13th International Conference on Applied Informatics
Eszterházy Károly Catholic University
Eger, Hungary, February 19–21, 2026

Retraining-Free Federated Unlearning for Credit Risk Modeling via Constraint-Controlled Masking

Pınar Yazgan[a], **Bálint Molnár**[b]

[a]Eötvös Loránd University (ELTE), Doctoral School of Informatics
fmx2hx@inf.elte.hu

[b]Eötvös Loránd University (ELTE), Faculty of Informatics
molnarba@inf.elte.hu

Abstract

Federated learning (FL) allows organizations to train a shared model without exchanging raw data, typically via FedAvg aggregation [3]. This is appealing for credit risk modeling, where cross-institutional data sharing is constrained by regulation and business requirements [1]. GDPR Article 17 (right to erasure) motivates federated unlearning, i.e., removing a participant's contribution after training [5]. However, full retraining is costly, and subtraction-style retraining-free methods (e.g., FedEraser) can be fragile under non-IID heterogeneity [2]. We present **LLMU-C**, a method that estimates a removed client's cumulative data footprint from auditable server logs and suppresses it using a learnable *soft mask* in parameter space. A single parameter, τ, controls attenuation strength. In a four-way non-IID partition of the UCI Credit Card Default dataset [7] (simulating different banks), LLMU-C preserves retained-client utility, reduces prediction drift relative to subtraction baselines, and yields near-chance membership inference on the removed client's data [6]. In our primary setting, LLMU-C achieves KEEP_F1=0.570, drift=0.042, and MIA_AUC=0.520.

Approach

- **Motivation.** Deletion requests are realistic in consortium FL and may be required post-training [5]. For credit systems, the unlearned model should

remain stable for retained clients, motivating conservative unlearning under non-IID heterogeneity [2].

- **Research questions.** Can we remove one client's contribution without full retraining while preserving retained-client utility? Under non-IID splits, does soft masking yield more stable outputs than subtraction-based unlearning? What server-side artifacts (per-round updates, secure aggregates, verifiable summaries) suffice for client-level deletion in regulated deployments [4, 5]?

- **Method.** After FL convergence, upon a deletion request, the server approximates the removed client's cumulative footprint from auditable logs and learns a soft mask that down-weights this component while staying close to the pre-unlearning model on retained data (calibration/distillation). The single knob τ trades off suppression strength vs. conservativeness and is tuned on a small retained validation split to satisfy a drift/utility budget. We evaluate utility (F1/AUC), stability via prediction drift against an oracle retraining reference, and privacy leakage via loss-based membership inference AUC [6].

- **Experimental setting.** We use the UCI Credit Card Default dataset [7] ($N = 30{,}000$, 23 features; positive rate ≈ 0.221) and simulate four banks via a non-IID label-quantile partition (Bank2 removed). Retained-client utility is measured on KEEP_TEST (F1/AUC); stability is drift Mean$|p - p_{\text{retrain}}|$ on retained samples, where p_{retrain} comes from RETRAIN_GT (oracle retrain without Bank2); privacy is loss-based membership inference AUC distinguishing removed-client TRAIN vs. removed-client HOLDOUT samples [6]. Baselines: FULL (pre-deletion model with all clients) and RETRAIN_GT (post-deletion oracle).

- **Contribution.** Constraint-controlled soft masking with a single parameter τ provides a stable retraining-free alternative to hard subtraction under non-IID heterogeneity, while making logging/audit assumptions explicit for regulated deployments.

Results

LLMU-C matches FULL on retained-client utility (KEEP_F1 0.570 vs. 0.567) while improving stability under non-IID heterogeneity (drift 0.052→0.042) and keeping membership inference near chance (MIA_AUC$\approx$ 0.5) [6], with practical runtime (2.91s). Compared to subtraction-style baselines, it avoids the severe instability/utility drop observed for FedEraser (drift 0.211, KEEP_F1 0.381).

Table 1. Retained-client utility (KEEP_F1), prediction drift vs. RETRAIN_GT, loss-based membership inference (MIA_AUC), and wall-clock time.

Method	KEEP_F1↑	Drift↓	MIA_AUC↓	Time(s)↓
FULL	0.567	0.052	0.522	26.17
RETRAIN_GT (oracle)	0.544	0.000	0.519	9.36
FedMask (λ=0.3)	0.537	0.058	0.515	< 0.01
FedEraser	0.381	0.211	0.506	< 0.01
LLMU-C (KeepDistill+ES, τ=0.1)	0.570	0.042	0.520	2.91

Conclusion

Constraint-controlled masking is a practical compromise for regulated FL: it supports client-level deletion without full retraining while making audit/logging assumptions explicit [5]. A key limitation is reliance on trustworthy server-side artifacts (e.g., per-round update traces or verifiable summaries), which may not be available in all deployments [4]. Next, we will validate across additional datasets and partitions, study repeated and multi-client deletions, and develop a principled selection of τ under stronger threat models and attacks beyond loss-based membership inference [4].

References

[1] S. Lessmann, B. Baesens, H.-V. Seow, L. C. Thomas: *Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research*, European Journal of Operational Research 247.1 (2015), pp. 124–136, DOI: 10.1016/j.ejor.2015.05.030.

[2] G. Liu, X. Ma, Y. Yang, C. Wang, J. Liu: *FedEraser: Enabling Efficient Client-Level Data Removal from Federated Learning Models*, in: 2021 IEEE/ACM 29th International Symposium on Quality of Service (IWQoS), IEEE, 2021, pp. 1–10, DOI: 10.1109/IWQOS52092.2021.9521274.

[3] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, B. Agüera y Arcas: *Communication-Efficient Learning of Deep Networks from Decentralized Data*, in: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS), vol. 54, Proceedings of Machine Learning Research, 2017, pp. 1273–1282, arXiv: 1602.05629 [cs.LG], URL: https://arxiv.org/abs/1602.05629.

[4] T. T. Nguyen, T. T. Huynh, P. L. Nguyen, A. W.-C. Liew, H. Yin, Q. V. H. Nguyen: *A Survey of Machine Unlearning*, 2022, arXiv: 2209.02299 [cs.LG], URL: https://arxiv.org/abs/2209.02299.

[5] *Regulation (EU) 2016/679 (General Data Protection Regulation)*, Official Journal of the European Union, 2016, URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj (visited on 12/31/2025).

[6] R. Shokri, M. Stronati, C. Song, V. Shmatikov: *Membership Inference Attacks Against Machine Learning Models*, in: 2017 IEEE Symposium on Security and Privacy (SP), IEEE, 2017, pp. 3–18, DOI: 10.1109/SP.2017.41.

[7] I. Yeh: *Default of Credit Card Clients*, UCI Machine Learning Repository (Dataset), 2009, DOI: 10.24432/C55S3H, URL: https://archive.ics.uci.edu/dataset/350/default+of+credit+card+clients (visited on 12/31/2025).