Beyond the Black Box: A Framework for Explainable and Attack-Resilient Federated Anomaly Detection in IIoT

Andrea Pinto
Systems and Computing Engineering Department, School of Engineering
Universidad de los Andes
Bogotá, Colombia
ya.pinto10@uniandes.edu.co

Yezid Donoso
Systems and Computing Engineering Department, School of Engineering
Universidad de los Andes
Bogotá, Colombia
ydonoso@uniandes.edu.co

Jairo A. Gutierrez
Department of Computer Science and Information Sciences
Auckland University of Technology
Auckland, New Zealand
jairo.gutierrez@aut.ac.nz

***Abstract*— Federated Learning (FL) is a promising paradigm for anomaly detection in Industrial Internet of Things (IIoT) environments. However, existing FL frameworks suffer from vulnerabilities such as model poisoning attacks, privacy leakage, and a lack of model interpretability, which is critical for IIoT environments. This paper introduces a novel framework, Federated Learning with Explainable Anomaly Signals (FL-EAS), designed to overcome these limitations. FL-EAS fundamentally alters the federated learning process by exchanging compact, 21-dimensional feature vector derived from the reconstruction errors of local, explainable models, rather than raw model parameters. The framework incorporates a server-side supervised classifier to detect and reject malicious contributions, thereby ensuring attack resilience. By propagating explainability from the client edge to the global model, FL-EAS provides transparent, human-interpretable results. The efficacy of this approach is contextualized for evaluation using the physical process data from the BATADAL 2.0 dataset, demonstrating a state-of-the-art F1-score of 0.9511 on concealed attacks, and demonstrating its potential for secure, efficient, and trustworthy anomaly detection in real-world Cyber-Physical Systems.**

Keywords—Federated Learning, Industrial Internet of Things (IIoT), Anomaly Detection, Explainable AI (XAI), Cybersecurity, Cyber-Physical Systems

I. Introduction And Related Work

The adoption of Federated Learning (FL) for intrusion detection in the Industrial Internet of Things (IIoT) is challenged by an expanding attack surface [1], privacy risks, vulnerability to concealed attacks, and the opaque 'black box' nature of deep learning models. To address this gap, we present the Federated Learning with Explainable Anomaly Signals (FL-EAS) framework. FL-EAS redefines the information exchanged in FL to simultaneously enhance security, privacy, and efficiency. Its novelty lies in using a reconstruction error-based feature vector as an anomaly signal, enabling a proactive defense against subtle threats by identifying anomalies that present a signature of near-zero reconstruction error, and an end-to-end explainable methodology that traces alerts to specific physical components [2], [3], [4].

While existing FL frameworks address challenges like client heterogeneity (e.g., FedMint [5], ZeKoC [6]) and security through computationally intensive cryptographic methods (e.g., DeepFed [7], DetectPMFL [8], DPAD [9]), they often overlook the specific threat of concealed attacks and lack integrated explainability. Some approaches add post-hoc explainability modules (e.g., AI4FIDS [10], IP2FL [11]), treating it as a separate component rather than part of the core design. FL-EAS diverges by using explainable features—the reconstruction errors—as the very basis of the federated exchange. This integrated approach allows for end-to-end interpretability and a lightweight, robust defense against poisoning attacks, closing a critical gap for trustworthy IIoT security.

II. Proposed Framework And Methodology

The framework's novelty lies in its hybrid approach, which leverages the strengths of both unsupervised and supervised models. It fundamentally redefines the federated exchange by using the reconstruction errors from local autoencoders as the input for a global supervised model, thereby enhancing security, efficiency, and explainability. We validated our framework on the BATADAL 2.0 dataset, a high-fidelity simulation of a water distribution network ideal for studying concealed attacks in a Cyber-Physical System [12].

A rigorous preprocessing pipeline was applied to prepare the raw simulation data. The workflow consists of four main steps: (1) Data Consolidation, where 51 weeks of normal operation data are combined; (2) Feature Selection, where only columns representing the physical state (Tanks, Junctions, Pumps, Valves) are retained; (3) Centralized Scaling, where a single MinMaxScaler is fitted on all normal data to establish a common [0, 1] range; and (4) Time-Series Sequencing, where the scaled data is transformed into sequences of 48 consecutive readings. To accurately simulate a real-world IIoT environment, the preprocessed data is then partitioned into 21 distinct clients based on physical components (e.g., Tank Monitors, Pump Controllers). This partitioning results in a highly non-IID data distribution, creating a realistic and challenging environment for the federated learning framework. The framework is composed of three primary stages: local client-side processing, a critical diagnostic pivot, and a supervised server-side aggregation and analysis.

A. Client-Side Processing and Diagnostic Analysis

Each of the 21 clients trains a local Long Short-Term Memory (LSTM) autoencoder on its time-series data to learn

a representation of its normal behavior. Initial experiments using the clients' high-dimensional latent space embeddings for global anomaly detection yielded a poor F1-score of approximately 0.58. The pivotal insight was that the scalar reconstruction error itself was a far more potent feature: subtle, concealed attacks produced an abnormally low reconstruction error ('too clean' signal), while noisy attacks resulted in a very high one. Our analysis confirmed this phenomenon, which we term the 'too clean' attack signature. Because concealed attacks often use simple, predictable signals, the local autoencoder can reconstruct them almost perfectly. This perfect reconstruction results in an abnormally low Mean Squared Error (MSE), a value often lower than that of normal operational data, which becomes a distinct feature for detection.

B. Federated Exchange and Server-Side Aggregation

Instead of exchanging high-dimensional embeddings, each client calculates its mean reconstruction error and transmits this single value to the central server. The global model's input is therefore a compact, 21-dimensional feature vector, with each component corresponding to one of the 21 client reconstruction errors. The server-side model is a supervised XGBoost (Extreme Gradient Boosting) classifier. The use of a supervised classifier like XGBoost is critical for this cybersecurity application. Unlike unsupervised algorithms like Isolation Forest or OCSVM, which are primarily designed to identify any deviation from a "normal" profile, XGBoost is a powerful supervised classifier that excels at learning complex, non-linear relationships and distinct patterns within data. This allows the model to specifically detect the subtle, deceptive signature of a concealed attack, a key advantage over previous unsupervised methods. It is trained on a composite dataset containing normal, baseline, and concealed attack data, enabling it to learn and classify distinct anomaly patterns rather than treating all deviations as a single class of anomaly.

III. Results And Discussion

Our framework was rigorously evaluated on a test set from the BATADAL 2.0 dataset, with a specific focus on its ability to detect concealed attacks. On the BATADAL 2.0 test set, FL-EAS achieved a state-of-the-art F1-score of 0.9511 on concealed attacks, with a recall of 0.9989 and precision of 0.9076, demonstrating its potential for reliable operational deployment. The FL-EAS framework offers significant and demonstrable advantages that represent a paradigm shift over traditional federated learning architectures:

A. Enhanced Privacy, Communication Efficiency and Attack Resilience

The framework's novel information exchange—transmitting only a scalar reconstruction error per client—provides a trifecta of benefits: enhanced privacy, communication efficiency, and attack resilience. First, privacy is significantly enhanced because the exchanged error value is an aggregated performance metric, not raw data or model parameters, making it nearly impossible to reverse-engineer the sensitive underlying system data. Second, communication efficiency is drastically improved by transmitting a single value per client instead of the thousands or millions of parameters typical in FL, and it avoids the significant computational overhead of complex cryptographic methods. Finally, this information-lean approach provides inherent attack resilience. By minimizing the data sent, the framework drastically reduces the attack surface for data poisoning, a common vulnerability in frameworks that exchange complex model weights.

B. End-to-End Explainability

The framework provides complete transparency by leveraging the intrinsic explainability of the local autoencoders and the feature importance analysis of the global XGBoost model. This is visually demonstrated in Fig. 1, which shows the XGBoost feature importance scores for each client's reconstruction error signal.

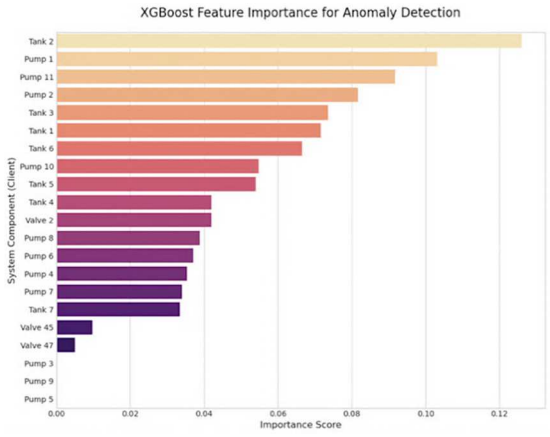

Fig. 1. XGBoost feature importance scores, highlighting the system components most influential in detecting anomalies.

The framework's transparency is demonstrated by the global XGBoost model's feature importance scores (Fig. 1), which show that critical components like Tank 2 and Pump 1 are the primary drivers for anomaly detection. This allows operators to immediately identify which parts of the system are most sensitive to disruption. This capability allows operators to trace an anomaly alert back to the specific physical components that caused it, a crucial feature for building trust and facilitating rapid incident response.

C. An End-to-End Explanation of a Concealed Attack

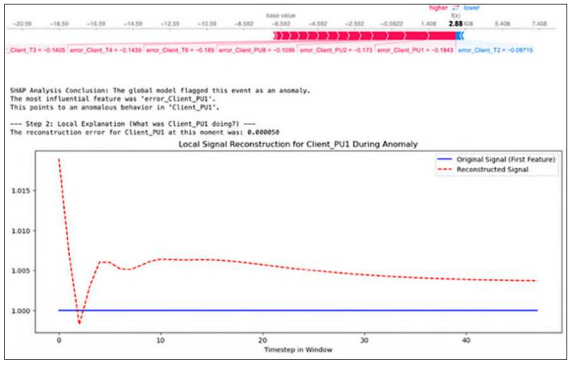

Fig. 2. End-to-end explanation of a concealed attack. (a) A SHAP force plot explains the global model's classification. (b) The corresponding local signal reconstruction at Client_PU1 reveals the 'too clean' attack signature.

The end-to-end interpretability of the framework is best demonstrated by a case study of a specific attack scenario that represents a classic learning-based concealment attack as described in the BATADAL 2.0 paper. Our framework provides both a global and a local explanation for a model's decision, allowing for a complete understanding of the event.

A SHAP force plot for this event (Fig. 2 (a)) explains why the global XGBoost model made its decision. The Base Value represents the average prediction score across the entire dataset. Each red arrow signifies a feature that pushed the model's prediction higher (towards "Attack"). The most influential features in this case were the error scores from Client_PU1, Client_PU8, Client_PU2, and Client_T6. Their low values strongly indicated an anomaly. Conversely, the blue arrows show features, such as the error from Client_T2, that were in a more normal range and slightly pushed the prediction back towards "Normal." The final prediction score of f(x)=2.88 is significantly higher than the base value, causing the model to confidently classify this event as an attack. Drilling down into the most influential client, Client_PU1, reveals the root cause (Fig. 2 (b)). The pump's original signal is a perfectly flat, unchanging line, which is unnatural for an operating sensor. The autoencoder reconstructed this simple signal almost perfectly, resulting in an extremely low reconstruction error (0.000050). This is the classic signature of a 'too clean' concealed attack, confirming the event was an attack that manipulated Pump 1 by feeding it an unnaturally stable signal.

D. Comparative Analysis

A direct quantitative comparison of FL frameworks is challenging due to dataset heterogeneity. While frameworks like DeepFed [7] and IP2FL [11] report high F1-scores, these results are on dated (SWaT) or general IoT datasets (Edge-IIoTset) not designed for modern IIoT threats. In contrast, FL-EAS achieves a competitive F1-score of 0.9511 on the more rigorous and relevant BATADAL 2.0 dataset. The primary novelty of our framework lies in its unique contributions. To our knowledge, this is the first work to specifically target concealed attacks using this federated anomaly detection technique. This contrasts with other XAI systems like AI4FIDS [10], which reported a lower F1-score of 0.8742 on a less challenging dataset. Furthermore, our approach provides this proactive defense without the significant computational overhead of cryptographic methods found in other systems, making it more suitable for resource-constrained IIoT devices.

CONCLUSION

This work validates that shifting the federated learning paradigm from exchanging model parameters to a concise vector of reconstruction errors creates a framework that is simultaneously attack-resilient, efficient, and transparent. By leveraging this information-lean exchange, the proposed FL-EAS framework effectively counters sophisticated concealed attacks by identifying anomalies that present a signature of near-zero reconstruction error. Evaluated on the rigorous BATADAL 2.0 dataset, the framework achieved a state-of-the-art F1-score of 0.9511 against these threats, with a high recall of 0.9989. Crucially, the framework provides end-to-end explainability with minimal communication overhead, allowing operators to trace an anomaly alert back to the specific physical components that caused it. This capability represents a significant step towards deploying secure and trustworthy AI that not only detects threats but also facilitates rapid and targeted incident response in critical IIoT infrastructure.

LIMITATIONS AND FUTURE WORK

While effective, this framework has limitations that present clear avenues for future research. The current reliance on centralized data scaling could be replaced by fully decentralized normalization techniques to better align with FL principles. Additionally, the framework was not benchmarked against a non-private, centralized model, which would serve as an ideal performance baseline. The supervised server model, which requires prior knowledge of attack types, could be evolved into a semi-supervised system capable of detecting novel, zero-day threats. Finally, future work will focus on creating a multi-modal framework that integrates industrial network traffic with the physical process data to build a more robust and holistic intrusion detection system.

REFERENCES

[1] N. Mehta, N. Bharot, J. G. Breslin, and P. Verma, "PPFL-DCS: Privacy-Preserving Federated Learning Using Neural Transformer and Leveraging Dynamic Client Selection to Accommodate Data Diversity," *IEEE Access*, vol. 13, pp. 94225–94238, 2025, doi: 10.1109/ACCESS.2025.3572605.

[2] V. C. Gogineni, S. Werner, F. Gauthier, Y. F. Huang, and A. Kuh, "Personalized Online Federated Learning for IoT/CPS: Challenges and Future Directions," *IEEE Internet of Things Magazine*, vol. 5, no. 4, pp. 78–84, Dec. 2022, doi: 10.1109/IOTM.001.2200178.

[3] H. Zhang *et al.*, "Large scale foundation models for intelligent manufacturing applications: a survey," *J Intell Manuf*, 2025, doi: 10.1007/s10845-024-02536-7.

[4] N. Sharma and P. G. Shambharkar, "Multi-layered security architecture for IoMT systems: integrating dynamic key management, decentralized storage, and dependable intrusion detection framework," *International Journal of Machine Learning and Cybernetics*, 2025, doi: 10.1007/s13042-025-02628-7.

[5] O. Wehbi *et al.*, "FedMint: Intelligent Bilateral Client Selection in Federated Learning With Newcomer IoT Devices," *IEEE Internet Things J*, vol. 10, no. 23, pp. 20884–20898, Dec. 2023, doi: 10.1109/JIOT.2023.3283855.

[6] Z. Chen, P. Tian, W. Liao, and W. Yu, "Zero Knowledge Clustering Based Adversarial Mitigation in Heterogeneous Federated Learning," *IEEE Trans Netw Sci Eng*, vol. 8, no. 2, pp. 1070–1083, Apr. 2021, doi: 10.1109/TNSE.2020.3002796.

[7] B. Li, Y. Wu, J. Song, R. Lu, T. Li, and L. Zhao, "DeepFed: Federated Deep Learning for Intrusion Detection in Industrial Cyber-Physical Systems," *IEEE Trans Industr Inform*, vol. 17, no. 8, pp. 5615–5624, Aug. 2021, doi: 10.1109/TII.2020.3023430.

[8] Z. Zhang, N. He, Q. Li, K. Wang, H. Gao, and T. Gao, "DetectPMFL: Privacy-Preserving Momentum Federated Learning Considering Unreliable Industrial Agents," *IEEE Trans Industr Inform*, vol. 18, no. 11, pp. 7696–7706, Nov. 2022, doi: 10.1109/TII.2022.3140806.

[9] S. Basak and K. Chatterjee, "DPAD: Data Poisoning Attack Defense Mechanism for federated learning-based system," *Computers and Electrical Engineering*, vol. 121, Jan. 2025, doi: 10.1016/j.compeleceng.2024.109893.

[10] A. Karampasi *et al.*, "Towards Transparent AI-Powered Cybersecurity in Financial Systems: The Deployment of Federated Learning and Explainable AI in the CaixaBank pilot," in *IEEE International Conference on Data Mining Workshops, ICDMW*, IEEE Computer Society, 2024, pp. 270–277. doi: 10.1109/ICDMW65004.2024.00041.

[11] D. Namakshenas, A. Yazdinejad, A. Dehghantanha, R. M. Parizi, and G. Srivastava, "IP2FL: Interpretation-Based Privacy-Preserving Federated Learning for Industrial Cyber-Physical Systems," *IEEE Transactions on Industrial Cyber-Physical Systems*, vol. 2, pp. 321–330, Jul. 2024, doi: 10.1109/ticps.2024.3435178.

[12] A. Erba, A. F. Murillo, R. Taormina, S. Galelli, and N. O. Tippenhauer, "On Practical Realization of Evasion Attacks for Industrial Control Systems," in *RICSS 2024 - Proceedings of the 2024 Workshop on Re-design Industrial Control Systems with Security, Co-Located with: CCS 2024*, Association for Computing Machinery, Inc, Nov. 2024, pp. 9–25. doi: 10.1145/3689930.3695213.