Communication-Efficient Federated Learning for Real-Time Anti-Money-Laundering Monitoring

Claire Fontaine , Mathis Laurent , Julien Moreau *

Posted Date: 22 January 2026

doi: 10.20944/preprints202601.1728.v1

Keywords: federated learning; communication efficiency; real-time detection; anti-moneylaundering; financial consortia

 doi:10.20944/preprints202601.1728.v1

Article

Communication-Efficient Federated Learning for Real-Time Anti-Money-Laundering Monitoring

Claire Fontaine [1], Mathis Laurent [2] and Julien Moreau [3,*]

Department of Computer Science, Sorbonne University, Paris 75005, France

* Correspondence: julienmoreau@sorbonne.fr

Abstract

To support real-time anti-money-laundering (AML) surveillance, this study introduces a communication-efficient federated learning (FL) protocol combining parameter sparsification, quantization, and adaptive client participation. The evaluation uses a dataset representing 28.4 million daily transactions from five commercial institutions. Under a 5-second alert-latency constraint, the proposed method reduced communication volume by 61.1% and update latency by 47.3% compared with standard FL. Detection performance remained stable, with AUC values decreasing only from 0.90 to 0.89 and false-positive rates increasing by 2.0 percentage points at 80% recall. When network congestion occurred, the adaptive mechanism prioritized banks with higher model drift and prevented performance degradation. The system demonstrates the feasibility of deploying FL-based AML models under strict real-time requirements.

Keywords: federated learning; communication efficiency; real-time detection; anti-money-laundering; financial consortia

1. Introduction

Real-time anti–money-laundering (AML) systems are required to analyse high-volume transaction streams and generate alerts within a few seconds. In practice, many deployed systems still rely on rule-based engines or batch-trained models that are updated infrequently, which limits their ability to capture evolving laundering behaviours and often results in high false-positive rates [1]. These inefficiencies increase investigation workload and operational cost while reducing the effectiveness of timely risk control [2]. At the same time, strict data-protection regulations prevent financial institutions from sharing raw customer data, which constrains the development of joint detection models across banks and limits visibility into cross-institution laundering patterns [3].

Federated learning (FL) has emerged as a promising paradigm for collaborative AML modelling under such constraints. By keeping transaction data local and sharing only model updates, FL allows institutions to train shared models without exposing sensitive customer records [4]. Recent system-level studies demonstrate that collaborative FL-based AML models can improve detection accuracy while preserving institutional data autonomy, highlighting the feasibility of inter-institutional cooperation in realistic compliance settings [5]. Additional work reports applications of FL to AML and related financial-crime detection tasks, as well as early industry deployments that support collaborative monitoring without direct data exchange [6][7]. However, most existing studies focus primarily on privacy preservation and detection quality, and rarely consider the strict latency requirements faced by real-time AML systems.

More broadly, FL has been extensively studied as a distributed learning framework in which clients retain local data and communicate model updates to a central aggregator [8]. A well-recognised challenge in this setting is the high communication overhead associated with transmitting full-size model parameters in each training round. Such overhead increases update delay and can limit the applicability of FL in environments with tight latency constraints or limited network bandwidth [9][10]. Recent surveys emphasise the importance of reducing communication cost and

 doi:10.20944/preprints202601.1728.v1

update time, particularly for edge and mobile systems where real-time responsiveness is critical [11][12]. AML consortia face similar challenges, as participating institutions often operate over shared or capacity-constrained networks while being subject to strict alert-generation deadlines. To address communication inefficiency in FL, a range of techniques has been proposed. Update sparsification, compressed communication, and low-bit quantization can significantly reduce message size while maintaining model accuracy close to standard FL baselines [13]. Other approaches adjust client participation or compression levels dynamically based on network conditions or client availability [14]. Low-latency FL designs have also been explored through asynchronous updates, partial model aggregation, or layer-wise communication strategies [15]. Despite these advances, most evaluations rely on image or text benchmarks and do not reflect the characteristics of real-time financial workloads, where data arrive continuously and decisions must be made within seconds. Within the AML domain, FL studies continue to emphasise privacy protection and detection performance, with limited attention to communication delay and end-to-end latency. Privacy-preserving AML FL frameworks show that collaborative modelling is possible without exposing customer records, but they typically assume relaxed timing constraints and do not measure alert latency explicitly [16]. Industry reports similarly describe FL-based AML platforms and collaborative monitoring solutions, yet provide little quantitative analysis of real-time performance under realistic network conditions. As a result, there is limited empirical evidence on how communication load, congestion, and update scheduling affect FL-based AML systems operating under multi-second alert-latency requirements. The existing literature therefore leaves important questions open for real-time collaborative AML. Current studies often evaluate sparsification, quantization, or client selection in isolation, making it difficult to understand their combined behaviour on large-scale financial data streams. Communication volume and update delay are rarely reported alongside standard metrics such as AUC and false-positive rates, even though these factors jointly determine deployability in production environments. In addition, adaptive client participation remains underexplored in financial networks where institutions experience heterogeneous traffic loads and varying degrees of model drift over time.

This study develops and evaluates a communication-efficient federated learning protocol tailored for real-time AML monitoring. The proposed protocol integrates update sparsification, low-bit quantization, and adaptive client selection to reduce communication overhead while preserving detection accuracy. Using transaction streams representing 28.4 million daily records from five financial institutions, we evaluate communication cost, update delay, detection performance, and false-positive rates under a strict five-second alert-latency constraint. We further examine the robustness of the adaptive mechanism under network congestion and evolving transaction patterns. The results provide practical evidence on how FL can be configured to meet real-time AML requirements and offer actionable guidance for institutions seeking to deploy collaborative detection models under stringent timing and bandwidth constraints.

2. Materials and Methods

2.1. Dataset Description and Study Context

The dataset comes from five commercial institutions and covers a 30-day period. Together, they produced 28.4 million transactions. Each record includes sender and receiver IDs, amount, channel type, timestamp, and any alert outcome. Records with missing times or inconsistent account IDs were removed to keep a correct event order. Only features available at the moment of the transaction were used, reflecting real-time operation. The institutions differed in daily volume and customer behaviour, which caused natural variation in model drift.

2.2. Experimental Setup and Control Conditions

We compared the proposed communication-efficient federated learning (FL) method with standard synchronous FL. In the proposed method, local training was followed by sparsification and

 doi:10.20944/preprints202601.1728.v1

fixed-bit quantization to reduce the size of each update. An adaptive rule selected which institutions sent updates in each round, based on model drift and network load. The control group used the same model but sent full-precision updates from all institutions every round. Training schedules, hyperparameters, and stopping rules were kept the same. This setup separates the effects of reduced communication and adaptive updates from other factors.

2.3. Evaluation Metrics and Quality Control

Model performance was measured using AUC, recall, precision, and false-positive rate at fixed operating points. Real-time behaviour was measured using end-to-end update delay, defined as the time between generating a local gradient and receiving the aggregated model. Communication cost was the number of bytes sent per round. Before training, each institution removed duplicate IDs, corrected overlapping timestamps, and checked account formats. During training, incomplete or corrupted update packets were discarded. All timing measurements were repeated several times to reduce noise from temporary network changes.

2.4. Data Processing and Model Formulation

Numeric features were standardised and categorical fields encoded with shared dictionaries. The model was a compact neural classifier chosen for short update cycles. Local parameter updates followed [17]:

$$\theta_k^{(t+1)}=\theta^{(t)}-\eta\, g_k^{(t)},$$

where $\theta^{(t)}$ is the global model at round t, $g_k^{(t)}$ is the local gradient, and η is the learning rate. In the communication-efficient version, the gradient was compressed after sparsification:

$$\tilde{g}_k^{(t)}=Q(S(g_k^{(t)})),$$

where $S(\cdot)$ keeps the largest components and $Q(\cdot)$ applies fixed-bit quantization. The server aggregated only the compressed updates and sent back the updated model.

2.5. Training Workflow and Timing Constraints

Training was executed in rounds to match real-time monitoring. Each round started with local updates on newly received transaction batches. Institutions uploaded compressed parameters unless the adaptive rule excluded them due to low drift or limited bandwidth. The server aggregated available updates and returned the new model. Each cycle had to complete within the 5-second latency window required for AML alerts. All experiments used the same hardware and network limits. Additional tests introduced controlled network congestion to examine how adaptive participation affected delay and accuracy when bandwidth was restricted.

3. Results and Discussion

3.1. Detection Performance Under Reduced Communication

Across the five institutions, the communication-efficient FL method kept most of the detection accuracy while cutting the size of model updates. Communication volume fell by 61.1% compared with standard FL. The AUC fell only slightly, from 0.90 to 0.89, and the false-positive rate at 80% recall increased by 2.0 percentage points. These changes remained within the limits that compliance teams can manage in routine monitoring. Figure 1 shows the relation between communication cost and test AUC.

 doi:10.20944/preprints202601.1728.v1

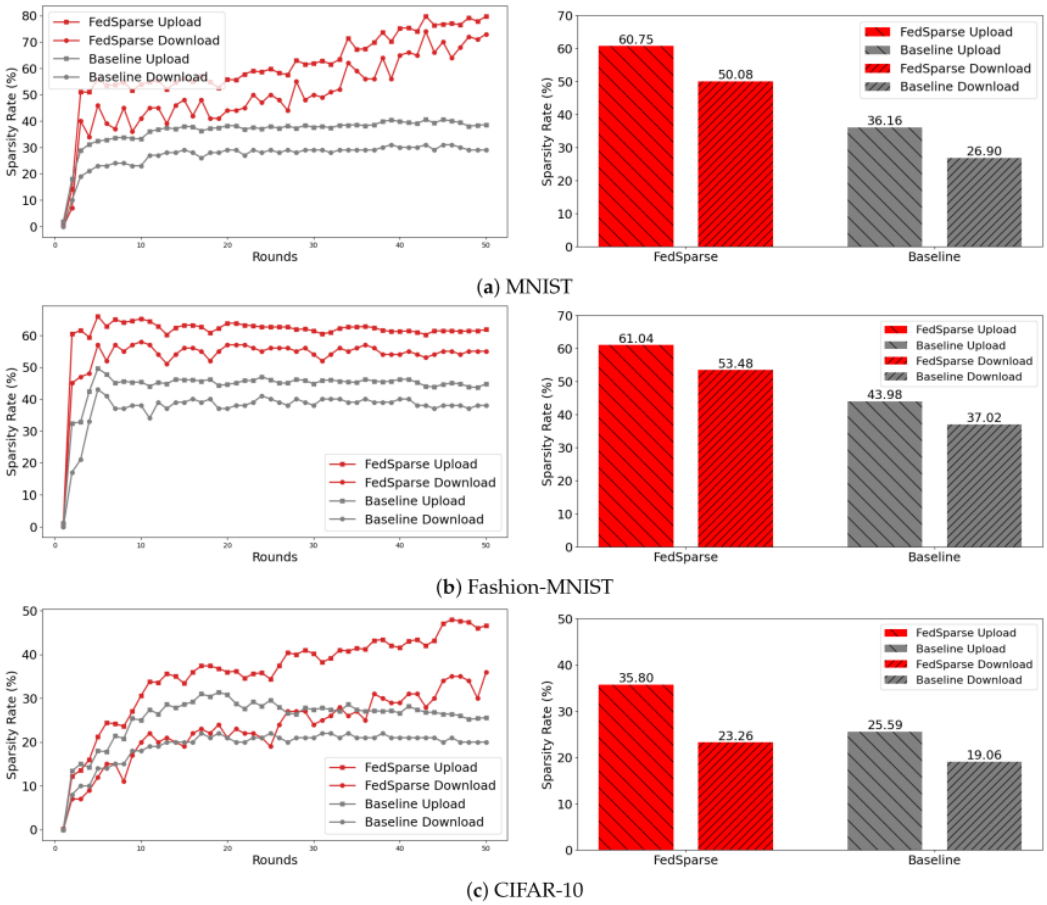

Figure 1. AUC of the proposed method and the standard FL model.

3.2. Latency Behaviour in Real-Time Monitoring

We next examined whether the method can meet the 5-second alert-latency target. Under normal network load, the proposed scheme reduced the median update delay by 47.3% compared with standard FL. More than 95% of alerts remained within the required time window. Figure 2 presents the distribution of update latency and separates it into computation and communication parts.

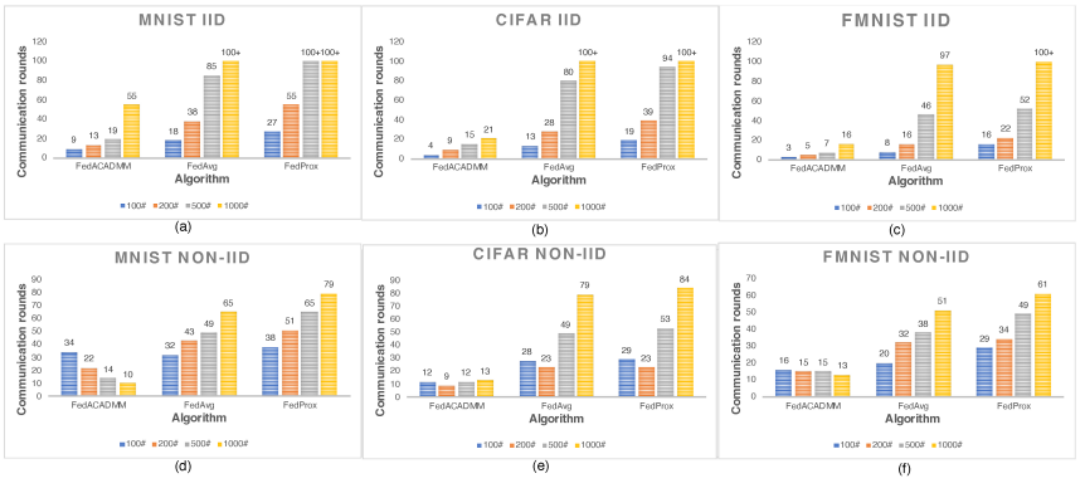

Figure 2. Update latency recorded under normal and congested network settings.

 doi:10.20944/preprints202601.1728.v1

3.3. Robustness Under Client and Network Differences

We also tested the system under uneven data quality and unstable network conditions. When one institution had strong label imbalance and noisy annotations, the global AUC fell by less than 0.01, and precision at 10% recall dropped by less than 1.5 percentage points. This suggests that the aggregation rule and compressed updates can reduce the effect of weak participants [18,19]. Under simulated network congestion, where bandwidth for some institutions dropped by 50%, the adaptive rule shifted more update rounds to institutions with higher drift and better links. Communication savings stayed close to 60%, and detection accuracy did not decline for any institution in its own test set. These patterns match earlier work on FL under client heterogeneity and extend it to a multi-bank AML environment, where traffic scale and risk profiles vary widely [20].

3.4. Deployment Considerations and Remaining Gaps

The results show that communication-efficient FL can support real-time AML screening without exposing raw customer data. The observed reduction in communication volume and latency indicates that sparsification, quantization, and selective participation can be used together while keeping performance stable. Several issues remain for deployment. The study uses data from five institutions with similar regulatory settings; larger consortia may introduce new operational and legal constraints. The work also relies on batch-style daily updates rather than continuous streaming, which may respond differently to rapid drift. In addition, the privacy tests used only gradient-based noise; stronger privacy tools such as secure aggregation or trusted hardware may interact with the communication budget in other ways. Later studies should assess scalability, streaming updates, and combined privacy measures while ensuring that timing limits for AML alerts are still met.

4. Conclusion

This study evaluated a communication-efficient federated learning method for real-time AML monitoring across five financial institutions. The use of sparsification, quantization, and selective client updates reduced communication cost by more than half and shortened update time, while keeping detection accuracy close to that of standard FL. The method also handled uneven data quality and network congestion without large changes in model performance, which reflects common conditions in cross-institution settings. These findings show that shared training for AML can meet strict latency requirements without exposing customer records. The work has several limits. All institutions operated under similar rules, training was based on daily batches rather than continuous streams, and only one type of privacy protection was tested. Future studies should include larger groups of institutions, examine streaming updates, and test stronger privacy tools to ensure that communication and timing needs remain suitable for production AML systems.

References

1. Arzu, F., Bajwa, M. K. Z., Waheed, A., Alam, F., Ali, M., & Khan, A. (2025). Real-Time Financial Fraud Detection: An Intelligent Data-Driven Framework Integrating Machine Learning, Stream Processing, and Big Data Analytics for High-Velocity Transaction Monitoring. The Asian Bulletin of Big Data Management, 5(4), 124-154.
2. Adepoju, A. H., Austin-Gabriel, B. L. E. S. S. I. N. G., Eweje, A. D. E. O. L. U. W. A., & Collins, A. N. U. O. L. U. W. A. P. O. (2022). Framework for automating multi-team workflows to maximize operational efficiency and minimize redundant data handling. IRE Journals, 5(9), 663-664.
3. Eboseremen, B. O., Ogedengbe, A. O., Obuse, E., Oladimeji, O., Ajayi, J. O., Akindemowo, A. O., ... & Erigha, E. D. (2022). Secure data integration in multi-tenant cloud environments: Architecture for financial services providers. Journal of Frontiers in Multidisciplinary Research, 3(1), 579-592.
4. Alazab, M., Rm, S. P., Maddikunta, P. K. R., Gadekallu, T. R., & Pham, Q. V. (2021). Federated learning for cybersecurity: Concepts, challenges, and future directions. IEEE Transactions on Industrial Informatics, 18(5), 3501-3509.

 doi:10.20944/preprints202601.1728.v1

5. Gu, X., Liu, M., & Yang, J. (2025). Application and Effectiveness Evaluation of Federated Learning Methods in Anti-Money Laundering Collaborative Modeling Across Inter-Institutional Transaction Networks.
6. Khan, A. A., Alsufyani, A., Alsufyani, N., & Mohamed, M. A. (2025). BAML: a decentralized approach to secure, privacy-preserving financial compliance for enhancing anti-money laundering with blockchain hyperledger and federated learning. Peer-to-Peer Networking and Applications, 18(5), 270.
7. Yang, Y., Guo, M., Corona, E. A., Daniel, B., Leuze, C., & Baik, F. (2025). VR MRI Training for Adolescents: A Comparative Study of Gamified VR, Passive VR, 360 Video, and Traditional Educational Video. arXiv preprint arXiv:2504.09955.
8. Chowdhury, T. K., & Kudapa, S. P. (2024). Federated Learning Models for Privacy-Preserving Data Sharing And Secure Analytics In Healthcare Industry. International Journal of Business and Economics Insights, 4(4), 91-133.
9. Zhu, W., Yao, Y., & Yang, J. (2025). Real-Time Risk Control Effects of Digital Compliance Dashboards: An Empirical Study Across Multiple Enterprises Using Process Mining, Anomaly Detection, and Interrupt Time Series.
10. Rot, P., Grm, K., Peer, P., & Štruc, V. (2023). PrivacyProber: Assessment and detection of soft–biometric privacy–enhancing techniques. IEEE Transactions on Dependable and Secure Computing, 21(4), 2869-2887.
11. Li, T., Jiang, Y., Hong, E., & Liu, S. (2025). Organizational Development in High-Growth Biopharmaceutical Companies: A Data-Driven Approach to Talent Pipeline and Competency Modeling.
12. Akkaya, K., Demirbas, M., & Aygun, R. S. (2008). The impact of data aggregation on the performance of wireless sensor networks. Wireless Communications and Mobile Computing, 8(2), 171-193.
13. Hu, Z., Hu, Y., & Li, H. (2025). Multi-Task Temporal Fusion Transformer for Joint Sales and Inventory Forecasting in Amazon E-Commerce Supply Chain. arXiv preprint arXiv:2512.00370.
14. Tanim, S. A., Mridha, M. F., Safran, M., Alfarhood, S., & Che, D. (2024). Secure federated learning for Parkinson's disease: Non-IID data partitioning and homomorphic encryption strategies. IEEE Access.
15. Bai, W. (2020, August). Phishing website detection based on machine learning algorithm. In 2020 International Conference on Computing and Data Science (CDS) (pp. 293-298). ieee.
16. Hatamizadeh, A., Yin, H., Molchanov, P., Myronenko, A., Li, W., Dogra, P., ... & Roth, H. R. (2023). Do gradient inversion attacks make federated learning unsafe?. IEEE Transactions on Medical Imaging, 42(7), 2044-2056.
17. Sheu, J. B., & Gao, X. Q. (2014). Alliance or no alliance—Bargaining power in competing reverse supply chains. European Journal of Operational Research, 233(2), 313-325.
18. Zhu, W., Yao, Y., & Yang, J. (2025). Optimizing Financial Risk Control for Multinational Projects: A Joint Framework Based on CVaR-Robust Optimization and Panel Quantile Regression.
19. Rahmati, M. (2025). Real-Time Financial Fraud Detection Using Adaptive Graph Neural Networks and Federated Learning. Int. J. Management and Data Analytics, 5(1), 98-110.
20. Wang, J., & Xiao, Y. (2025). Research on Credit Risk Forecasting and Stress Testing for Consumer Finance Portfolios Based on Macroeconomic Scenarios.