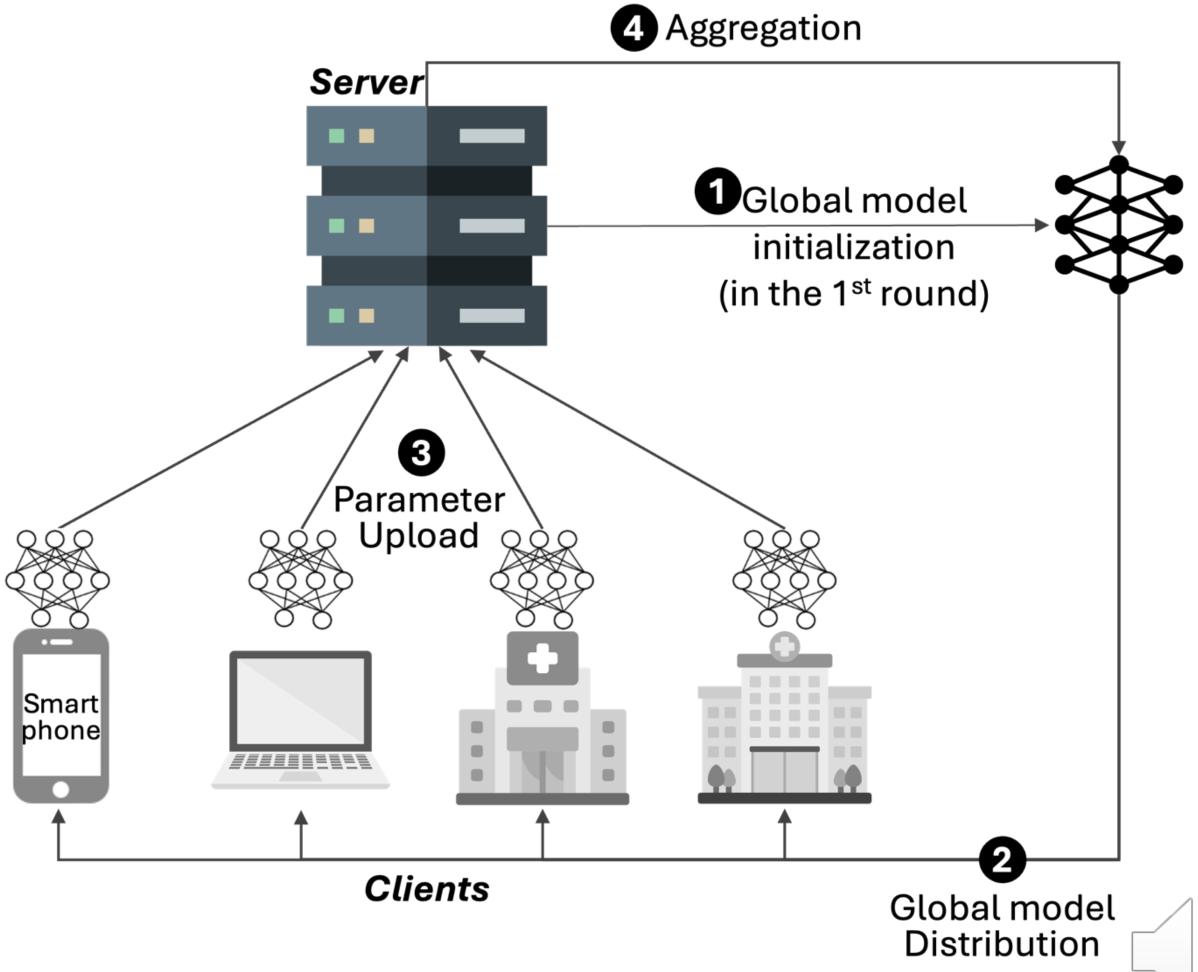
FedCC: Robust Federated Learning against Poisoning Attacks

Hyejun Jeong (UMass Amherst),
Hamin Son (UC Davis), Seohu Lee (Johns Hopkins Univ.),
Jayun Hyun (Hippo T&C Inc.), Tai-Myoung Chung (Hippo T&C Inc.)

SecureComm 2025

What is Federated Learning?

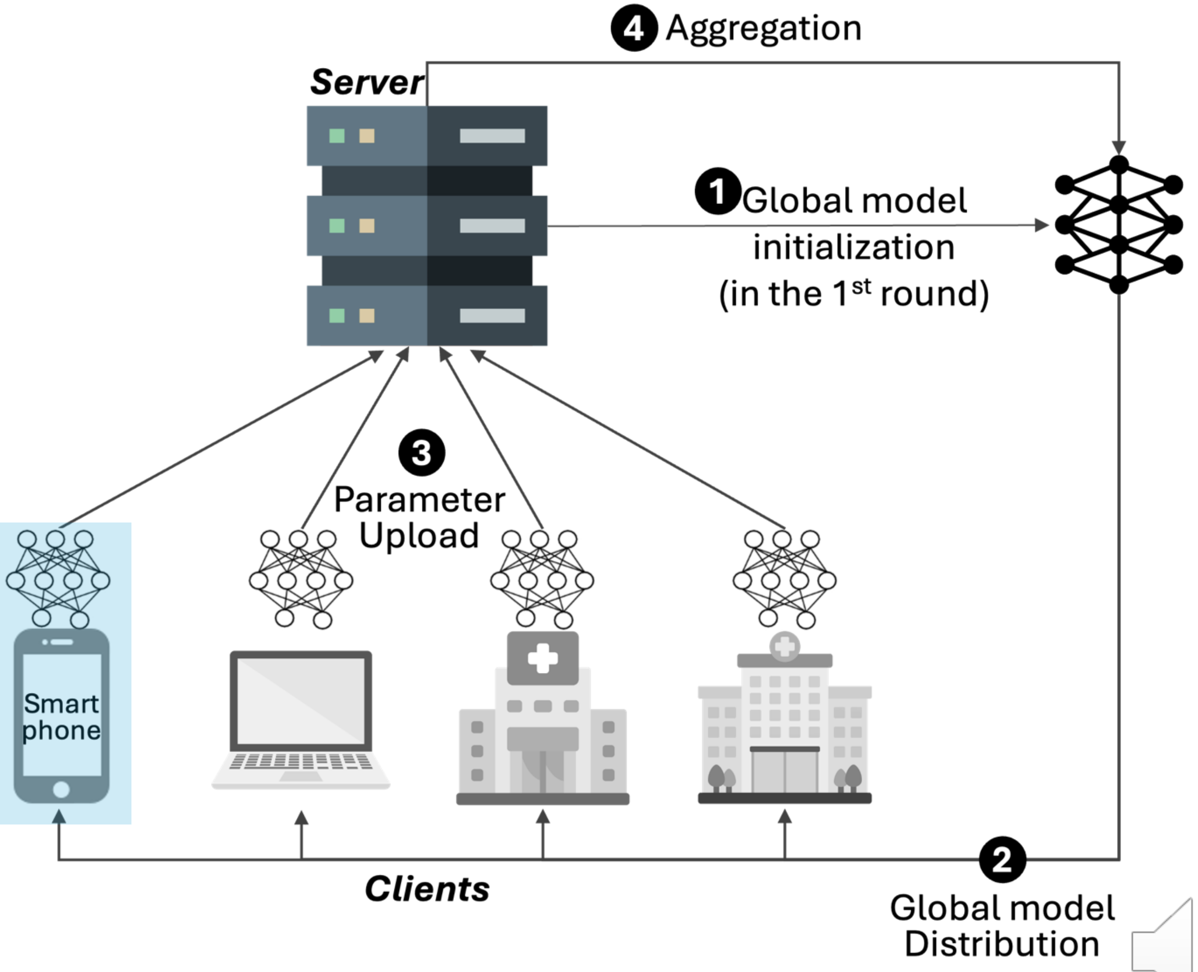
- Local data stays on device, only model weights are shared
- Use cases: mobile phones, hospitals, IoT
- Benefits: privacy, decentralization
- But introduces new attack surfaces



Threats in Federated Learning

- **Untargeted poisoning**: degrade model performance (Fang-Krum, Fang-Med)
- **Targeted attacks / Backdoors**: misclassify specific inputs
- **Challenge**: These attacks are harder to detect under **non-IID data**

Vulnerable to poisoning attacks

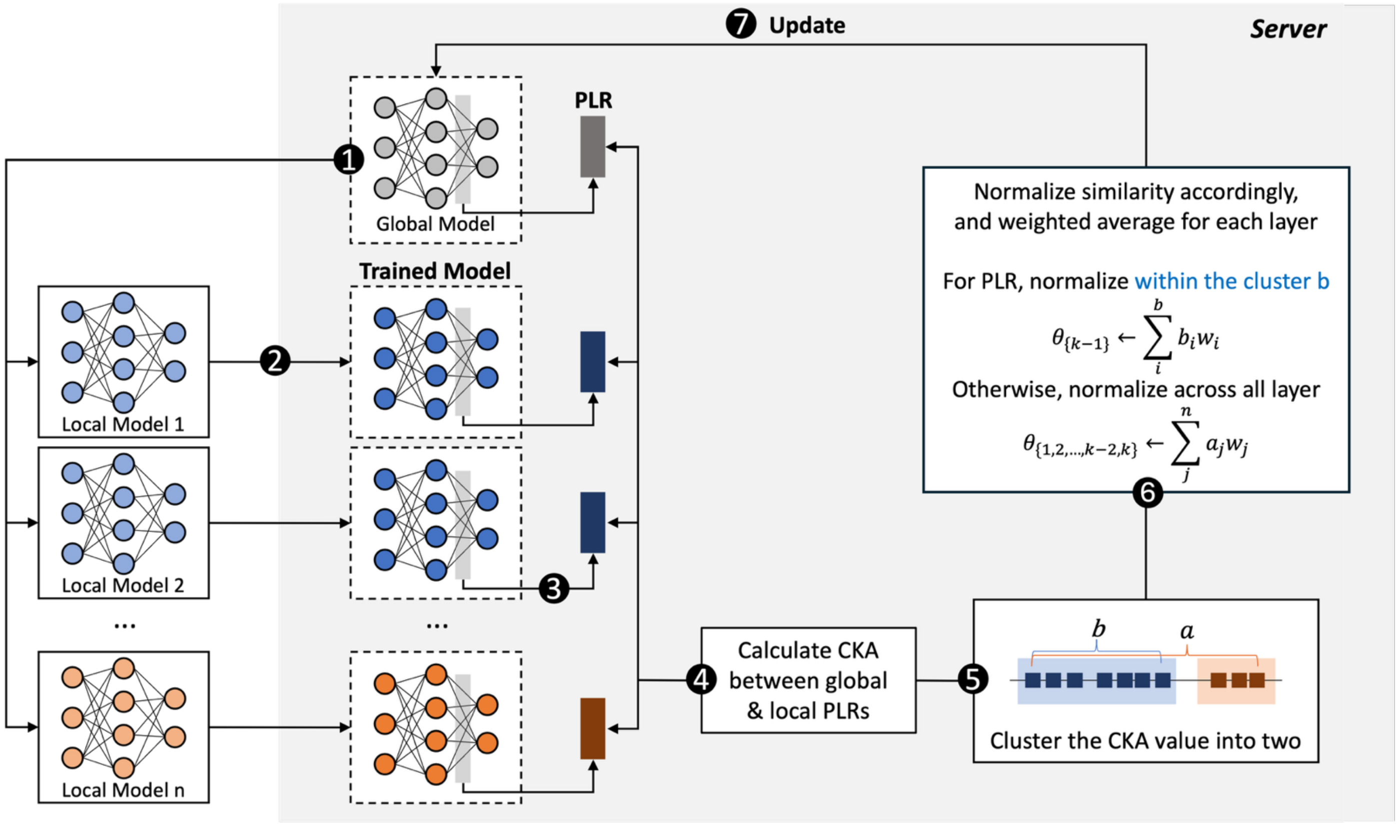


Motivation and Challenge

- Similar Most defenses assume **IID data** or require **manual thresholds**
- Non-IID client data → benign clients look diverse → hard to detect attackers
- Need a defense that:
  - Is **threshold-free**
  - Works under **non-IID**
  - Doesn't require **access to data**

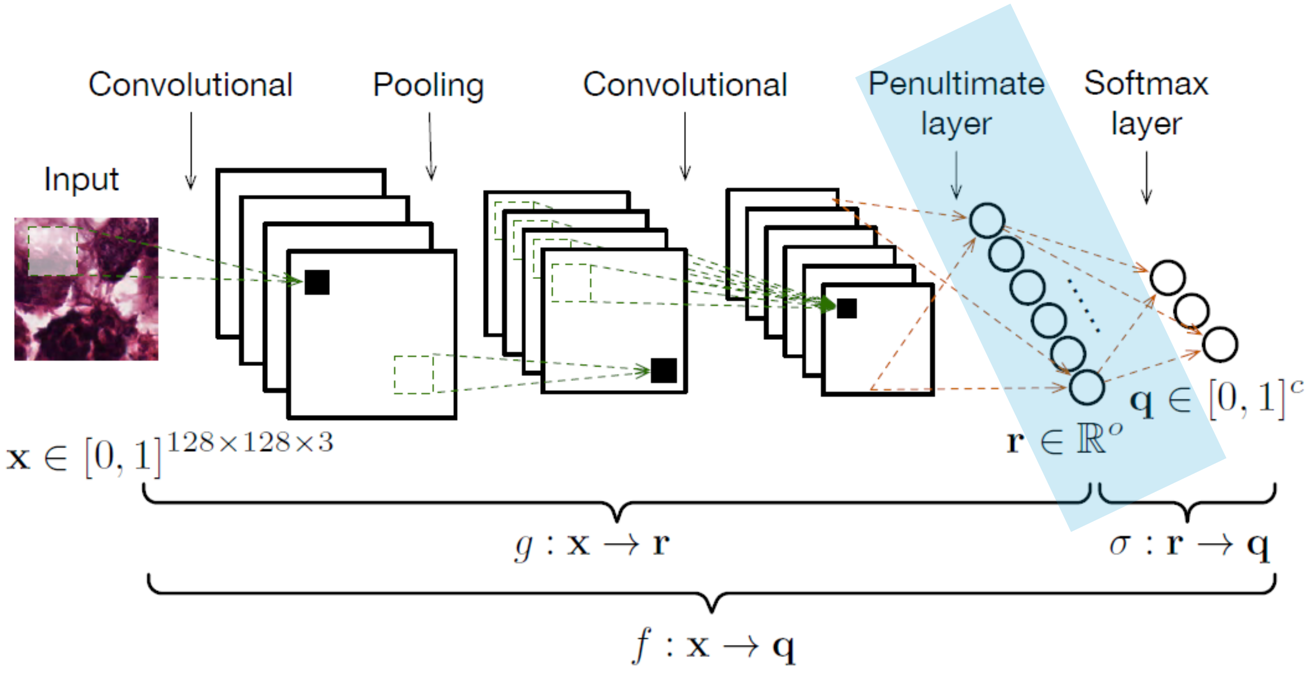
Overview of FedCC

- Core idea: Use **CKA similarity** on **PLRs**
- Use clustering to softly weight (not reject) client updates
- Works under **any client distribution**



Why Penultimate Layer Representations (PLR)?

- Later layers are more sensitive to local data.
- PLRs differentiate the poisonous models [1].
- Backdoor patterns cluster in a penultimate layer latent space [2].



[1] Wang, Ning, et al. "FLARE: Defending Federated Learning against Model Poisoning Attacks via Latent Space Representations." *Proceedings of the 2022 ACM on Asia Conference on Computer and Communications Security*. 2022.
[2] Xiang, Zhen, David J. Miller, and George Kesidis. "A benchmark study of backdoor data poisoning defenses for deep neural network classifiers and a novel defense." *2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2019.

Why Centered Kernel Alignment (CKA)?

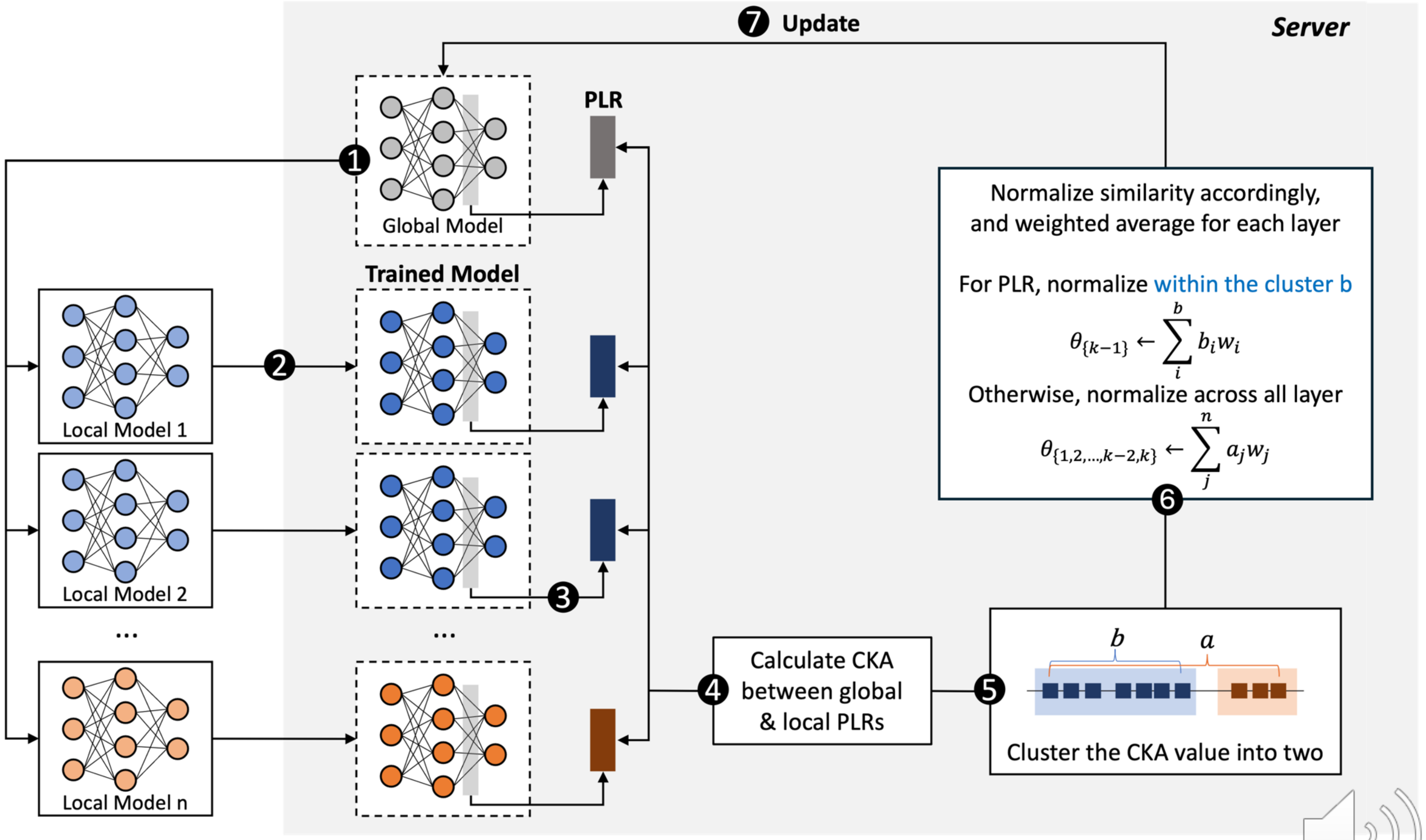
- Compares **representations** across models robustly
- Better than cosine, Euclidean, or MMD
- Handles **scaling, rotations**, and **different weight magnitudes**
- Works well even with **non-IID data**

Table 1: Comparison of Performance with Various Similarity Metrics

Method	Fang-Med		Fang-mKrum		Targeted	
	IID	NIID	IID	NIID	IID	NIID
Kernel CKA	**69.20**	**41.00**	**70.22**	**43.24**	**71.44/6e-07**	**54.62/0.0118**
Linear CKA	10.00	13.13	64.09	39.55	71.02/0.0007	49.53/0.0616
MMD	63.39	40.90	69.69	32.27	70.85/1e-09	50.51/9e-05
Cosine	68.82	33.90	68.81	10.04	69.76/0.0002	53.66/0.0529
Euclidean	69.06	27.82	68.54	41.57	69.17/0.0221	52.20/0.0015

FedCC Aggregation Procedure

1. Send a global model
2. Send local models
3. Extract PLR for each client
4. Compute **CKA similarity** to global model
5. Run **clustering**
6. Apply **within-cluster** normalization on PLRs, **across-cluster** for others
7. **Layer-wise weighted aggregation**



Experimental Setup

- Datasets: fMNIST, CIFAR-10, CIFAR-100
- Architectures: Lightweight CNNs
- Non-IID simulation: a Dirichlet distribution with $\alpha = 0.2$
- Attacks: Fang-Krum, Fang-Med, Targeted Backdoor, DBA
- Baselines: FedAvg, Krum, Coomed, Multi-Krum, Bulyan, FLARE, FLTrust
- Metrics: Accuracy, Backdoor Confidence

Results: Untargeted Attacks (Non-IID)

- FedCC achieves highest accuracy across all datasets
- Other methods misidentify benign clients → lower performance

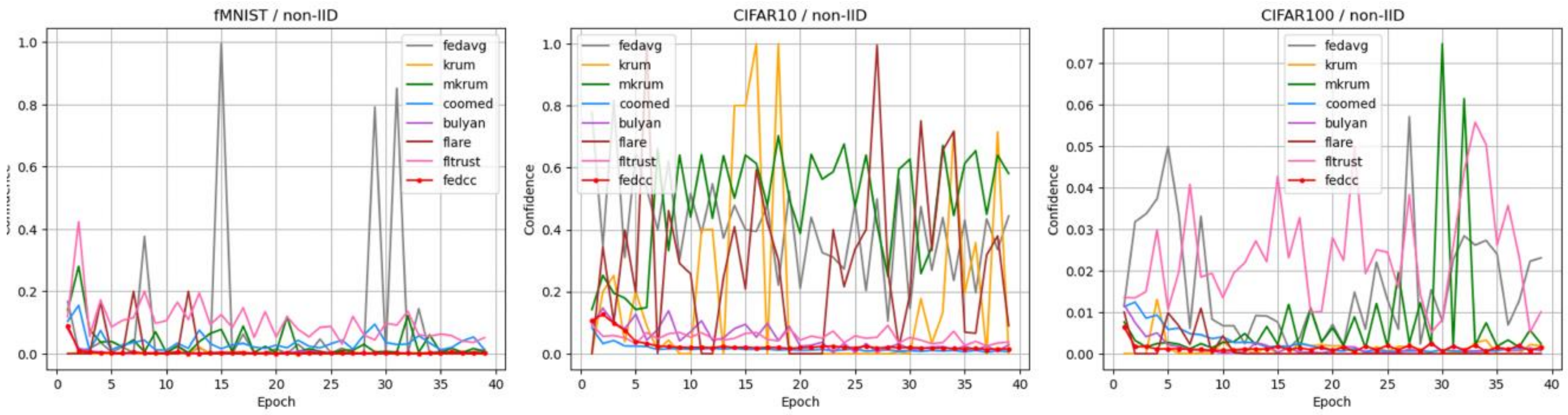
Table 3: Test Accuracy under untargeted attacks in Non-IID setting.

Case	data	FedAvg	Krum	MKrum	Coomed	Bulyan	FLTrust	FLARE	FedCC
Fang	fM	57.14	16.51	45.88	57.12	13.39	60.8	49.54	**71.13**
-Krum	C10	33.69	15.38	20.5	35.7	19.23	41.87	17.03	**52.06**
non-IID	C100	2.27	1	4.95	7.85	0.98	11.04	7.46	**14.51**
Fang	fM	16.32	49.33	66.84	68.9	64.12	18.96	52.25	**72.76**
-Med	C10	10.02	25.06	45.44	40.23	32.47	10	14.59	**47.85**
non-IID	C100	1	6.24	14.52	10.27	6.91	1.09	1	**16.12**

Results: targeted Attacks (Non-IID)

- FedCC reduces **backdoor confidence to near zero**
- Also maintains high main task accuracy
- DBA (distributed backdoor) handled effectively

Case	data	FedAvg	Krum	MKrum	Coomed	Bulyan	FLTrust	FLARE	FedCC
Target non-IID	fM	75.65	45.27	65.97	71.70	57.96	61.82	64.31	**75.66**
	C10	36.16	14.98	30.72	48.97	40.11	44.06	10.18	**51.56**
	C100	4.46	6.18	6.90	12.04	11.16	12.95	1.14	**15.26**
DBA	C10	38.56	24.94	7.09	44.45	34.19	51.49	38.73	**52.28**



Results: IID Setting

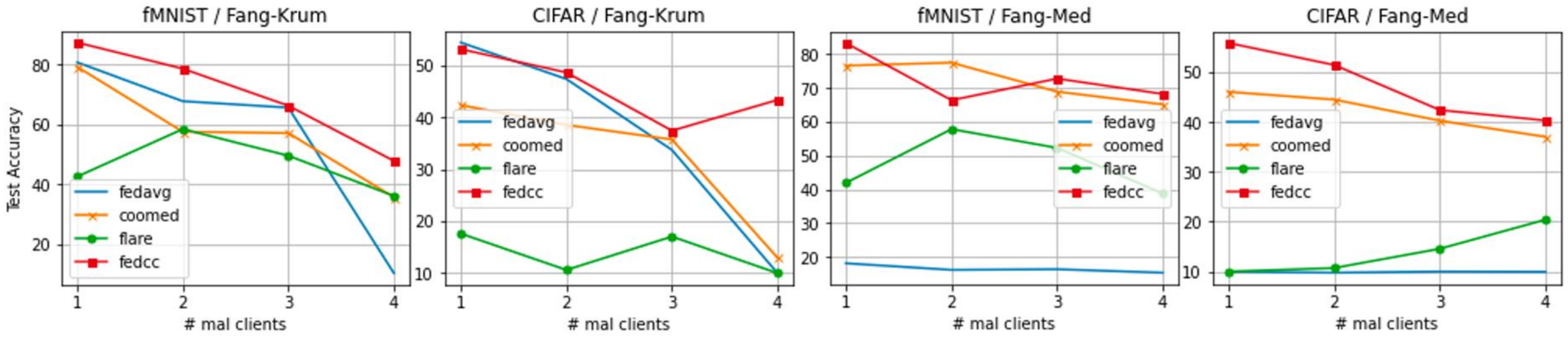
- FedCC also outperforms others under IID
- Indicates generalizability

Case	data	FedAvg	Krum	MKrum	Coomed	Bulyan	FLTrust	FLARE	FedCC
Fang -Krum IID	fM	75.55	31.66	87.78	87.62	50.30	89.53	79.16	**89.57**
	C10	49.67	40.86	63.42	57.40	12.67	68.25	25.77	**69.84**
	C100	13.72	1.04	7.64	6.17	1.59	17.14	7.49	**18.47**
Fang -Med IID	fM	20.86	85.33	89.53	86.70	87.45	21.36	71.08	**89.66**
	C10	9.51	54.28	69.68	59.20	57.69	9.92	49.82	**70.52**
	C100	0.87	12.27	16.52	14.43	12.56	1.16	5.83	**17.83**

Case	data	FedAvg	Krum	MKrum	Coomed	Bulyan	FLTrust	FLARE	FedCC
Target IID	fM	88.27	86.63	87.03	89.41	89.45	89.59	75.29	**90.01**
	C10	64.68	57.69	71.19	69.85	68.76	68.61	11.09	**71.64**
	C100	13.83	5.72	17.57	13.24	15.11	17.08	1.03	**18.61**
DBA	C10	10.00	35.58	51.40	10.00	16.16	56.69	10.00	**58.04**

Results: Robustness and Scalability

- Varying numbers of attackers
- Different participation rates



(a) Fang-Krum (b) Fang-Med

		Untargeted-Krum				Untargeted-Med			
Frac	**Data**	**FedAvg**	**Med**	**FLARE**	**FedCC**	**FedAvg**	**Med**	**FLARE**	**FedCC**
0.1	fM	55.31	49.83	34.02	**64.83**	16.57	66.41	52.24	**69.52**
	C10	10.06	**22.55**	14.50	20.49	10.00	15.33	10.00	**29.81**
0.3	fM	64.22	57.52	10.00	**73.55**	16.26	58.07	10.00	**61.12**
	C10	24.24	12.59	10.00	**27.81**	10.98	22.61	10.00	**38.27**
0.5	fM	62.37	58.20	10.00	**76.41**	18.36	62.49	10.00	**69.05**
	C10	23.99	17.28	10.06	**34.32**	9.87	27.83	10.00	**37.27**

Comparison Summary

Criteria	FedCC	Krum	Coomed	FLARE
Non-IID Robustness	✅	❌	⚠️	❌
No data access	✅	✅	✅	❌
Backdoor defense	✅	⚠️	✅	⚠️
Threshold-free	✅	❌	❌	❌

Limitations & Future Work

- Only tested on CNNs and small datasets
- Assumes homogeneous models
- CKA computation is not lightweight
- No formal guarantees (only empirical + theoretical insight)

Conclusion

- **FedCC** introduces a new aggregation method using **CKA over PLRs**
- Robust to both untargeted and backdoor attacks
- Especially effective under **non-IID**, which is common in practice

Thank You

Questions to hjeong@umass.edu

https://github.com/HyejunJeong/FedCC