Paper ID: 5008

INTERNATIONAL JOINT CONFERENCE ON NEURAL NETWORKS
IJCNN2025
30 JUNE - 5 JULY 2025 | ROME, ITALY
INTERNATIONAL NEURAL NETWORK SOCIETY

FedCTTA: A Collaborative Approach to Continual Test-Time Adaptation in Federated Learning

Rakibul Hasan Rajib*, Md Akil Raihan Iftee*, Mir Sazzat Hossain*, A.K.M. Mahbubur Rahman*, Sajib Mistry†, M. Ashraful Amin*, Amin Ahsan Ali*

**Center for Computational & Data Sciences, Independent University, Bangladesh*
†Curtin University, Australia

International Joint Conference on Neural Networks (IJCNN) - 2025

Presenter: Rakibul Hasan Rajib

Motivation

⏩**Federated Learning(FL)**

- Collaborative model training across clients
- No raw data sharing → ensures privacy

⏩**Challenge:** Performance degradation due to distribution shifts

Test-time Adaptation (TTA) Offers a Promising Solution

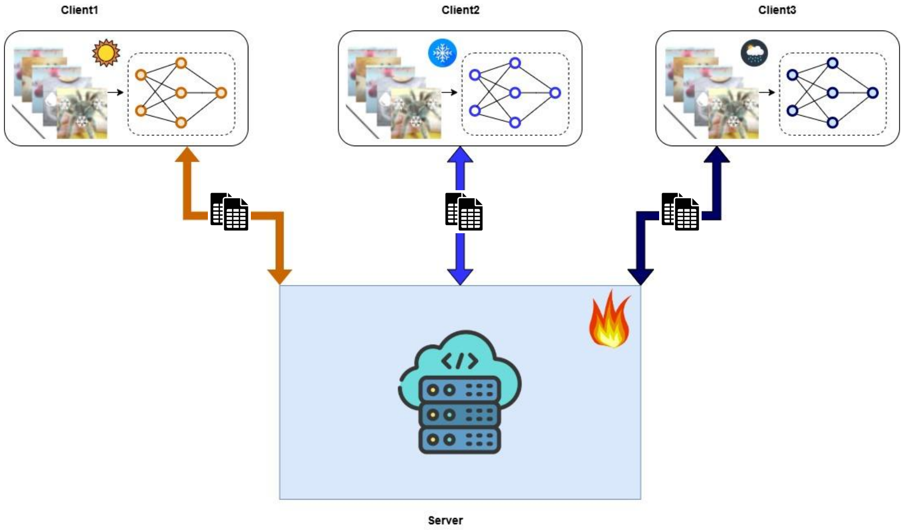
- Models adapt using only test samples
- Adapts to distribution shifts at inference

⏩Challenges of TTA in FL

- Heterogenous and evolving distributions
- Privacy risks from feature sharing
- Scalability issues

⏩Proposed Method

- **FedCTTA** – a privacy-preserving and computationally efficient framework for continual test-time adaptation



Limitations of Prior Work

FedICON:

- High computational demands

ATP:

- Assumes static test-time distributions
- No inter-client knowledge sharing

FedTHE+:

- Struggles with severe out-of-distribution (OOD) data

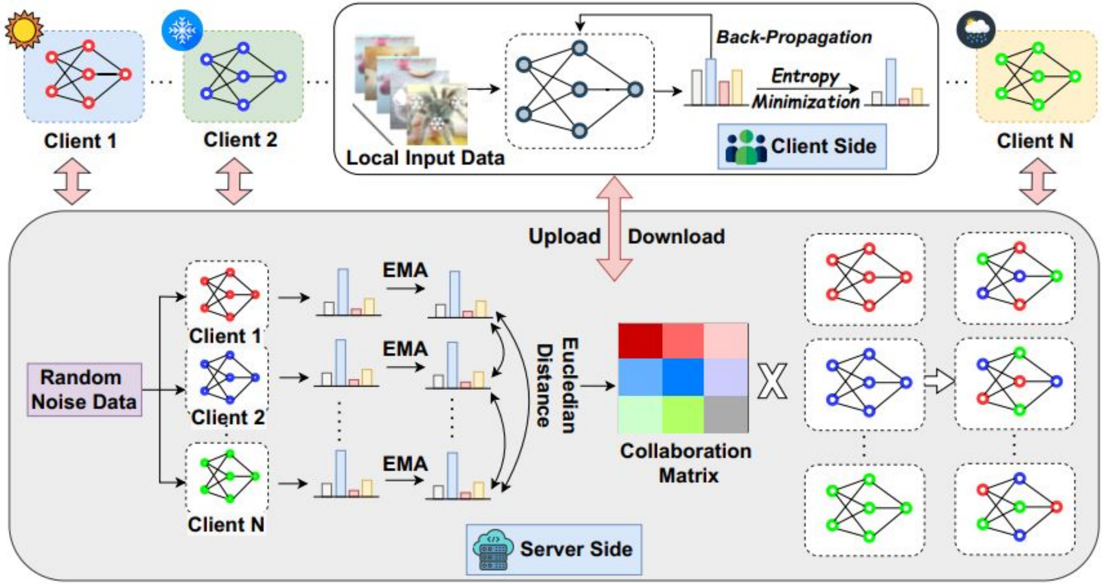
FedTSA:

- Privacy risks from sharing local feature stats
- Requires server-side learning
- Scalability issues due to memory bank overhead

Key Contributions

- **Similarity-aware aggregation based on functional similarity**
- **No sharing of local feature embeddings, ensuring data security and mitigating privacy risks**
- **Eliminates server-side training, reducing computational overhead**
- **Constant memory footprint, enabling scalability to many clients**

Federated Continual Test-Time Adaptation (FedCTTA)



- Local adaptation via entropy minimization or BN statistics updates
- Server computes similarity using model outputs on random noise samples
- Personalized aggregation without sharing raw data or features

Federated Continual Test-Time Adaptation (FedCTTA)

⏩Client Side - Local Adaptation

- TTA-grad: minimizes entropy and updates all model parameters

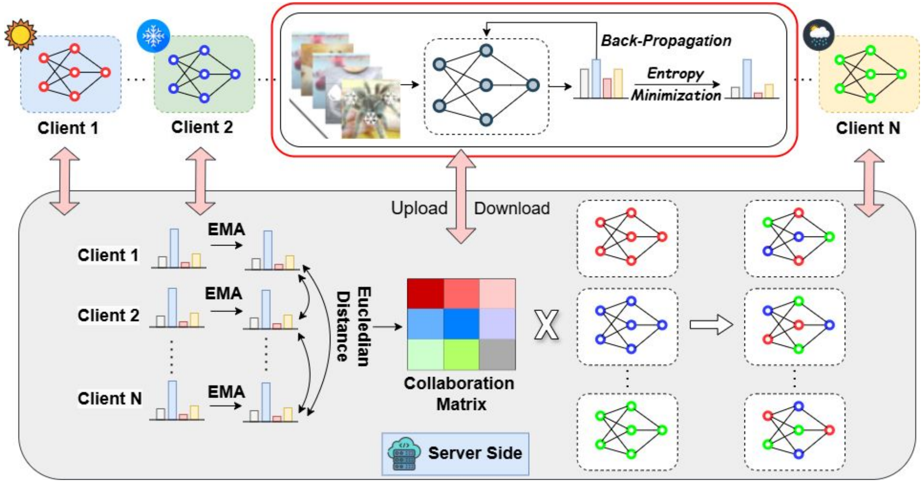
$$H(p) = -\sum_{k=1}^{K} p_k \log(p_k)$$

$$\mathcal{L}_{\text{ent}} = \frac{1}{|\mathcal{D}_t^{(i)}|} \sum_{x \in \mathcal{D}_t^{(i)}} H(f_{\theta_i}(x))$$

- TTA-bn: updates only BatchNorm activation statistics without requiring backpropagation

$$\mu_i^{\text{new}} = \mathbb{E}_{x \sim \mathcal{D}_t^{(i)}}[x]$$

$$\sigma_i^{2,\text{new}} = \text{Var}_{x \sim \mathcal{D}_t^{(i)}}(x)$$



Federated Continual Test-Time Adaptation (FedCTTA)

⏩Server Side - Similarity-aware Aggregation

- Server aggregates models based on functional similarity.
- For each client i, server computes mean logits using random noise samples

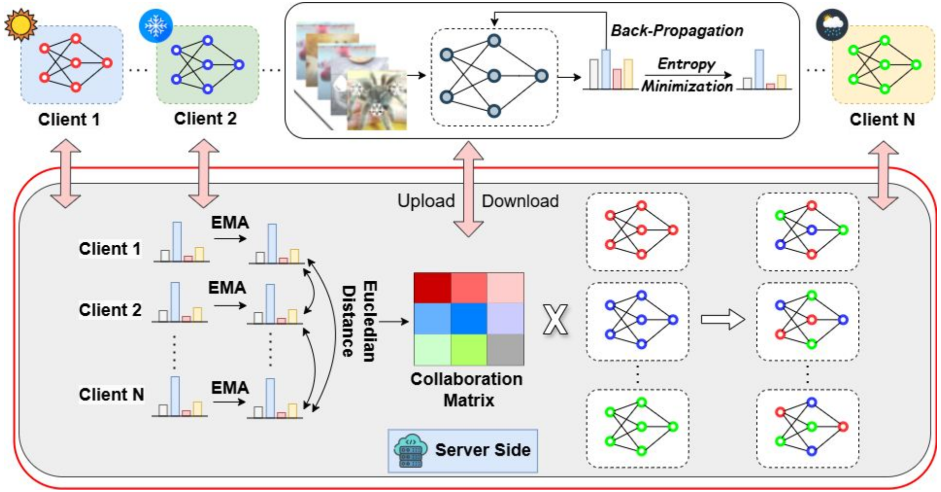
$$\mu_i = \frac{1}{M}\sum_{k=1}^{M} f_{\theta_i}(z_k)$$

- For clients i and j, similarity is:

$$D_{ij} = -\|\mu_i - \mu_j\|_2$$

- New model for client i using weighted aggregation:

$$\theta_i^{\text{new}} = \sum_{j=1}^{K} \frac{\exp(D_{ij})}{\sum_{k=1}^{K}\exp(D_{ik})}\theta_j$$



Experimental Setup

- **Datasets:** CIFAR10-C and CIFAR100-C (15 corruptions, 5 severity levels; results at severity 5)
- **Models:** Pretrained ResNeXt-29 (CIFAR100-C) and ResNet-8 (CIFAR10-C).
- **FL Setting:** 20 clients, streaming test data in batches of 10.
- **TTA Setups:** TTA-grad and TTA-bn

⏩ Heterogeneity Simulation

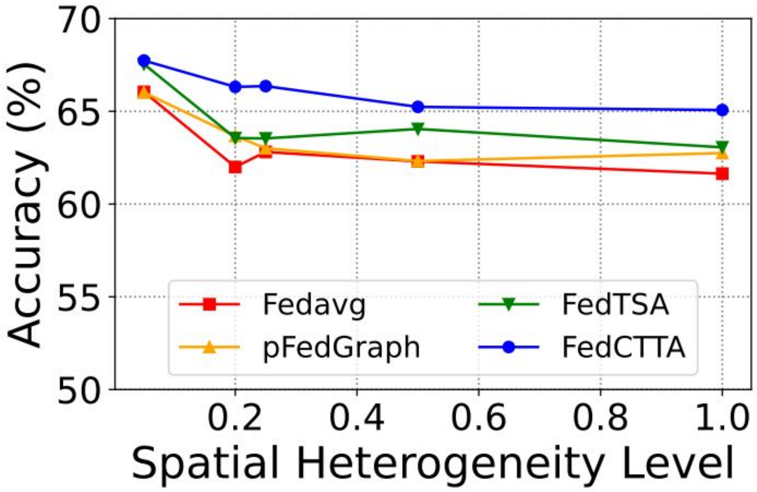
- **Spatial Heterogeneity (SH_t):** Measures diversity among client data distributions.
  - NIID: SH_t = 0.2 (4 clusters)
  - IID: SH_t = 0.05 (single cluster)
- **Temporal Heterogeneity (TH_i):** Measures frequency of distribution changes in streaming data.
  - Constant at 0.02 for both scenarios.

Results

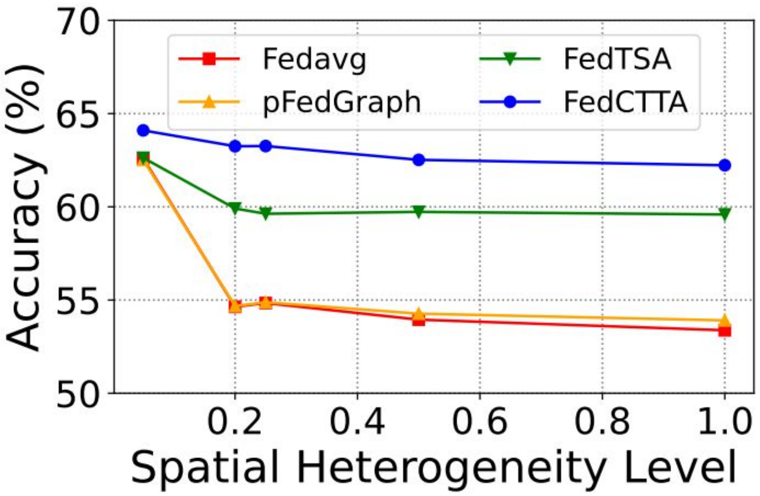
Method	NIID				IID			
	CIFAR10-C		CIFAR100-C		CIFAR10-C		CIFAR100-C	
	TTA-grad	TTA-bn	TTA-grad	TTA-bn	TTA-grad	TTA-bn	TTA-grad	TTA-bn
No-Adapt	58.47±0.19	58.61±0.17	30.22±0.12	30.22±0.12	58.64±0.22	58.55±0.21	30.22±0.12	30.22±0.12
Local	63.82±0.31	64.65±0.29	52.85±0.32	55.99±0.34	63.96±0.33	64.79±0.31	52.94±0.31	56.05±0.34
FedAvg	61.15±0.24	61.45±0.23	51.63±0.17	57.13±0.43	66.12±0.26	67.41±0.27	62.54±0.31	63.96±0.31
FedAvg+FT	63.82±0.27	61.45±0.23	47.83±0.58	57.13±0.43	63.79±0.30	67.41±0.27	61.72±0.59	63.96±0.31
FedProx	61.68±0.22	61.45±0.23	53.00±0.38	57.13±0.43	66.12±0.24	67.41±0.27	62.33±0.67	63.96±0.31
FedAvgM	61.50±0.25	61.37±0.19	52.31±0.46	57.13±0.43	63.60±0.28	67.41±0.27	54.66±0.27	63.96±0.31
MOON	61.58±0.23	61.45±0.23	54.26±0.27	57.13±0.43	66.05±0.25	67.41±0.27	62.40±0.23	63.96±0.31
pFedSD	61.31±0.21	61.45±0.23	53.33±0.37	57.13±0.43	66.14±0.26	67.41±0.27	62.32±0.33	63.96±0.31
pFedGraph	62.38±0.26	64.21±0.25	57.01±0.38	58.73±0.38	66.10±0.29	64.42±0.28	62.48±0.30	58.75±0.63
LDAWA	61.85±0.23	61.45±0.23	53.61±0.33	57.13±0.43	65.92±0.26	67.41±0.27	62.37±0.41	63.96±0.31
FedTSA	63.39±0.27	66.19±0.26	58.03±0.38	62.93±0.29	66.29±0.28	67.51±0.27	62.62±0.36	63.70±0.34
FedCTTA	**66.23±0.28**	**66.50±0.27**	**64.81±0.29**	**63.39±0.28**	**66.64±0.29**	**67.78±0.28**	**64.15±0.28**	**64.52±0.28**

- FedCTTA consistently outperforms FedTSA and state-of-the-art FL methods
- FedCTTA consistently achieves higher accuracy while preserving privacy.

Robustness to Spatial Heterogeneity



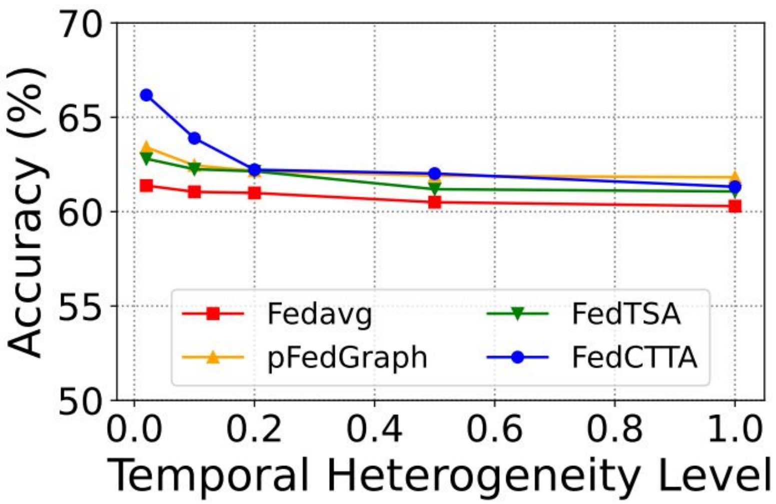
a) CIFAR-10C



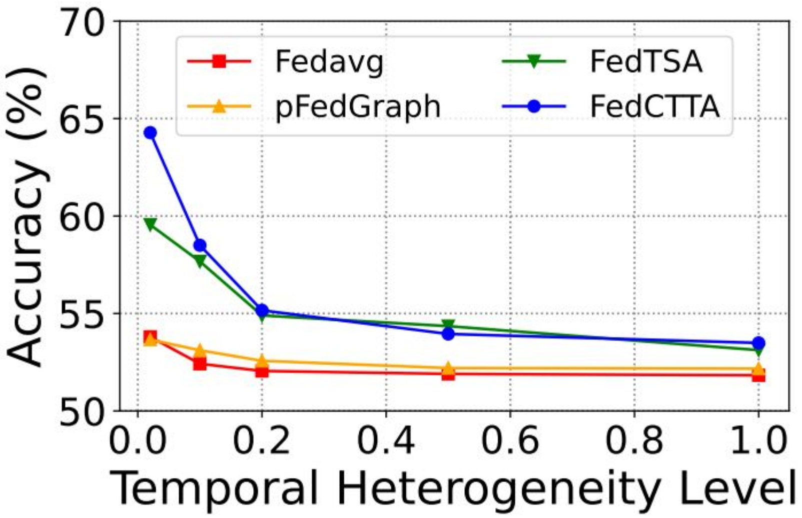
b) CIFAR-100C

- Accuracy declines for all methods with increasing SH_t
- FedAvg shows the steepest drop. FedCTTA shows minimal performance degradation
- Demonstrates strong adaptability to non-IID client distributions

Robustness to Temporal Heterogeneity



c) *CIFAR-10C*

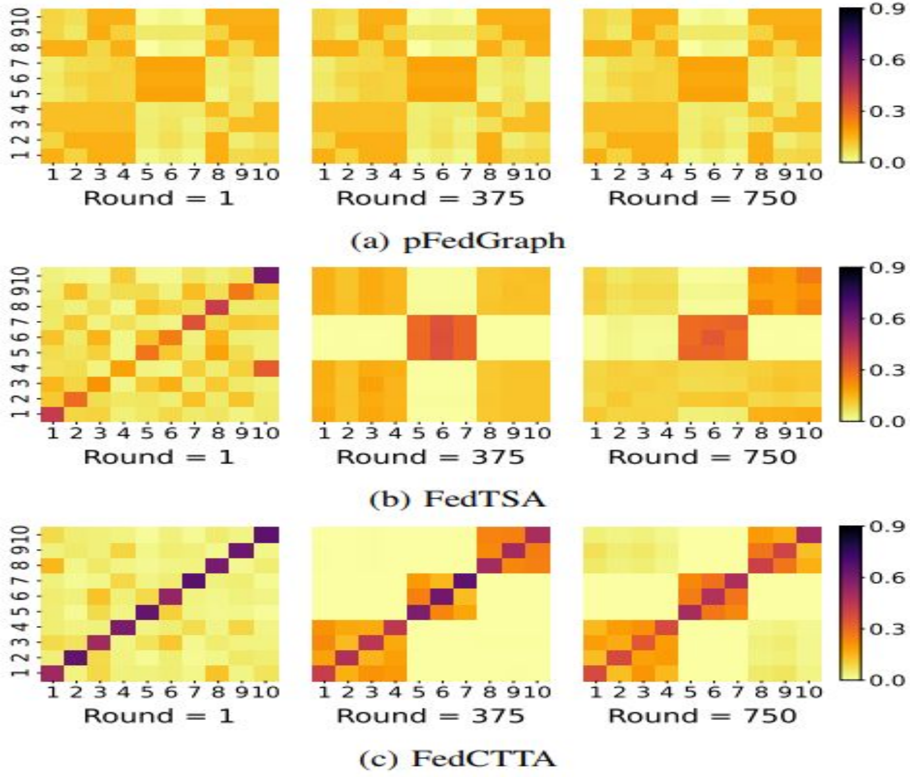


d) *CIFAR-100C*

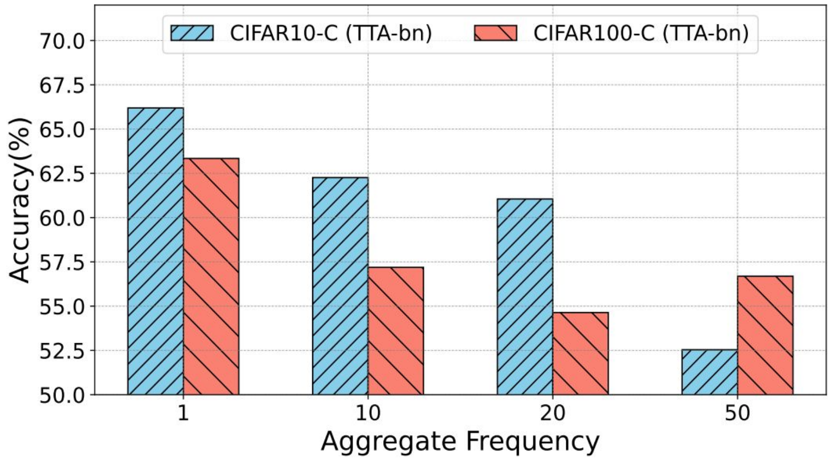
- Strong Robustness against temporal heterogeneity
- Strength lies in adaptive aggregation using temporal similarity
- Performs best when temporal shifts are gradual

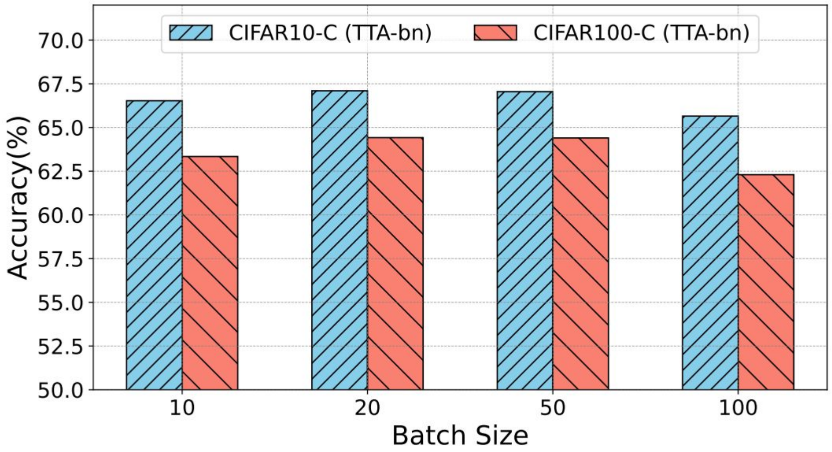
Collaboration Matrix Analysis

- 10 clients (CIFAR10-C), 3 groups by shift sequence.
- **pFedGraph:** Scattered, unstructured collaboration
- **FedTSA:** Initially self-focused, weak clustering
- **FedCTTA:** Naturally forms structured, adaptive clusters



Ablation Study - Aggregate Frequency & Batch Size





⏩Aggregation Frequency:

- Higher interval negatively impacts accuracy.
- Frequent updates are crucial for performance.

⏩Batch Size:

- Very low (10) or very high (100) sizes are suboptimal. Moderate sizes (20, 50) yield best results.
- **Too large**: frequent shifts, reduced performance.
- **Too small**: unstable updates.

Similarity Metrics & Auxiliary Data

TABLE IV: Comparison of test accuracy using distance measures for output logits and feature embeddings on CIFAR10-C dataset with the TTA-grad method under the NIID setting.

Data	Output Logit (Acc. %)				Feature (Acc. %)	
	Euclid	KL-div	CE	Cosine	Euclid	Cosine
Random Noise	66.19	61.62	61.60	61.62	62.07	61.63
Selected (CIFAR)	65.92	61.65	61.64	61.63	61.80	61.63

⏩**Optimal Combination**: Output logits from random noise samples + Negative Euclidean Distance.

Conclusion

- Enables adaptive inter-client collaboration without sharing raw data or features
- Demonstrates robust performance under spatial and temporal heterogeneity
- Efficient, scalable, and privacy-preserving

Thank You

Contact: rakibul@iub.edu.bd