University of Glasgow
School of Computing Science

ICDCS 2025 GLASGOW

Knowledge & Data Engineering Systems

# Resilient Inference for Personalized Federated Learning in Edge Computing Environments


***Ke Xiao, Qiyuan Wang, Christos Anagnostopoulos, Kevin Bryson***

**School of Computing Science, University of Glasgow**



**Abstract:** Federated Learning (FL) and Edge Computing (EC) enable distributed learning systems that prioritize data privacy and low latency. Personalized FL (PFL) enhances this paradigm by tailoring models to individual participants. **However, when edge servers fail, existing task rescheduling methods often ignore the impact of inherent model differences (a byproduct of PFL) on inferential tasks, leading to suboptimal performance**.

To address this, we propose **SOIR**, a framework that integrates model similarity into the rescheduling process. We formulate the inference task rescheduling problem as a Mixed Integer Nonlinear Programming model and introduce an efficient algorithm to solve it. Experimental results demonstrate that SOIR is both applicable and effective in FL-based resilient edge environments.


## Introduction

**Background:** FL and EC are cutting-edge paradigms for distributed learning systems that ensure data privacy and low-latency communication.

- ✓ FL enables collaborative model training while preserving local data privacy.
- ✓ EC brings cloud capabilities to the network edge, reducing latency for real-time apps.
- ✓ PFL personalizes models per participant, enhancing local inference performance.

**Challenges:**

- ✓ Node Vulnerability: Resource-constrained edge nodes prone to hardware/software failures → threatens service reliability.
- ✓ Resilience-Accuracy Tradeoff: Failed nodes require task rescheduling to surrogates, BUT PFL personalization causes: system metrics-only selection (e.g., latency/energy) → Model mismatch → Accuracy plummets.

**Our Contribution:** Existing edge task schedulers ignore model discrepancies' impact on inference quality. We propose SOIR:

- ✓ First framework to integrate model similarity into inference rescheduling optimization → Ensures EC system efficiency + high-accuracy resilience for PFL.

## Observation

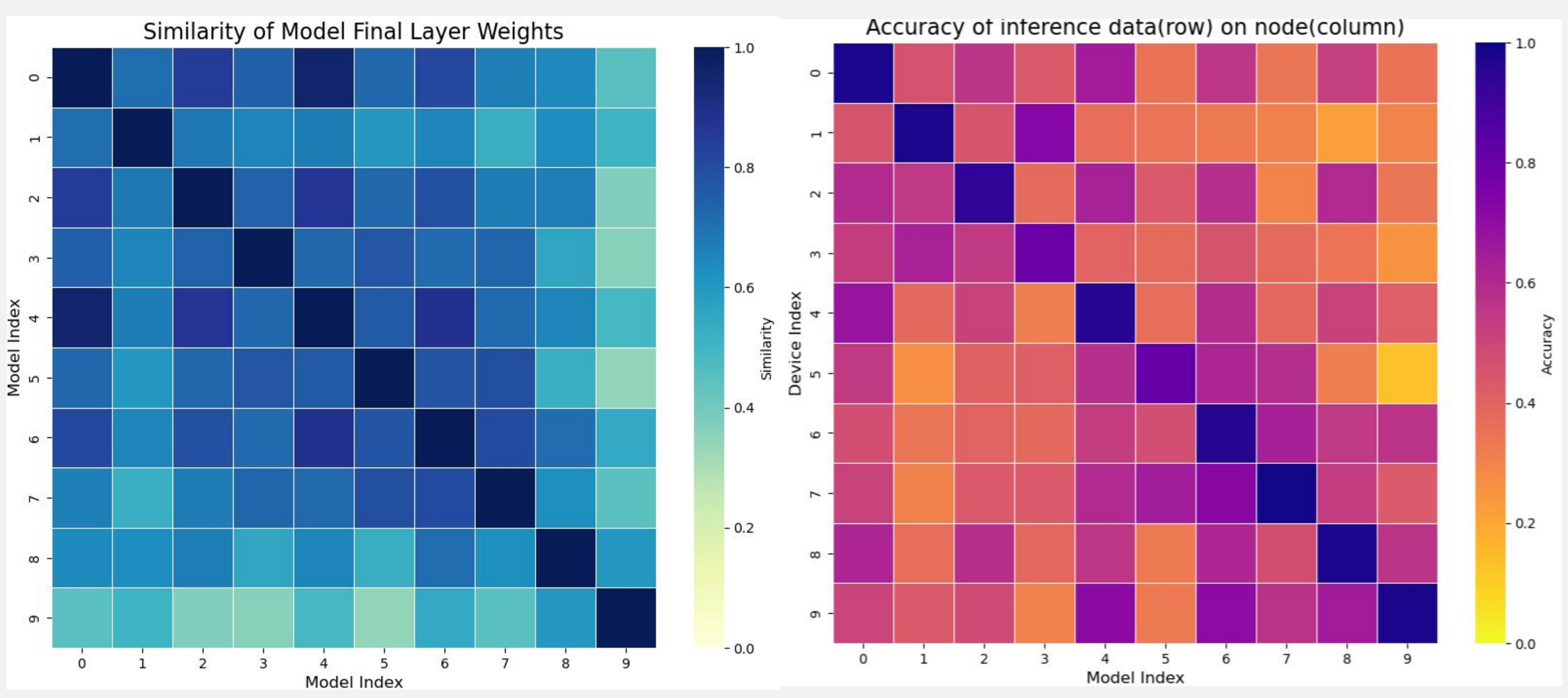


Figure 1 The pairwise similarity (a) and inference accuracy (b) between personalized DL models

**The Link Between Model Similarity & Inference Accuracy:**

**Experimental Setup:** We trained 10 distinct personalized models and analyzed the relationship between their similarity and their inference accuracy on each other's test datasets. (Darker colors indicate higher similarity/accuracy)

**Key Finding:** As illustrated in *Fig. 1*, we observed a significant positive correlation.

- ✓ *Fig.1a* shows the pairwise similarity between the classifiers of the 10 models.
- ✓ *Fig.1b* shows the inference accuracy of these models on each other's corresponding test data.
- ✓ It is clear that when the similarity between two models is high (a dark square in *Fig.1a*), the inference accuracy of their test data also tends to be high (a corresponding dark square in *Fig.1b*).

## Problem Formulation

We formulate resilient inference rescheduling as a Mixed-Integer Nonlinear Programming problem. The objective is to derive an optimal policy assigning surrogate models to failed devices, minimizing holistic cost (system + model metrics) under operational constraints.

**Objective:**

- ✓ **System Costs:** Includes Data Transmission Latency, Inference Latency, Transmission Energy, and Inference Computation Energy.
- \+ **Node Load:** A cost that penalizes selecting nodes with fewer available resources.
- \+ **Model Dissimilarity:** CKA-based classifier divergence

**Constraints:**

- ✓ **Assignment:** Each task → exactly one surrogate
- ✓ **Latency SLO:** Total delay ≤ device tolerance
- ✓ **Energy Cap:** Transmission energy ≤ device budget
- ✓ **Capacity:** Task demand ≤ surrogate resources

## Methodology

**SOIR Algorithm:** This rolling optimization heuristic prioritizes tasks by SLO urgency, then iteratively assigns each to the surrogate node minimizing combined system costs (latency/energy) and model dissimilarity (CKA-based) while satisfying operational constraints, with dynamic resource updates between assignments.

## Experiments & Results

**Experimental Setup:** We simulated a 10-node edge environment with 20 devices, using non-i.i.d. CIFAR-10-trained EfficientNet-B0 models to reflect data heterogeneity, and compared against three baselines: Similarity Only (SO), System Cost Only (SCO), and Random surrogate selection.

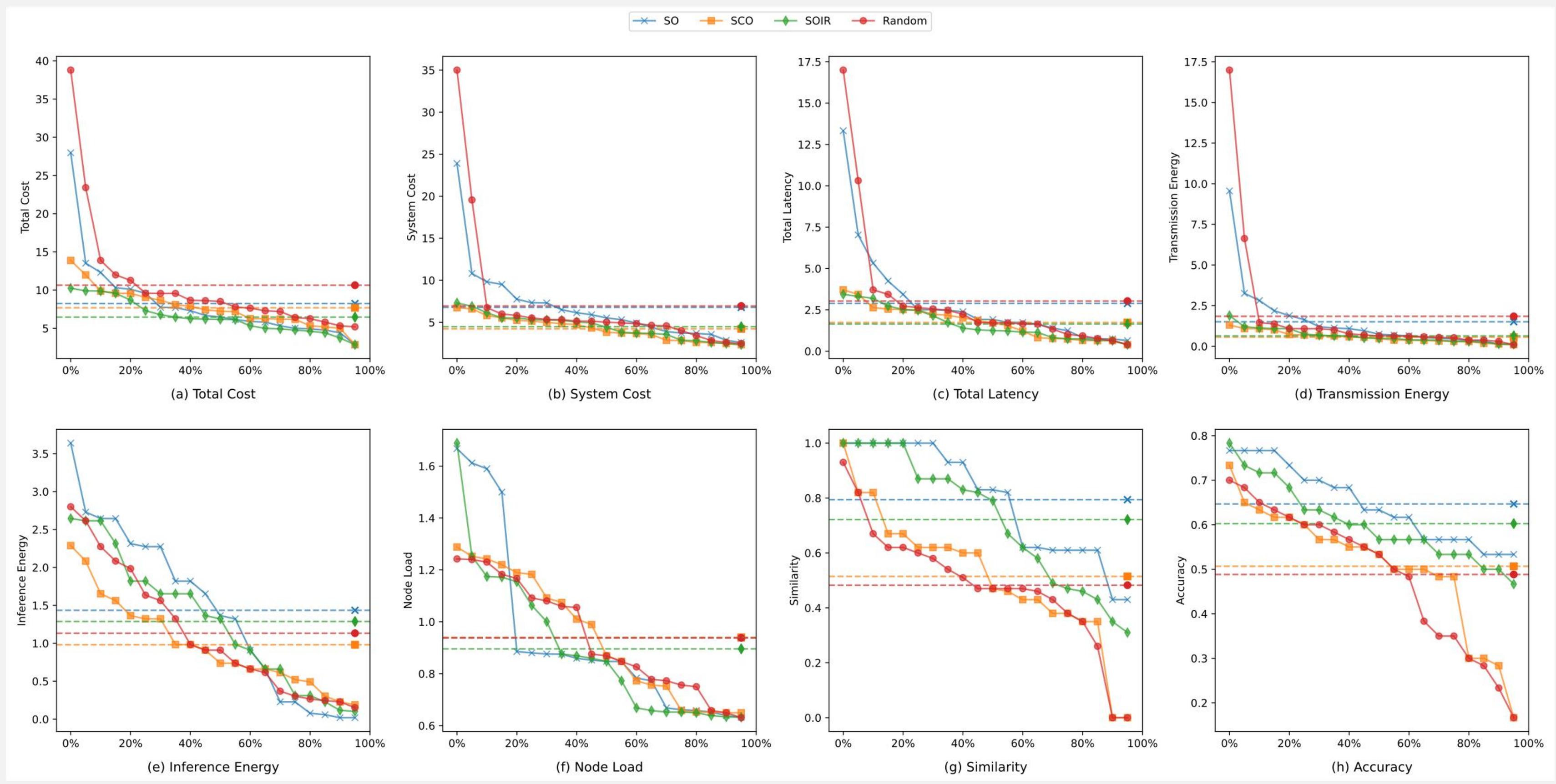


Figure 2 Performance of SO, SCO, Random and SOIR on different metrics

## Contact Information


Ke XIAO, BSc, MSc, PhD Candidate
University of Glasgow
Email: **k.xiao.1@research.gla.ac.uk**

School of Computing Science | University of Glasgow
S114 Sir Alwyn Williams Building | G12 8RZ