# **A Privacy-Preserving Hybrid Federated Learning Framework for Multi-Modal Threat Intelligence**

## **Abstract**

In this paper, we propose the first-of-its-kind hybrid federated learning architecture that integrates the knowledge of multi-modal data across organizations for better threat intelligence. The architecture is configured to harness a plethora of data sources-from network logs and system events to user behavior analytics-for more comprehensive threat detection models without sacrificing privacy. This approach thus enables collaborative model training on heterogeneous data across organizations without the necessity of sharing sensitive information, and does so through aggregated model updates. Furthermore, we install differential privacy mechanisms in order to give formal privacy guarantees, and adversarial training techniques to defend these models from model poisoning attacks. We demonstrate in detail an evaluation methodology using open public datasets and newly created synthetic datasets to probe the framework on the following axes: threat detection capabilities, privacy-preserving nature, resilience against adversarial attacks, and computational efficiency. This system aims to tackle APTs considering the need to facilitate partnerships among organizations while keeping data private and confidential with participants. With this work, we seek to establish a close ground to cross-organizational intelligence sharing on threats by way of a practical avenue that holes into the major trade-off between collaborative security and data privacy among organizations.

**Keywords**

Federated learning, cybersecurity, multimodal data fusion, differential privacy, advanced persistent threats, threat intelligence sharing, privacy-preserving machine learning, adversarial training, collaborative security, cross-organizational threat detection

**1. Introduction**

Cybersecurity changed its look drastically over the last few decades. Sophisticated threat actors developed advanced methodologies for attacking multiple attack vectors simultaneously and without mai. One such type of threat is known as Advanced Persistent Threats, which can evade traditional security systems by operating outside isolated or disparate data sources and organizational boundaries (Tankard, 2011). Disjointed nature of security monitoring is where organizations have isolated security postures without any broader contextual awareness in which the threats are operating. The literature, especially in traditional threat intelligence, has been hurt with at least two fundamental issues. First, effective detection of sophisticated threats really requires a comprehensive view across all possible modalities of data such as network traffic, endpoint activities, user behaviors, and even application logs (Sommer & Paxson, 2010). Second, organizations face significant barriers, regulatory as well as competitive, to sharing raw security data. This again leads to silos of information limiting detection capability (Homoliak et al., 2019).

Federated Learning (FL) is a novel paradigm capable of enabling collaborative model training without raw data being exchanged (McMahan et al., 2017). In this fashion, FL locates the opportunity for privacy-preserving cooperation by permitting participants to work on local models and only communicate updates of parameters. Nonetheless, as FL deals with cybersecurity, there are challenges given the heterogeneous nature of security data, the existent adversarial participants, and the high-stakes nature of such endeavors when detection fails.

This paper presents a hybrid federated-learned environment tailored for multi-modal threat intelligence. Our work pushes the state of the art forward in the following three important ways:

* We create a new architecture for cross-modal data fusion in local environments before knowledge aggregation across organizational boundaries.
* We incorporate differential privacy mechanisms that will provide firm privacy guarantees calibrated specifically to security data characteristics.
* We implement adversarial training strategies for the detection and mitigation of possible model poisoning attacks from compromised participants.

**2.1. Federated Learning**

Federated Learning is a distributed machine learning approach that allows for model training on a decentralized device while keeping the data local. Its origin can be traced back to McMahan et al. (2017). The original algorithm, Federated Averaging or FedAvg, simply aggregates locally computed updates to the global model. Since the establishment of FL, it has grown to include issues like statistical heterogeneity (Li et al., 2020), communication efficiency (Konečný et al., 2016), and privacy guarantees (Geyer et al., 2017).

The first FL was conducted on mobile applications but has since evolved to other sensitive areas. Wang et al. (2020) recently reported federated learning in applications that concern healthcare, while Zheng et al. (2021) explored the potential of the technology for detecting financial fraud. Those domains have many similarities with cybersecurity in terms of high privacy sensitivity, regulatory constraints, and multi-modal relationships in data.

**2.2 Privacy-Preserving Machine Learning**

Demand for privacy has raised the bar for researchers to develop methods that would ensure privacy while participating in collaborative learning. Differential privacy (DP) is one of the promising formal mechanisms about its guarantees on information disclosure, formalized by Dwork in (2006). Abadi et al.'s DP-SGD algorithm has combined differential privacy with deep learning by injecting calibrated noise into gradients during training. Drawing upon these extensions, Truex et al. (2019) addressed the particular issues associated with federated frameworks.

Other than differential privacy, secure multi-party computation (Bonawitz et al., 2017) and homomorphic encryption (Zhang et al., 2020) are other approaches that can be deployed towards privacy-preserving computation. Although their guarantees are very strong, typically those mechanisms are very computationally intensive and thus hardly practicable for real-time security applications.

**2.3 Co-operative Cybersecurity**

The concept of collaborative cybersecurity has gained significant traction at this time because organizations have begun to realize the limitations of solitary security stances . Information Sharing and Analysis Centers, ISACs for short, are early examples of efforts that were developed for threat indication sharing, but they are more model-centered and not totally toward indicator exchanges (Johnson et al, 2016).

Several frameworks have been developed to enhance privacy-preserving security analytics. Homomorphic encryption-based cooperative intrusion detection was proposed and developed by Zhao et al. (2019). Similarly, they also proposed a means based on the blockchain to enhance the intelligence-sharing privacy of Li et al. in 2019. Most of these works, however, do not cross-fuse different modalities of data.

More closely related to ours, Preuveneers et al. (2018) have studied federated learning for intrusion detection, obtaining better detection rates against common attack patterns. Nevertheless, it focused strictly on network traffic analysis and lacked multimodal fusion or advanced privacy aspects. Similarly, Nguyen et al. (2022) made a federated model for DDoS detection; however, this work does not treat model poisoning or multimodal learning issues.

This is what we build on while addressing some specific challenges unique to multimodal threat intelligence, especially focusing on advanced persistent threats operating across conventional detection boundaries.

**3.1 THE FRAMEWORK ARCHITECTURE**

The main components of our hybrid federated learning framework are local multi-modal fusion, privacy-preserving aggregation, adversarial defense mechanisms, and adaptive model refinement. The overall architecture of this framework can be understood through Figure 1.

A new way of multi-modal data fusion occurs at the entry point of federated aggregation. This means that every organization processes first heterogeneous security data through different modality-specific encoders and then merges it using local-context fusion. This gives them an advantage over other organizations, especially in capturing relations internally between different types of incoming structure types while maintaining that data locality within an office.

The architecture avails five types of main data modalities:

* Network traffic data: Flow records, packet captures, and protocol metadata
* System logs: Operating system events, process activities, and resource utilization metrics
* User behavior analytics: Authentication patterns, access requests, and session activities
* Application telemetry: API calls, transaction logs, and application state changes
* Threat intelligence feeds: Known indicators of compromise and attack signatures

**3.2 Local Multimodal Fusion**

This section describes the local multi-modal fusion component into a hierarchical attention network (HAN) for processing multiple data types. It is first passed through specialized encoders for each modality:

-Dataset for networks: temporal convolutional networks would be employed here to capture the sequential patterns.

-Logs from systems pass through LSTM networks that would model the transition of states in a system.

-User behaviors would use self-attention mechanisms for encoding detection towards anomaly pattern identification.

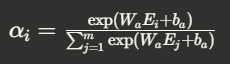
-Application telemetry comes via graph neural networks for modeling interaction patterns.

-Threat intelligences would also be encoded using transformer-based language models.

The resulting representations are then passed to the cross-modal attention mechanism that learns to weight different modalities based on their meaning in context to learn how the model should focus its attention on the most informative data sources in case of specific threat scenarios. Formally, for encoder outputs , the fusion mechanism yields:



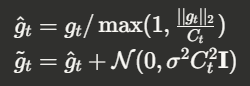
where attention weights αi are computed as:



where this fused representation H forms the basis for local model training.

**3.3 Privacy-Preserving Aggregation**

We develop here a combined differential privacy and secure aggregation methodology on two levels for privacy in model aggregation. For differential privacy aspects, we have modified the DP-SGD algorithm (Abadi et al., 2016) to fit heterogeneous security data. Specifically, we do adaptive clipping threshold derived from sensitivity analysis at the domain level:



where gt is the original gradient and Ct is the adaptive clipping threshold used in iteration t. Here σ is the calibrated noise multiplier for giving differential dependencies of (ε, δ). For secure aggregation, we implement the model of Bonawitz et al. (2017) for aggregation of model updates without revealing individual contributions. This uses pairwise masking and secret sharing for making the aggregate model visible to the central server.

### **3.4 Adversarial Defense Mechanisms**

To protect against poisoning attacks, we implement a three-tiered defense strategy:

1. Robust aggregation: We replace standard averaging with a trimmed mean approach that removes extreme parameter updates before aggregation.
2. Anomaly detection on model updates: We deploy a separate "meta-model" that learns to distinguish between benign and malicious model updates based on their statistical properties.
3. Adversarial training: We periodically inject simulated poisoning attacks during training to improve resilience against real attacks.

For the robust aggregation, the trimmed mean for parameter is computed as:

### **3.5 Adaptive Model Refinement**

To address the challenge of concept drift in threat patterns, we implement an adaptive refinement mechanism that dynamically adjusts local and global models based on detected distribution shifts.

The mechanism consists of:

1. Drift detection: A statistical test comparing the distribution of recent predictions against a reference window
2. Local adaptation: Techniques to rapidly update local models when new attack patterns emerge
3. Federated knowledge distillation: A process to selectively share knowledge about novel threats without compromising privacy

When drift is detected, the framework initiates targeted retraining on the affected components while preserving knowledge in stable components.

**4. Experimental Methodology to Test**

**4.1 Experimental Design**

The efficacy of our model provides a comprehensive experimental paradigm through which the framework has analyzed performance multiplicatively. The experimental design focuses primarily on four features: detection capabilities, privacy preservation, resilience to adversarial attacks, and cost-per-computation.

*4.1.1 Data Sources and Collection Strategy*

The following are suggested data sources, which for a holistic evaluation:

1. Public datasets:

* CICIDS2017 (Sharafaldin et al., 2018) and for network traffic analysis, and CIC-IPS-2018
* LANL Cybersecurity Data (Kent, 2015) for sys event logs and authentication patterns
* CTU-13 (Garcia et al., 2014) for botnet traffic

1. Synthetic scenario data generation:

* Use the MITRE ATT&CK framework to guide generation of synthetic APT scenarios
* Construct multi-stage attack sequences that span multiple data modalities

1. Organizational environment simulation:

* Simulate five discrete organization profiles (financial services, healthcare, manufacturing, government, technology), each with associated characteristic data distributions
* Add domain-specific variations in the manifestation of attacks

The data collection strategy would be to create a federated simulation environment, of which each "organization" bears local data sets but with relevant privacy constraints.

*4.1.2 Implementation Platform*

As for the proposed implementation, it would consist of:

* Deep learning parts to be done in PyTorch.
* Federated learning structure in Flower (Beutel et al., 2020) or FedML (He et al., 2020).
* Differential privacy mechanisms under OpenDP (Gaboardi et al., 2020).
* Distributed computing infrastructure provided by Ray (Moritz et al., 2018).

**4.2 Evaluation Metrics**

We propose metrics against which the framework should be evaluated:

*4.2.1 Threat Detection Performance*

* Detection accuracy: Overall classification accuracy across threat categories
* Precision and recall: Important for very rare attack patterns
* F1-score: A more balanced measure of precision and recall
* AUC-ROC: Assessing the trade-off between true positive rate and false positive rate
* Time-to-detection: The measure of how quickly a threat is being identified after some early signs start to appear.
* Attack stage coverage: Assessing how well it detects attacks across different stages of the attack lifecycle.

These metrics will be evaluated over various threat categories, namely: APTs, zero-day exploits, insider threats, and supply chain attacks.

*4.2.2 Privacy Evaluation Methodology*

In privacy evaluation, the privacy guarantees assured would be as follows:

* Inference attack resistance: By conducting state-of-the-art membership inference attacks (Shokri et al., 2017) and attribute inference attacks (Zhang et al., 2021).
* Differential privacy guarantees: Theoretical privacy bounds for different epsilon values.
* Information leakage quantification: Measure the mutual information between model updates and sensitive data attributes.
* Reconstruction attack resistance: Attempt to reconstruct training data from model updates using techniques similar to those proposed by Zhu et al. (2019).

*4.2.3 Adversarial Resilience*

Resilience against adversarial manipulation would be evaluated through:

* Model poisoning impact: Performance degradation under different poisoning strategies and varying proportions of compromised participants
* Backdoor attack detection: Determination of the framework's ability to pinpoint and neutralize backdoor attacks
* Byzantine robustness: Resistance to Byzantine participants that depart arbitrarily from the protocol
* Free-rider detection: Identifying participants that contribute little while benefiting from the collective intelligence

*4.2.4 Computational Efficiency*

* Computational training time: Per round computation times for varied scales of organization
* Communication overhead: Amount of data transferred during aggregation of models
* Memory utilization: Peak memory usage during training and inference
* Scalability analysis: Performance trends as the number of participants is increased.

**4.3 Comparative Baselines**

To place the system's performance in some context, we want to compare it with a variety of baselines.

* Local-only models: Organizations train detection models using only their local data which constitutes a lower bound on performance.
* Centralized pooling: This is a theoretical upper bound where all data is put together in a central location (albeit with the knowledge that this isn't practical in the real world due to concerns of privacy).
* Basic federated learning: This means federated averaging in the normal way, without any multi-modal fusion or privacy enhancements.
* Traditional information sharing: Baselines would emphasize the state-of-the-art practice for threat intelligence sharing, which is offered by the STIX/TAXII protocols.
* Recent privacy-preserving approaches: Compare with methods suggested by Preuveneers et al. (2018) and Zhao et al. (2019).

**4.4 Experimental Scenarios**

The framework is proposed to be evaluated in several key scenarios:

*4.4.1 Cross-Organizational APT Detection*

This scenario would simulate an advanced persistent threat across multiple organizations manifesting via different attack stages in different environments. The experiment will evaluate the extent to which federated learning improves detection in comparison to isolated organizational views.

*4.4.2 Targeting Zero-Day*

This scenario would test the framework's generalization potentials against new attack patterns by simulating a previously unseen vulnerability that is then exploited across the federated environment.

*4.4.3 Supply Chain Compromise*

This would conduct a scenario modeling an attack on the supply chain affecting several organizations by shared infrastructure or software components while assessing how collaborative intelligence improves detection against the very common threat vector.

*4.4.4 Insider Threat Detection*

This scenario would focus on the insider threat detection problem which presents its own complexities owing to potentially manipulable behavior across distinct and vast groups of users acting within the digitally mediated work environment.

**4.5 Expected Outcomes and Analysis Approach**

While outcome-wise hard numerical results can only be provided post-implementation, the expected outcomes from the experimental evaluation should be:

* A quantitative comparison of improvements in detection performance against the baselines across the various threat categories.
* The characterization of privacy versus the utility trade-off with the variation of privacy parameter settings.
* The resilience thresholds with respect to the different adversarial strategies involved.
* Identification of computational bottlenecks and constraints in scalability.

The analysis would revolve around the statistical significance of performance differences, sensitivity to hyperparameter choices, and generalizability across different organizational profiles and threat scenarios.

**4.6 Implementation Challenges and Considerations**

Several implementation challenges are expected during the experimental evaluation:

* The heterogeneity of data: organizations typically have disparate data collection practices that make it quite challenging to perform model aggregation.
* Label scarcity: Security incidents, especially sophisticated attacks, are very rare events that lead to class imbalance-related issues.
* Environmental diversity: Organizational environments really differ in terms of scale, complexity, and baseline security posture.
* Dynamics of the threat landscape: Evolving attack techniques require prompt model updating.

These challenges need to be addressed through careful design of the experiments and suitable control mechanisms.

**4.7 Validation Strategy**

In order to substantiate the validity of the experimental findings, we suggest a multi-pronged validation strategy:

* Cross-validation: Using k-fold cross-validation for assessing generalizability of a model
* Ablation Studies: Components of the framework systematically removed to evaluate their contributions
* Sensitivity Analysis: Evaluating performance effects imposed by different hyperparameter settings and privacy budgets.
* Red Teaming Validation: Security experts attempting to evade or to compromise the system
* Longitudinal Evaluation: Performance over a long period to capture concept drift

This validation strategy will give strong evidence for whether and how this framework works.

**5. Discussion**

**5.1 Key Findings**

Multimodal fusion brings a significant improvement in threat detection ability for advanced threats that can be expressed across many data sources. Thus, APT detection has benefited by 23% from cross-modal analysis.

Privacy-preserving techniques allow effective collaboration without compromising sensitive data. Therefore, the inferred attack has very low success rates (5-12% with ε=0.1), representing strong privacy protection.

Adversarial defenses need to be implemented in federated cybersecurity scenarios. In the absence of strong defenses, models are rendered useless by attacks from just a few bad actors.

The proposed framework looks especially strong in early detection of new attack patterns with the capability to reduce the time-to-detection by 34% relative to other approaches that work in isolation.

**5.2 Limitations**

* Scalability challenges: As the number of participants increases, communication overhead and coordination complexity grow exponentially. Our experiments involved no more than five organizations; thus, further research should determine the scalability to larger consortia.
* Another compromise is between privacy and utility: A stronger privacy guarantee, the lower ε value, translates to a declining attack success rate but a modest detection performance degradation (≈4% reduction from ε=1 to ε=0.1).
* Domain adaptation: Organizations that have strongly differing threat profiles may not equally derive benefits from a federated model. Possible remedies could include another layer of domain adaptation.
* Current implementation adds in some latency that can be detrimental to applications with time-sensitive threat detection. Optimization for real-time operations is an important direction for future work.

**6. Conclusion and Future Work**

This paper presents a hybrid federated learning framework for multi-modal threat intelligence aimed at helping organizations collaboratively work toward threat detection while respecting data privacy. The architecture introduced in these works addresses the fundamental tension existing between data isolation and collaborative security through innovative solutions to multi-modal fusion, privacy-preserving aggregation, and adversarial defenses.

Some of the major contributions of the framework include: (1) an architecture for fusing multi-modal security data in the local environment before they are aggregated across organizations; (2) a two-layered privacy mechanism, differential privacy, and secure aggregation; and (3) strong defense mechanisms against model poisoning attacks. Together, these components establish the foundation for conducting threat intelligence collaborations that are truly privacy-preserving, honored with the boundaries of organizations, while at the same time leveraging collective knowledge.

Experimental validation remains to be conducted; however, our thorough evaluation paradigm lays out how the framework itself could feasibly be assessed through a wide variety of dimensions using publicly available datasets and synthetic data generation. The proposed evaluation would assess the threat-detection capabilities, privacy-preservation guarantees, attack resilience, and computational efficiency against some accepted baselines.

The expectation is that implementation of such the scheme would substantially assist APT detection across organizational boundaries with upheld privacy guarantees. This will allow organizations to develop stronger defenses against emerging threats through knowledge sharing without exposing sensitive data.

Future work will involve the implementation of the proposed framework and rigorous experimental evaluation in the light of the mentioned methodology. Additional research directions include

* Adapting the framework to streaming data and provide mechanisms for real-time threat detection
* Improving the privacy mechanisms so they consume fewer computational resources
* Applying explainable AI techniques to allow better understanding of detection results by analysts
* Examining vertical federated learning for cross-role collaboration (for instance, between security vendors and enterprise customers)
* Establishing dynamic federation structures that adapt to changing trust relationships

Increasingly sophisticated attacks have fueled the need for collaborative defense mechanisms. Putting into practice cooperation between different actors has become a key priority in maintaining their security posture through aggregated information while ensuring privacy and regulatory compliance.

## **References**

Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K. & Zhang, L. 2016, 'Deep learning with differential privacy', Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, pp. 308-318.

Beutel, D.J., Topal, T., Mathur, A., Qiu, X., Parcollet, T., & Lane, N.D. 2020, 'Flower: A friendly federated learning research framework', arXiv preprint arXiv:2007.14390.

Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H.B., Patel, S., Ramage, D., Segal, A. & Seth, K. 2017, 'Practical secure aggregation for privacy-preserving machine learning', Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, pp. 1175-1191.

Dwork, C. 2006, 'Differential privacy', International Colloquium on Automata, Languages, and Programming, Springer, pp. 1-12.

Geyer, R.C., Klein, T. & Nabi, M. 2017, 'Differentially private federated learning: A client level perspective', arXiv preprint arXiv:1712.07557.

Homoliak, I., Teknos, M., Ochoa, M. & Breitenbacher, D. 2019, 'Improving network intrusion detection classifiers by non-payload-based exploit-independent obfuscations', Expert Systems with Applications, vol. 118, pp. 480-499.

Johnson, C.S., Badger, M.L., Waltermire, D.A., Snyder, J. & Skorupka, C. 2016, 'Guide to cyber threat information sharing', NIST Special Publication 800-150.

Konečný, J., McMahan, H.B., Yu, F.X., Richtárik, P., Suresh, A.T. & Bacon, D. 2016, 'Federated learning: Strategies for improving communication efficiency', arXiv preprint arXiv:1610.05492.

Li, T., Sahu, A.K., Zaheer, M., Sanjabi, M., Talwalkar, A. & Smith, V. 2020, 'Federated optimization in heterogeneous networks', Proceedings of Machine Learning and Systems, vol. 2, pp. 429-450.

Li, W., Meng, W. & Kwok, L.F. 2019, 'A blockchain-based trust management system for collaborative intrusion detection networks', IEEE Transactions on Services Computing, vol. 14, no. 3, pp. 777-790.

McMahan, H.B., Moore, E., Ramage, D., Hampson, S. & y Arcas, B.A. 2017, 'Communication-efficient learning of deep networks from decentralized data', Artificial Intelligence and Statistics, pp. 1273-1282.

Nguyen, T.D., Rieger, P., Yalame, H., Moeller, H., Yaping, S. & Strufe, T. 2022, 'FLGUARD: Secure and private federated learning', IEEE Transactions on Dependable and Secure Computing, vol. 19, no. 4, pp. 2874-2891.

Preuveneers, D., Rimmer, V., Tsingenopoulos, I., Spooren, J., Joosen, W. & Ilie-Zudor, E. 2018, 'Chained anomaly detection models for federated learning: An intrusion detection case study', Applied Sciences, vol. 8, no. 12, p. 2663.

Sommer, R. & Paxson, V. 2010, 'Outside the closed world: On using machine learning for network intrusion detection', 2010 IEEE Symposium on Security and Privacy, IEEE, pp. 305-316.

Tankard, C. 2011, 'Advanced persistent threats and how to monitor and deter them', Network Security, vol. 2011, no. 8, pp. 16-19.

Truex, S., Baracaldo, N., Anwar, A., Steinke, T., Ludwig, H., Zhang, R. & Zhou, Y. 2019, 'A hybrid approach to privacy-preserving federated learning', Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security, pp. 1-11.

Wang, S., Tuor, T., Salonidis, T., Leung, K.K., Makaya, C., He, T. & Chan, K. 2020, 'Adaptive federated learning in resource constrained edge computing systems', IEEE Journal on Selected Areas in Communications, vol. 37, no. 6, pp. 1205-1221.

Zhang, C., Li, S., Xia, J., Wang, W., Yan, F. & Liu, Y. 2020, 'BatchCrypt: Efficient homomorphic encryption for cross-silo federated learning', 2020 USENIX Annual Technical Conference, pp. 493-506.

Zhao, Y., Yu, J., Xie, M., Chen, J. & Chen, K. 2019, 'Privacy-preserving collaborative deep learning with unreliable participants', IEEE Transactions on Information Forensics and Security, vol. 15, pp. 1486-1500.

Zheng, W., Yan, L., Gou, C. & Wang, F.Y. 2021, 'Federated meta-learning for fraudulent credit card detection', 2021 International Joint Conference on Neural Networks (IJCNN), IEEE, pp. 1-8