Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 4, 2026

WHEN CLIENTS DRIFT: FEDERATED SLA-RISK FORECASTING ACROSS UNSEEN 6G RAN REGIMES

Paras Mangi[*1], Sadaf Bibi[2], Ali Nawaz[3], Sadia Bibi[4]

[*1]*Institute of Computer Science, Shah Abdul Latif University, Khairpur Mirs, Sindh, Pakistan*
[2]*Department of Artificial Intelligence, Aror University of Art, Architecture, Design & Heritage, Sukkur, Sindh, Pakistan*
[3]*Department of Electrical Engineering Technology, The Benazir Bhutto Shaheed University of Technology & Skill Development, Khairpur Mirs, Sindh, Pakistan*
[4]*Department of Software Engineering, Comera LLC, Abu Dhabi, UAE*

[*1]72scholar.paras110@gmail.com, [2]f24ari184@aror.edu.pk, [3]engralitechno@gmail.com, [4]sadia.bibi@mycomera.com

DOI: https://doi.org/10.5281/zenodo.19723844

Keywords
Federated learning, 6G radio access networks, SLA risk forecasting, client heterogeneity, unseen-regime robustness.

Article History
Received: 28 February 2026
Accepted: 07 April 2026
Published: 24 April 2026

Copyright @Author
Corresponding Author: *
Paras Mangi

Abstract
In 6G radio access networks, the service patterns and operating environments of different clients vary significantly, posing a major challenge to the generalization of federated SLA risk prediction. This paper investigates federated next-window SLA risk prediction under unseen regimes using a leave-one-regime-out evaluation setting. We propose a federated method that combines client profile information and regime-aware aggregation to model cross-layer 6G RAN telemetry data. Experimental results show that centralized models remain stronger in AUROC, with Centralized-GRU reaching 0.882 and HistGB reaching 0.872; however, the proposed method performs best on metrics that more closely reflect practical decision quality, achieving the highest average balanced accuracy of 0.676, macro balanced accuracy of 0.676, and the best worst-regime F1 score of 0.485. Compared with vanilla FedProx-GRU, the method improves AUROC, F1, and balanced accuracy. Ablation experiments further show that client profiling and regime-aware weighting both contribute to robust prediction under unseen regimes. These results suggest that although the method is not the strongest overall ranker, it is more practically valuable for robust federated decision-making in unseen 6G RAN environments.

1. Introduction
With the development of 6G and AI-native RAN, the network side increasingly relies on intelligent control, continuous monitoring, and rapid closed-loop optimization. For operators, it is no longer sufficient to respond only after an anomaly occurs; the ability to identify potential SLA risks in advance is important for service stability, resource scheduling, and user experience [1], [2]. Existing research shows that prediction methods based on KPI, QoS, and multi-layer telemetry data can support performance estimation, QoS prediction, and fault warning [3], [4]. However, most methods assume that training and testing data come from similar operating conditions, so performance under unseen regimes remains unclear. This is especially important for deployment because the RAN environment itself is dynamic [2], [4].

At the same time, network data are naturally distributed, and clients differ substantially in service load, wireless conditions, and behavior patterns. Federated learning offers an attractive direction, but non-IID data, statistical heterogeneity, and client drift can all affect training stability and generalization [5]–[9]. Especially when distribution shift or concept drift occurs, conventional federated methods can degrade under unseen conditions [6], [9].

Based on this motivation, this paper studies federated next-window SLA risk prediction on cross-layer 6G RAN telemetry using a leave-one-regime-out protocol, where one regime is held out as an unseen test environment. The proposed portrait-conditioned FedProx-GRU combines client profile information with regime-aware aggregation to improve federated robustness in heterogeneous environments. Results show that while centralized baselines remain stronger on AUROC, the proposed method achieves the best average balanced accuracy, macro balanced accuracy, and worst-regime F1, indicating stronger thresholded decision-making under unseen regimes. Recent operational screening work in other domains also highlights that strong ranking alone is not sufficient and that calibrated, auditable decision support is important in practice [11], [12].

The main contributions of this paper are threefold. First, it formulates federated next-window SLA risk prediction on cross-layer 6G telemetry as a leave-one-regime-out generalization problem under unseen regimes. Second, it designs a portrait-conditioned FedProx-GRU with regime-aware aggregation to model client heterogeneity. Third, it shows experimentally that although centralized models remain stronger on AUROC, the proposed method performs best on balanced accuracy and worst-regime F1, making it more suitable for robust federated decision-making under unseen regimes.

2. Related Work

2.1 SLA/QoS/Telemetry-Driven Network Prediction

Recent work on intelligent networks for 5G/6G and O-RAN increasingly uses KPI, QoS, and multi-layer telemetry data to support service assurance, performance optimization, and fault prediction [1]–[4]. Some studies use 5G/B5G KPIs to predict network performance, while others use federated learning for QoS forecasting or combine multiple telemetry layers in O-RAN settings [3], [4]. These works demonstrate the value of telemetry-driven network management, but most focus primarily on predictive accuracy rather than robustness under unseen deployment conditions [2]–[4].

2.2 Federated Learning for Heterogeneous Clients

Federated learning enables shared model training without exchanging raw client data [5]. In practice, however, clients often differ in sample size, label balance, feature distributions, and system capacity, creating strong non-IID effects [6], [7]. FedProx is a representative approach for heterogeneous federated optimization because it stabilizes training through a proximal term [6]. Subsequent work has further examined non-IID behavior, client drift, and concept drift in federated settings [7]–[9]. Even so, most prior work treats heterogeneity mainly as an optimization issue and less often uses client profile information to improve robustness under unseen regimes.

2.3 Distribution Shift and Unseen-Environment Evaluation

Distribution variation and unseen-environment generalization are core issues in reliable machine learning [8]–[10]. Prior reviews show that OOD shift, distribution shift, and concept drift can substantially affect deployment reliability [8]–[10]. In federated settings, the challenge becomes more difficult because distribution changes are compounded by client heterogeneity [8], [9]. Nevertheless, many network-side prediction studies still evaluate on seen distributions, with less emphasis on thresholded decision robustness under unseen regimes, such as balanced accuracy and worst-regime F1, which are more directly connected to operational decision quality.

3. Problem Formulation

We study next-window SLA risk prediction for 6G RAN telemetry data. For client c at time window t, the telemetry observation is defined as follows:

$$x_t^{(c)} \in \mathbb{R}^d \tag{1}$$

Here, d denotes the telemetry feature dimension. Each client also has a relatively stable profile vector describing heterogeneous attributes:

$$p_c \in \mathbb{R}^m, \tag{2}$$

where m is the profile feature dimension. Each sample also belongs to a network operating regime:

$$r \in \mathcal{R}, \tag{3}$$

where R denotes the set of operating conditions, such as light, medium, heavy, and impaired. To predict the service risk for the next window, the model uses a historical telemetry sequence of length L as input. For client c at time t, the sequence is defined as:

$$X_t^{(c)} = \left[x_{t-L}^{(c)}, x_{t-L+1}^{(c)}, \ldots, x_{t-1}^{(c)}\right]. \tag{4}$$

The prediction target is the SLA risk label for the next window:

$$y_t^{(c)} \in \{0, 1\}, \tag{5}$$

where y_t^(c) = 1 indicates SLA risk in the next window and y_t^(c) = 0 indicates normal service. The learning objective is then written as:

$$\hat{y}_t^{(c)} = f_\theta\left(X_t^{(c)}, p_c\right), \tag{6}$$

where fθ(·) is the prediction model. To evaluate generalization under unseen conditions, the paper adopts a leave-one-regime-out setting. For any held-out regime r*:

$$\mathcal{D}_{\text{train}}(r^\star) = \left\{\left(X_t^{(c)}, p_c, y_t^{(c)}\right) \mid r \neq r^\star\right\}, \tag{7}$$

$$\mathcal{D}_{\text{test}}(r^\star) = \left\{\left(X_t^{(c)}, p_c, y_t^{(c)}\right) \mid r = r^\star\right\}. \tag{8}$$

This setting exposes the model only to seen regimes during training while requiring direct evaluation on a completely unseen operating condition. It therefore measures not only fitting under seen conditions but also robust decision-making under unseen regimes, which is closer to the operational drift scenarios expected in AI-native 6G networks.

4. Proposed Method

The goal of the proposed method is not simply to maximize overall ranking performance, but to improve balanced decision-making ability and worst-case robustness for federated SLA risk prediction under unseen operating regimes. The task is built on cross-layer 6G RAN telemetry data with a leave-one-regime-out protocol, where one regime is held out as an unseen test environment and the remaining regimes are used for training and validation. The final method combines client heterogeneity modeling, regime-aware aggregation, and validation-stage calibration to better support federated prediction under unseen conditions.

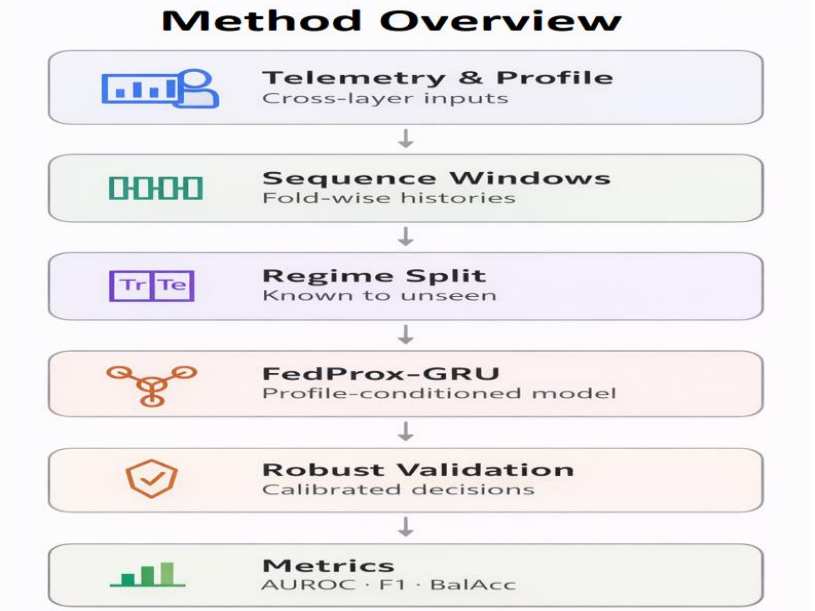

Figure 1. Overview of the proposed method for unseen-regime federated SLA-risk forecasting.

4.1 Sequence Encoder

Let the latest L telemetry windows for a client at time t be X_(t-L+1:t) ∈ R^(L×F), where F is the cross-layer feature dimension. A single-layer GRU encodes the recent telemetry history, and the hidden state at the final time step is used as the sequence representation:

$$\mathbf{h}_t = \mathrm{GRU}(\mathbf{X}_{t-L+1:t}). \tag{9}$$

The motivation is that SLA risks are often associated with recent changes in traffic load, wireless status, and quality of service, so short temporal windows provide a compact and effective dynamic context. The sequence length was searched fold-wise and the most stable value was selected for each held-out fold.

4.2 Portrait Conditioning

Relying only on temporal telemetry makes it difficult to fully express statistical heterogeneity among clients. The client portrait vector p_c is therefore extracted from the heterogeneity table and mapped into an embedding space through a lightweight MLP:

$$\mathbf{z}_c = \mathrm{MLP}(\mathbf{p}_c). \tag{10}$$

The sequence representation h_t is then concatenated with the portrait representation z_c and passed to the prediction head to produce the risk logit:

$$\hat{y}_t = g([\mathbf{h}_t; \mathbf{z}_c]). \tag{11}$$

This design lets the model share global parameters while remaining aware of long-term differences across clients. Ablation results show that removing portrait conditioning reduces AUROC, F1, and balanced accuracy, indicating that it supports robust prediction under unseen regimes.

4.3 Federated Optimization

Training uses FedProx-based federated optimization. In each communication round, several eligible clients are sampled, updated locally for a small number of epochs, and then aggregated into the global model. Given local parameters w and global parameters w^(g), the client optimization objective is:

$$\mathcal{L}_c = \mathcal{L}_{\text{task}} + \frac{\mu}{2}\|\mathbf{w} - \mathbf{w}^{(g)}\|_2^2, \quad (12)$$

Here, μ is the proximal coefficient of FedProx. The task loss uses class-balanced focal BCE with label smoothing to reduce bias caused by class imbalance. The study also evaluates Centralized-GRU, HistGB, and vanilla FedProx-GRU so that the contribution of the enhanced design can be separated from the effect of federated training itself.

4.4 Regime-Aware Aggregation and Decision Calibration

Beyond standard federated averaging, the method introduces regime-aware client weighting together with EMA-style global updates to reduce the excessive influence of a single client or regime on the aggregated result. The purpose is not to become the strongest global ranker, but to obtain better balanced thresholded decisions under unseen regimes. During validation, probabilities are calibrated using temperature scaling and, when needed, isotonic regression. Threshold selection is also made regime-aware by considering global F1, macro balanced accuracy, and worst-regime performance. Final evaluation reports AUROC, AUPRC, F1, balanced accuracy, Brier score, and macro and worst-regime metrics.

5. Experimental Setup

5.1 Dataset and Task

The experiment uses two CSV files: one containing time-organized cross-layer telemetry features and another containing client heterogeneity profiles. The dataset covers 200 clients across four operating regimes: heavy, impaired, light, and medium. The task is defined as next-window SLA status prediction and is further framed as binary risk prediction, indicating whether SLA risk will occur in the next window. Inputs consist of recent telemetry histories together with client-level portrait features.

5.2 Evaluation Protocol

To evaluate generalization under unseen regimes, the paper adopts a leave-one-regime-out protocol, yielding four folds. In each fold, one regime is held out as the test set and the remaining three regimes are used for training and validation. For the sequence model, the sequence length is searched within each fold over {8, 12, 16}, and the value with the best validation performance is selected. This protocol more directly reflects robustness under new operating conditions than evaluation on seen distributions alone.

5.3 Baselines

Four methods are compared: HistGB as the non-sequential tree baseline, Centralized-GRU as the centralized sequence upper-bound reference, FedProx-GRU as the standard federated sequence baseline, and Proposed_Full as the complete portrait-conditioned and regime-aware federated method. This setup supports comparison between centralized and federated learning, as well as between vanilla and enhanced federated designs.

5.4 Metrics

The evaluation reports ranking metrics, thresholded decision metrics, and probabilistic quality metrics, including AUROC, AUPRC, F1, balanced accuracy, and Brier score. To better reflect cross-regime robustness, the paper also reports macro balanced accuracy, worst-regime F1, and worst-regime AUROC. Because the study emphasizes robust decision-making under unseen regimes, balanced accuracy and worst-regime metrics receive particular attention in the analysis.

5.5 Implementation Details

The model is implemented with PyTorch and scikit-learn and trained in a CUDA environment. The federated sequence model uses a GRU encoder with FedProx optimization. The complete model additionally incorporates portrait conditioning, regime-aware weighting, and EMA-style aggregation. Validation-stage probability calibration and threshold search are applied under the same leave-one-regime-out protocol for all compared methods unless otherwise noted.

6. Results and Discussion

This section analyzes the results from overall comparison, unseen-regime robustness, and ablation studies. Overall, the findings are

consistent with the paper's central claim: centralized models remain stronger in AUROC, whereas the proposed method performs best on the balanced decision metrics that are more directly relevant to deployment and clearly outperforms the vanilla federated baseline.

6.1 Main Comparison

The overall comparison under leave-one-regime-out evaluation shows that Centralized_GRU achieves the highest mean AUROC of 0.882, followed by HistGB at 0.872. Proposed_Full reaches a mean AUROC of 0.832. Although it is not the strongest ranker, it performs best on several metrics more closely tied to decision quality: macro balanced accuracy reaches 0.676, mean F1 reaches 0.708, and worst-regime F1 reaches 0.485. Compared with vanilla FedProx-GRU, Proposed_Full consistently improves AUROC, F1, and balanced accuracy, indicating that profiling and regime-aware design improve the quality of federated risk decisions.

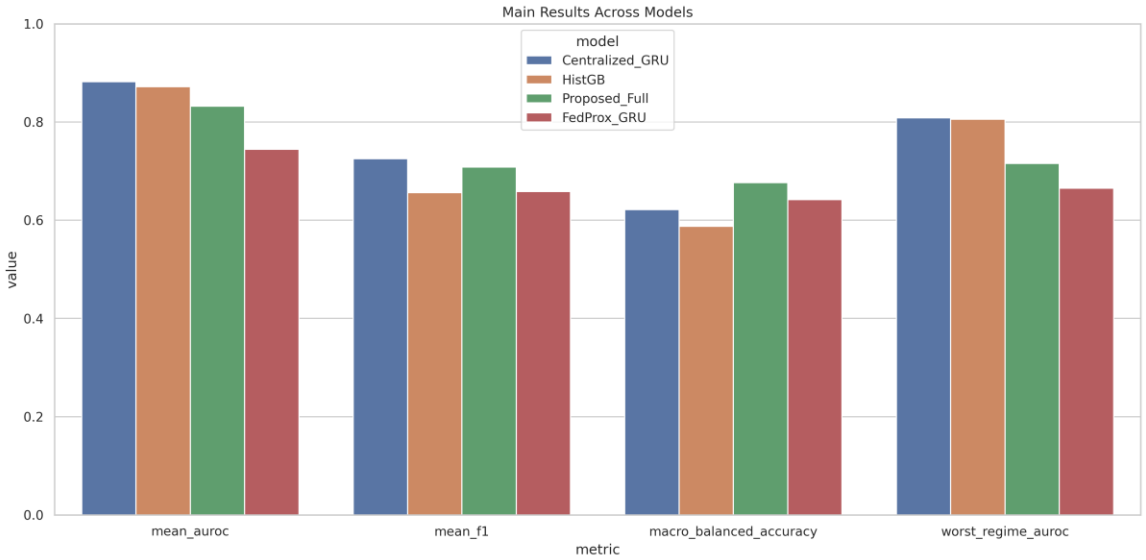

Figure 2. Main comparison across models under leave-one-regime-out evaluation.

From an application perspective, the paper is less concerned with which model has the highest overall ranking score and more concerned with which model provides more balanced and robust risk decisions under unseen regimes. Under that criterion, Proposed_Full clearly outperforms the vanilla federated baseline and also compares favorably against some methods with higher AUROC but weaker balanced decision-making performance.

6.2 Unseen-Regime Robustness

Per-held-out-regime F1 results show that the light fold is the most challenging scenario across methods. HistGB reaches only 0.148 on this fold, Centralized-GRU reaches 0.414, FedProx-GRU reaches 0.442, and Proposed_Full achieves the highest value at 0.485. Because the light regime contains a very low proportion of positive cases, it better reveals model robustness under unseen and imbalanced conditions. Compared with earlier experimental versions, the final method avoids severe collapse on the hardest fold, indicating more stable thresholded decision-making.

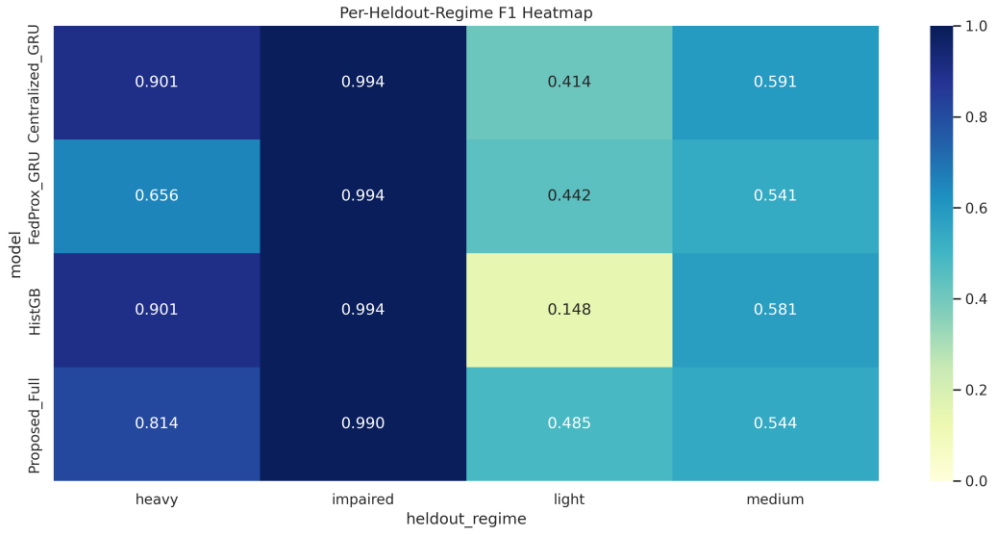

Figure 3. Per-held-out-regime F1 results under leave-one-regime-out evaluation.

Meanwhile, all models are close to saturation on the impaired fold, which means that this fold alone cannot distinguish model quality very well. For this reason, the paper emphasizes macro indicators and worst-regime indicators rather than relying only on a single average score. Overall, Proposed_Full provides stronger balanced decision-making under unseen regimes, which is one of the key findings of the study.

6.3 Ablation and Sequence-Length Findings

Ablation results show that removing regime-aware weighting reduces AUROC to 0.762 and balanced accuracy to 0.639, while removing portrait conditioning further reduces AUROC to 0.713 and balanced accuracy to 0.648. Both values are lower than those of Proposed_Full, indicating that both components contribute to the final performance. Portrait conditioning appears to have a larger effect on AUROC, whereas regime-aware weighting is particularly important for balanced decision quality.

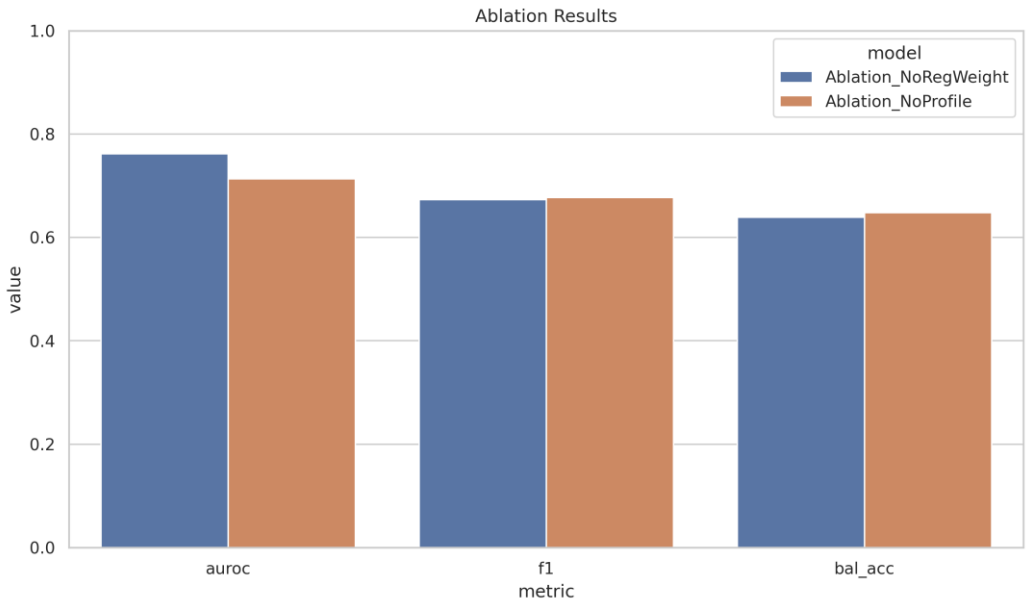

Figure 4. Ablation results for portrait conditioning and regime-aware weighting.

Fold-wise sequence-length search further shows that the heavy, impaired, and medium folds tend to favor shorter historical windows (L = 8), while the light fold selects a longer window (L = 16). This suggests that short-term context is often sufficient, but longer histories can still help in the most difficult and sparse regimes. Accordingly, the paper does not claim that a fixed short sequence is always optimal; instead, the choice of sequence length should depend on the statistical characteristics of the held-out regime.

6.4 Discussion and Limitations

The main strength of the proposed method is not that it is the best overall ranker, but that it provides more balanced and robust federated risk decisions under unseen regimes. Centralized models still remain stronger on AUROC, indicating that room for improvement remains in federated settings. The overall calibration of Proposed_Full is also not claimed to be the best. In addition, the current data still exhibit structured characteristics, whereas real-world large-scale 6G deployment environments are likely to be more complex. External validity therefore still requires further validation in more realistic network scenarios.

7. Conclusion

The proposed portrait-conditioned federated method is not the strongest model in terms of overall ranking ability, but under leave-one-regime-out regime-shift evaluation it performs best in robustness-oriented balanced decision-making and worst-regime decision quality within federated settings. The paper studies next-window SLA risk prediction on cross-layer 6G RAN telemetry and proposes a portrait-conditioned FedProx-GRU that combines client profile modeling with regime-aware aggregation. Experimental results show that although centralized baselines remain stronger on AUROC, the proposed method achieves better balanced decision performance in federated settings and the strongest worst-regime F1, suggesting more practical thresholded decision robustness under regime shift. Future work should further reduce the gap with centralized upper bounds and test generalization and deployment stability under more realistic and complex network scenarios.

Acknowledgment and Declarations

Acknowledgment

The authors would like to thank all collaborators, as well as family and friends, for their continued support, encouragement, and understanding during the research and writing process.

Compliance with Ethical Standards

This study did not involve human subjects, animals, or interventional procedures requiring ethics approval. The data and experiments were used only for method validation and model evaluation, and the work was conducted in accordance with academic integrity and research ethics standards.

AI Assistance Disclosure

AI tools were used only for language polishing, formatting, and technical writing support, and not for data generation, experiments, result interpretation, or reference generation. The research design, analysis, and conclusions were completed independently by the authors.

IEER
Institute for Excellence in Education & Research

Funding

No funding was received for this study.

Conflict of Interest

The authors declare no conflicts of interest.

REFERENCES

[1] B. Brik, H. Chergui, L. Zanzi, F. Devoti, A. Ksentini, M. S. Siddiqui, X. Costa-Perez, and C. Verikoukis, "Explainable AI in 6G O-RAN: A Tutorial and Survey on Architecture, Use Cases, Challenges, and Future Research," IEEE Communications Surveys and Tutorials, 2024, doi: 10.1109/COMST.2024.3510543.

[2] M. M. Islam, K. Hasan, and S. H. Jeong, "Performance evaluation and optimization for 6G networks: A survey of KPIs, Tools, and AI models," ICT Express, 2025, doi: 10.1016/j.icte.2025.12.012.

[3] N. P. Tran, O. Delgado, B. Jaumard, and F. Bishay, "ML KPI Prediction in 5G and B5G Networks," in 2023 Joint European Conference on Networks and Communications and 6G Summit (EuCNC/6G Summit), 2023, doi: 10.1109/EuCNC/6GSummit58263.2023.10188363.

[4] N. Baganal-Krishna, R. Lubben, E. Liotou, K. V. Katsaros, and A. Rizk, "A federated learning approach to QoS forecasting in cellular vehicular communications: Approaches and empirical evidence," Computer Networks, vol. 242, art. no. 110239, 2024, doi: 10.1016/j.comnet.2024.110239.

[5] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, PMLR 54, pp. 1273-1282, 2017, doi: 10.48550/arXiv.1602.05629.

[6] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated Optimization in Heterogeneous Networks," in Proceedings of Machine Learning and Systems, vol. 2, 2020.

[7] H. Zhu, J. Xu, S. Liu, and Y. Jin, "Federated learning on non-IID data: A survey," Neurocomputing, vol. 465, pp. 371-390, 2021, doi: 10.1016/j.neucom.2021.07.098.

[8] M. Kang, S. Kim, K. H. Jin, and C. D. Yoo, "FedNN: Federated learning on concept drift data using weight and adaptive group normalizations," Pattern Recognition, vol. 149, art. no. 110230, 2024, doi: 10.1016/j.patcog.2023.110230.

[9] O. A. Mahdi, E. Pardede, S. Bevinakoppa, and N. Ali, "Federated Learning Under Concept Drift: A Systematic Survey of Foundations, Innovations, and Future Research Directions," Electronics, vol. 14, no. 22, art. no. 4480, 2025, doi: 10.3390/electronics14224480.

[10] J. Yang, K. Zhou, Y. Li, and Z. Liu, "Generalized Out-of-Distribution Detection: A Survey," International Journal of Computer Vision, vol. 132, no. 12, pp. 5635-5662, 2024, doi: 10.1007/s11263-024-02117-.

[11] B. Raza, A. Maitlo, Z. H. Shar, and I. Hyder, "Operational Android Malware Filtering: Calibrated Probabilities and Distribution-Free Guarantees," Kashf Journal of Multidisciplinary Research, vol. 2, no. 12, pp. 58-73, 2025, doi: 10.71146/kjmr778.

[12] I. Hyder, R. A. Shaikh, R. H. Arain, Z. Hussain, and B. Raza, "Audit-Ready Healthcare Fraud Screening: Split-Safe Provider Aggregation and Explainable Boosted Risk Triage," Southern Journal of Computer Science, vol. 2, no. 1, pp. 18-28, 2026.

IEER
Institute for Excellence in Education & Research