ReSoFed: Reliability-Guided Model Souping for Robust Federated Learning in Heterogeneous Classroom Environments

Muhammad Rafsan Kabir, Md Shopon, Marina L. Gavrilova
University of Calgary, Canada

BTLab

Code

CVPR JUNE 3-7, 2026 DENVER COLORADO

Introduction

Research Problems

- Privacy regulations prevent educational institutions from sharing centralized classroom video data, requiring federated learning solutions.
- Heterogeneous classroom environments (camera quality, lighting, viewpoints) create non-IID data distributions that degrade federated learning performance [1].
- Standard aggregation methods like FedAvg [2] equally weight all clients, leading to negative transfer from poorly generalizing models.

Main Contributions

- **Proposed ReSoFed:** A privacy-preserving, reliability-guided federated aggregation framework using a heterogeneous server-side validation set to estimate client generalization.
- **Two-Stage Aggregation:** Greedy model-soup–based client selection followed by reliability-aware weighted aggregation to suppress poorly generalizing updates and prioritize stronger ones.
- **Performance:** Consistent improvements over FedAvg across six backbone architectures on the SCB dataset while preserving privacy and compatibility with standard FL pipelines.

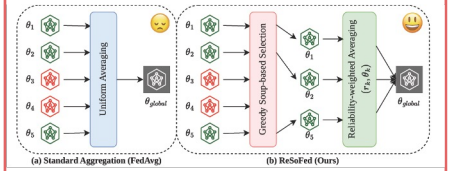


Fig. 1: Comparison between (a) FedAvg and (b) the ReSoFed framework.

Acknowledgments

Authors acknowledge the support from NSERC, Alberta Innovates, NFRF, and Research Excellence Chair in Trustworthy AI.

rafsan.kabir@ucalgary.ca md.shopon@ucalgary.ca mgavrilo@ucalgary.ca

Proposed Methodology

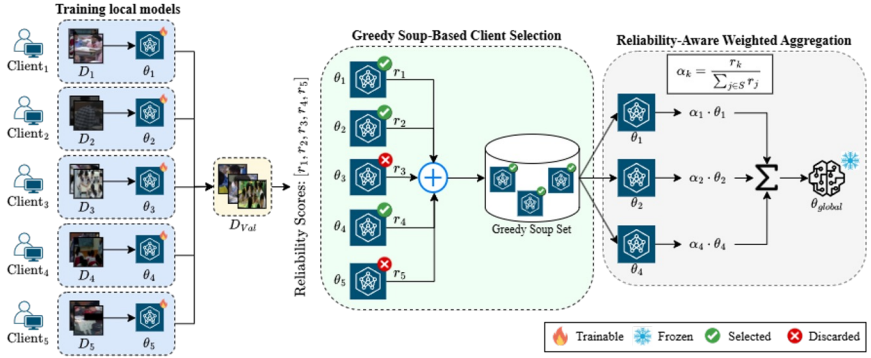


Fig. 2: Overview of the proposed ReSoFed framework.

Cross-Domain Reliability Estimation

Each client model is evaluated on the server-held validation set to compute a reliability score, measuring how well the client generalizes beyond its local domain.

Greedy Soup-Based Client Selection

The soup set is initialized with the most reliable client; remaining clients are iteratively added only if their inclusion improves cross-validation performance.

Reliability-Aware Weighted Aggregation

Selected clients in the set are aggregated using reliability-proportional weights, resulting in a global model with stronger cross-domain generalization.

By combining greedy soup-based client selection with reliability-weighted aggregation, ReSoFed effectively filters poorly generalizing clients while prioritizing models that demonstrate stronger cross-domain reliability.

Results & Analysis

Method	CNN Models					Transformer-based Models				
	Model	Acc (%)	Pre (%)	Rec (%)	F1 (%)	Model	Acc (%)	Pre (%)	Rec (%)	F1 (%)
Standard FL (FedAvg)	ResNet-50	73.82	77.34	73.67	72.83	ViT B-16	80.47	80.75	80.36	80.22
	EfficientNet-V2-S	78.93	80.31	78.79	78.45	Swin-V2-Tiny	81.23	82.56	81.10	80.80
	ConvNeXt-Base	81.82	81.97	80.78	81.37	Swin-V2-Base	81.96	83.26	81.73	81.41
ReSoFed	ResNet-50	78.50	78.38	78.44	78.06	ViT B-16	81.00	81.18	80.91	80.85
	EfficientNet-V2-S	81.61	81.49	81.57	81.22	Swin-V2-Tiny	81.68	83.01	81.58	81.18
	ConvNeXt-Base	**82.75**	**83.00**	**82.68**	**82.35**	Swin-V2-Base	**82.35**	**82.55**	**82.29**	**82.16**

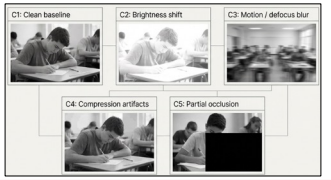
Table 1: Performance comparison between standard federated learning (FedAvg) and the proposed ReSoFed framework on the SCB dataset. Results are reported across CNN and transformer architectures. **Bold** values indicate the best performance within each architecture group.

Experiments

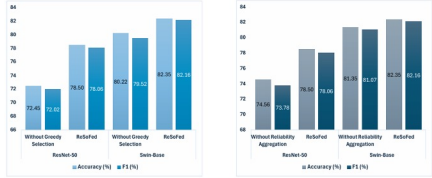
Experimental Setup

- **Dataset:** Student Classroom Behavior (SCB) ~90,000 samples across 14 categories
- **CNN Backbones:** ResNet-50, EfficientNet-V2-S, ConvNeXt-B
- **Transformer Backbones:** Swin-V2-T, Swin-V2-B, ViT B-16
- **Classification Metrics:** Accuracy, Precision, Recall, F1 Score

Heterogeneity Simulation



Ablation Study



Conclusion & Future Work

- ReSoFed mitigates negative transfer in heterogeneous federated learning via greedy model-soup client selection and reliability-aware aggregation, preserving data privacy throughout.
- Experiments on SCB dataset show consistent performance gains over FedAvg across multiple CNN and transformer backbones.
- Future work will explore multi-round federated learning, advanced reliability estimation, and evaluation on real-world settings.

References

[1] Qu et al. Rethinking architecture design for tackling data heterogeneity in federated learning. In *CVPR*, 2022.
[2] McMahan et al. Communication-efficient learning of deep networks from decentralized data. In *AISTATS*, 2017.