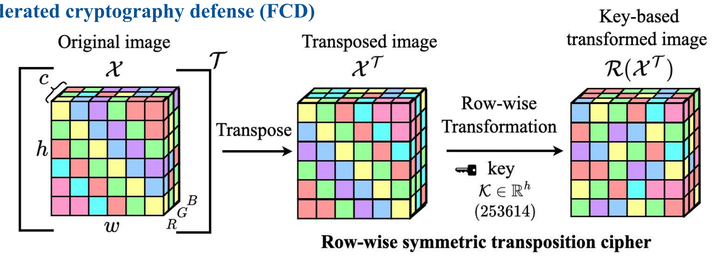

Fig 2. Proposed row-wise symmetric transposition cipher-based FCD transformation

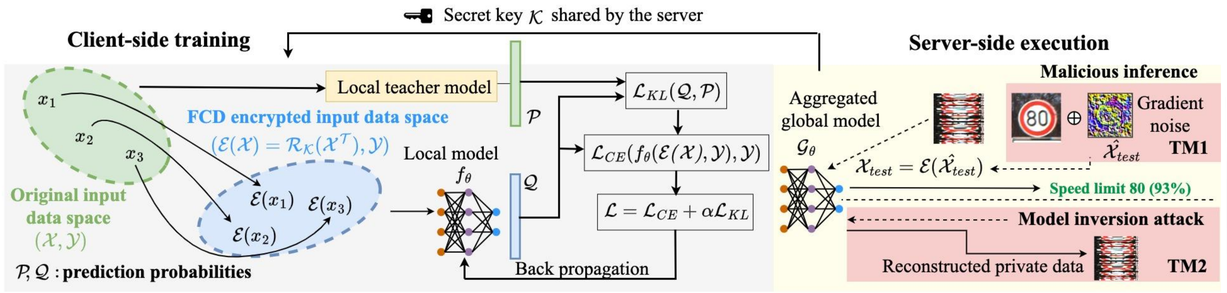

Fig 3. Overview of the FCD-integrated FL system, with a focus on one client's training and server-side execution. We compute the cross-entropy loss $\mathcal{L}_{CE}$ for the local model on encrypted data and calculate the distillation loss using KL divergence $\mathcal{L}_{KL}$ between the pretrained 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.

Lemma: 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.

Table 1. 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 (%), $A_p = \frac{\rho}{Nc} \times 100$	Model	Global epochs	Local epochs	α	Metrics
TM1	GTSRB [61]	3, 5,	3, 5,	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 [40]	10, 15, 25	10, 15, 25						
	CIFAR10 [30]	100	40 and 70		ResNet18 [21]	500	10		
	EMNIST [9]	10000	100 and 500		LeNet5 [34]				
TM2	CIFAR100 [30]	Exact setup used in [35], n = m = 2		-	VGG11 [59]	200	1		MSE

Table 2. Comparison of evasion defenses in terms of impact on utility (U) ↓ under M-SimBA attack across different FL configurations. ND denotes an FL system without defense. Bold indicate best results.

Dataset	Setting	$A_p \rightarrow$ / $n = m$	30% ND	30% FAT [81]	30% RS [10]	30% FCD (ours)	50% ND	50% FAT [81]	50% RS [10]	50% FCD (ours)	100% ND	100% FAT [81]	100% RS [10]	100% FCD (ours)
GTSRB [61]	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**

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 $(\theta_g^t, \mathcal{K})$:
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 $\theta_k^t \leftarrow \theta_g^t$
  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)**
    $\theta_k^t \leftarrow \theta_k^t - \eta \nabla_{\theta_k^t} \mathcal{L}_k$
  $\nabla\theta_k^t \leftarrow \theta_k^t - \theta_g^t$
  return $\nabla\theta_k^t$
Server execution $(\nabla\theta_k^t)$:
Share θ_g^t, $\mathcal{K}$ to all the clients
Receive model updates from selected clients $\leftarrow \nabla\theta_k^t$
Perform model aggregation using FedAvg **(Eq. 5)**
Update the global model parameter: θ_g^{t+1}
$\hat{\mathcal{X}}_{test} \leftarrow$ **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 $\mathcal{A}_g$

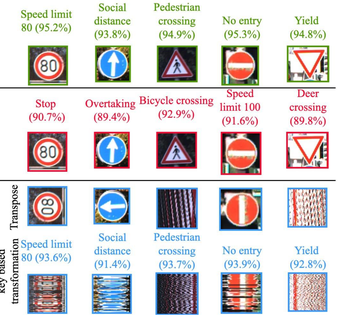

Fig 3. Visualization of FCD transformed M-SimBA adversarial samples on the GTSRB dataset in TM1.

cs19m20p000001@iith.ac.in, rmitra@semo.edu, and ckm@cse.iith.ac.in

Connect with me on LinkedIn: 👉