



Federated Learning for Privacy-Preserving Threat Detection: Collaborative Attack Detection Without Sharing Data in Healthcare and Finance

VINUTHA G R

Dept of Computer Science and Engineering
VTU University, Muddenahalli,
Chikkaballapur Karnataka
vinuthagr786@gmail.com

Dr.Nirmala Hiremani

Dept of Computer Science and Engineering
VTU University, Muddenahalli,
Chikkaballapur Karnataka
nirmalavtupg@gmail.com

Abstract—Cyber threats targeting healthcare and financial institutions have grown sharply with the digitization of critical services, yet effective detection increasingly depends on intelligence shared across organizations—intelligence that privacy regulations such as HIPAA, GDPR, and PCI-DSS forbid these organizations from pooling. Federated Learning (FL) resolves this tension by enabling collaborative model training in which only model updates, never raw data, are exchanged. This paper presents a comprehensive survey of FL-based threat detection and proposes the Privacy-Preserving Federated Threat Intelligence Framework (PPFTIF) for healthcare and financial ecosystems. PPFTIF integrates federated learning, differential privacy, secure aggregation, Byzantine-resilient aggregation, and Explainable AI (XAI) to achieve collaborative threat detection while preserving data confidentiality. We formalize a sector-weighted federated objective, specify the end-to-end training algorithm with its privacy and robustness steps, and analyze the framework against membership-inference, gradient-leakage, and model-poisoning attacks. A worked cross-sector example shows how a hospital and a bank jointly improve detection of a shared attack campaign without exposing any record, and a layered evaluation methodology with explicit datasets, metrics, and a privacy-budget protocol is provided. The contribution is an integrated, explainable, and regulation-aware blueprint rather than empirically validated outcomes; the design directly addresses cross-sector collaboration, poisoning resilience, explainability, and simultaneous regulatory compliance.

Index Terms—Federated learning, cybersecurity, threat detection, privacy preservation, healthcare security, financial security, differential privacy, secure aggregation, Byzantine resilience, explainable AI.

I. INTRODUCTION

Healthcare organizations, banks, insurers, and other financial institutions generate enormous volumes of security-relevant data every day. Detecting sophisticated, multi-stage cyberattacks frequently requires correlating indicators across many organizations, because an attack campaign that looks like noise to one victim may be unmistakable when its traces are pooled. Yet the very regulations that protect this data—HIPAA for protected health information, GDPR for personal data, and PCI-DSS for cardholder data, among sector-specific financial frameworks—restrict the direct sharing of the sensitive records on which such correlation would depend.

Traditional centralized machine learning resolves the correlation problem by aggregating data into a single repository, but in doing so it creates a cluster of new problems: it manufactures privacy violations and regulatory non-compliance, concentrates breach risk into a single high-value target, raises questions of data ownership and liability, and imposes high storage and transfer costs. For regulated sectors, these costs are frequently prohibitive, and so the collaboration that would most improve detection simply does not happen.

Federated Learning (FL) offers a way out. In FL, each organization trains a model locally on its own data and shares only model updates with a coordinating server, which combines them into a global model that is then redistributed. Raw data never leaves the organization. FL has shown strong potential across privacy-preserving cybersecurity tasks—intrusion detection, malware detection, anomaly detection, and fraud analytics—but a deployable framework that simultaneously serves both healthcare and finance, resists poisoning, explains its alerts, and respects multiple regulatory regimes at once remains absent. This paper surveys the field and proposes such a framework, PPFTIF. The contributions of this paper are as follows:

- A structured survey of FL-based threat detection across healthcare, finance, and IoT security, and an analysis of the research gaps that remain (Sections III–IV).
- PPFTIF: a six-layer privacy-preserving federated threat-intelligence architecture combining differential privacy, secure aggregation, Byzantine-resilient aggregation, and XAI (Section V).
- A formal, sector-weighted federated objective and the end-to-end training algorithm, with explicit privacy and robustness steps (Sections VI–VII).
- A security analysis against membership-inference, gradient-leakage, and model-poisoning attacks, and a worked cross-sector example (Sections VIII–IX).
- A layered evaluation methodology with datasets, metrics, and a privacy-budget protocol, together with expected outcomes and open problems (Sections X–XII).

II. BACKGROUND

A. Federated learning

Federated learning is a distributed machine-learning paradigm in which N participants collaboratively train a shared model while keeping their local datasets private. Centralized learning assumes access to the union of all data,

$$D = D_1 \cup D_2 \cup \dots \cup D_N \quad (1)$$

which is precisely what regulation forbids for sensitive sectors. Federated learning instead keeps each D_i resident at participant i and exchanges only model parameters with the server. The canonical aggregation rule, Federated Averaging (FedAvg), forms the global model as a sample-weighted mean of the local models:

$$w_{t+1} = \sum_i (n_i / n) w_i \quad (2)$$

where w_i are the local weights at participant i , n_i is the number of local samples, and $n = \sum_i n_i$ is the total. FedAvg lets organizations that cannot pool data nonetheless converge to a model informed by all of it.

B. Cybersecurity threat detection

Threat-detection systems identify network intrusions, malware infections, insider threats, fraudulent transactions, and anomalous behavior. In a federated setting, each participant trains its detector on local telemetry—network flows, endpoint logs, or transaction records—and contributes to a shared detector that generalizes across organizations. Recent surveys indicate that FL-based intrusion detection can match centralized detection performance while sharply reducing privacy-leakage risk, motivating its use as the substrate for cross-organization threat intelligence.

III. LITERATURE SURVEY

Research on FL for security and privacy spans healthcare analytics, collaborative intrusion detection, IoT security, privacy mechanisms, secure aggregation, and differential privacy. We review these strands and then summarize representative works in Table I.

FL in healthcare. Surveys of privacy-preserving FL for smart healthcare establish the regulatory motivation and report that FL preserves confidentiality while retaining analytic utility [1], though non-identically-distributed (non-IID) data across hospitals degrades convergence [1], [6].

Collaborative intrusion detection. FL-based intrusion-detection systems demonstrate collaborative detection across organizations and IoT environments [2], [3], [12], but several remain vulnerable to poisoning by malicious participants [2] and incur substantial communication overhead [3].

FL security and privacy. Security surveys catalogue the FL threat landscape and defensive strategies [4], [7], and anomaly-detection-based defenses identify malicious updates [8]; privacy-focused reviews analyze leakage and mechanisms in depth [5], [9] but raise scalability concerns.

Privacy mechanisms. Differential privacy provides provable guarantees at an accuracy cost [8], secure-aggregation protocols hide individual updates from the server at the price of added latency [11], and homomorphic encryption enables computation on encrypted updates [10]. These mechanisms are complementary rather than mutually exclusive.

Critical comparison. The literature thoroughly characterizes individual mechanisms and single-sector deployments, but no surveyed work integrates differential privacy, secure aggregation, Byzantine resilience, and explainability into one cross-sector threat-intelligence framework that addresses healthcare and

financial compliance simultaneously. That integration gap is the target of PPFTIF.

TABLE I. Comparative Survey of FL-Based Threat-Detection Systems

Ref.	Year	Domain	Method	Key contribution	Limitation
[1]	2022	Healthcare	FL	Privacy-preserving medical analytics	Non-IID data issues
[2]	2024	Cybersecurity	FL-IDS	Collaborative intrusion detection	Vulnerable to poisoning
[3]	2024	IoT security	Privacy-preserving FL	Secure IDS architecture	Communication overhead
[4]	2024	FL security	Security survey	Threat taxonomy	No deployment framework
[5]	2024	Privacy	Privacy-preserving FL	Comprehensive privacy analysis	Scalability concerns
[6]	2025	Healthcare	Secure FL	Regulatory compliance	Computational cost
[7]	2025	Secure aggregation	Aggregation protocols	Enhanced privacy	Latency increase
[8]	2025	Differential privacy	DP-FL optimization	Better privacy guarantees	Accuracy trade-off

IV. RESEARCH GAP AND DESIGN GOALS

Synthesizing the survey, existing solutions exhibit four recurring limitations, from which we derive the goals that PPFTIF must meet.

- **G1 — Cross-sector collaboration.** Most frameworks address either healthcare or cybersecurity, not both; PPFTIF must support joint training across heterogeneous sectors.
- **G2 — Poisoning resilience.** Many FL systems remain vulnerable to Byzantine participants and malicious updates; aggregation must be robust to them.
- **G3 — Explainability.** Security analysts require human-readable justifications for AI-generated alerts, so explanation must be a first-class output.
- **G4 — Regulatory compliance.** Few frameworks address HIPAA and financial compliance at once; privacy guarantees must be provable and auditable across regimes.

V. PROPOSED FRAMEWORK (PPFTIF)

The Privacy-Preserving Federated Threat Intelligence Framework (PPFTIF) is organized as six layers, shown in Fig. 1. Telemetry is processed locally; only privacy-protected model updates ever leave an organization; and the global model, its explanations, and compliance reports flow back to participants. The end-to-end exchange between participants and the aggregation server in a single training round is depicted in Fig. 2.

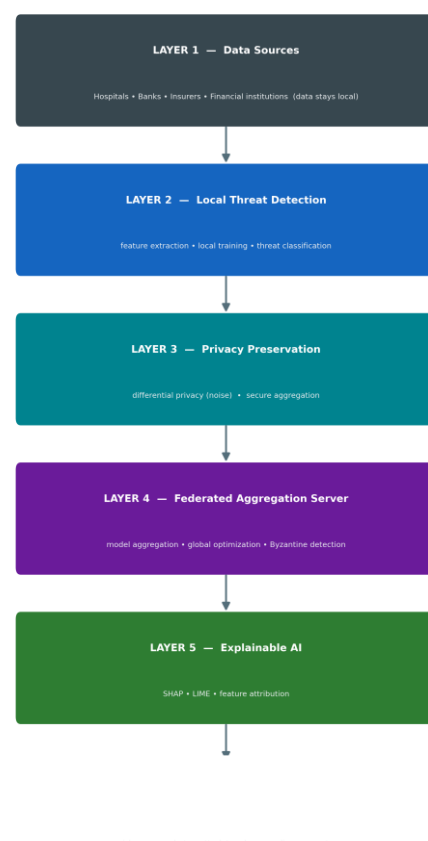


Fig. 1. PPFTIF six-layer architecture. Only privacy-protected model updates leave Layer 3; raw data never leaves the participant.

A. Layer 1: Data sources

Participants—hospitals, banks, insurers, and other financial institutions—retain all local data within their own boundaries. No raw record is ever transmitted.

B. Layer 2: Local threat-detection engine

Each participant performs feature extraction, local model training, and threat classification on its own telemetry, which may include public intrusion datasets such as CICIDS2017, CICIDS2018, and UNSW-NB15 alongside healthcare network logs and financial transaction logs.

C. Layer 3: Privacy-preservation layer

Before any update leaves the organization it is protected in two complementary ways. Differential privacy perturbs each update with calibrated Gaussian noise,

$$w_i' = w_i + N(0, \sigma^2) \quad (3)$$

where the noise scale σ controls the privacy–utility trade-off and yields a provable privacy guarantee. Secure aggregation then ensures that the server observes only the aggregate

$$W = \sum_i w_i' \quad (4)$$

and never any individual participant's update, so that even a curious server cannot single out one organization's contribution.

D. Layer 4: Federated aggregation server

The server performs model aggregation and global optimization and runs Byzantine-attack detection to identify and discard malicious updates before they reach the global model.

E. Layer 5: Explainable AI layer

SHAP, LIME, and feature-attribution methods explain each threat classification, so analysts see why an alert was raised rather than an opaque verdict—directly serving goal G3.

F. Layer 6: Threat-intelligence dashboard

The dashboard surfaces risk scores, threat alerts, attack-trend analytics, and compliance reports, turning the global model's

output into actionable, auditable intelligence for each participant.

VI. MATHEMATICAL MODEL

PPFTIF minimizes the standard federated objective, a sample-weighted sum of local losses:

$$\min F(w) = \sum_i (n_i / n) F_i(w) \quad (5)$$

where $F_i(w)$ is the local loss at participant i . To support cross-sector collaboration (goal G1) without exposing records, the global objective is expressed as a convex combination of a healthcare term F_H and a finance term F_F ,

$$F_{\text{global}}(w) = \alpha F_H(w) + \beta F_F(w), \quad \alpha + \beta = 1 \quad (6)$$

The weights α and β let operators tune how strongly each sector influences the shared model, enabling sector-specific knowledge to be blended without any raw data crossing organizational—or sectoral—boundaries.

VII. PROPOSED ALGORITHM

Algorithm 1 specifies one full PPFTIF training cycle. It interleaves local training with the privacy-preservation steps of Layer 3 and the robustness checks of Layer 4, repeating until the global model converges.

```

Algorithm 1: PPFTIF Federated Training
Input: local threat datasets  $D_1 \dots D_N$ , noise scale  $\sigma$ ,
       sector weights  $\alpha, \beta$ 
Output: global threat-detection model  $W$ 

1. Initialize global model  $W_0$ ;  $t \leftarrow 0$ 
2. repeat
3.   Distribute  $W_t$  to all participants
4.   for each participant  $i$  in parallel:
5.      $w_i \leftarrow \text{Local Train}(W_t, D_i, E \text{ epochs})$  //  $E$  local epochs
6.      $w_i' \leftarrow w_i + \text{Gaussian Noise}(0, \sigma^2)$  // diff. privacy, Eq.(3)
7.      $u_i \leftarrow \text{Encrypt}(w_i')$  // secure aggregation
8.     Send  $u_i$  to server
9.   Verify Update Integrity( $\{u_i\}$ )
10.   $\{w_i'\} \leftarrow \text{Secure Aggregate}(\{u_i\})$  // server sees sum only
11.   $\{w_i'\} \leftarrow \text{Byzantine Filter}(\{w_i'\})$  // median / robust agg
12.   $W_{t+1} \leftarrow \text{Weighted Combine}(\{w_i'\}, \alpha, \beta)$  // Eq.(2), (6)
13.   $t \leftarrow t + 1$ 
14. until convergence
15. return  $W_t$ 

```

VIII. SECURITY ANALYSIS

PPFTIF is designed against the principal attacks on federated systems. Fig. 2 shows the protected exchange that underlies the analysis: updates leave each participant already noised and encrypted, and the server can act only on their aggregate.

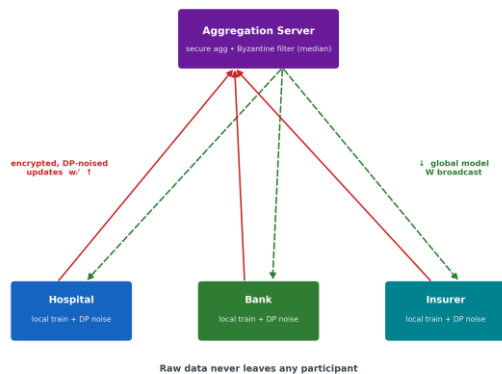


Fig. 2. One PPFTIF training round. Participants send encrypted, differentially-private updates; the server applies secure aggregation and Byzantine filtering and broadcasts the global model.

A. Membership-inference attacks

An adversary may try to determine whether a particular record participated in training. Differential privacy bounds how much any single record can influence an update, and secure

aggregation prevents the server from isolating any one participant's contribution; together they sharply limit membership inference.

B. Gradient-leakage attacks

Attackers may attempt to reconstruct private inputs from shared gradients. PPFTIF counters this with gradient clipping, calibrated noise injection (Eq. 3), and, where stronger guarantees are required, homomorphic encryption of updates, so that the information available to an attacker is too degraded to invert.

C. Model-poisoning attacks

Malicious participants may submit corrupted updates to degrade or backdoor the global model. Rather than simple averaging, PPFTIF aggregates with a robust estimator such as the coordinate-wise median,

$$\text{Agg} = \text{Median}(w_1, w_2, \dots, w_N) \quad (7)$$

which is far less sensitive to a minority of outlying updates and therefore substantially improves resilience to Byzantine participants (goal G2).

IX. WORKED EXAMPLE: CROSS-SECTOR DETECTION

To make the framework concrete, consider an attack campaign that touches two organizations. A hospital observes unusual outbound connections from a medical-device subnet, while a bank sees a small wave of anomalous transactions originating

from the same set of network indicators. Individually, each signal is weak: the hospital's traffic could be a misconfigured device, and the bank's transactions fall just under its alerting threshold.

Under PPFTIF the two organizations contribute to a shared detector without exchanging any record. Each trains locally on its own logs (Layer 2), perturbs and encrypts its update (Layer 3), and sends it to the aggregation server, which combines the healthcare and finance contributions through the sector-weighted objective of Eq. 6 and filters any anomalous updates

X. PERFORMANCE EVALUATION PLAN

Although this paper proposes a design rather than reporting measurements, we specify a reproducible evaluation protocol so that the framework's claims can be tested and compared.

A. Datasets

Cybersecurity detection is evaluated on CICIDS2017, CICIDS2018, and UNSW-NB15; healthcare on medical-IoT datasets and hospital network logs; and finance on the IEEE-CIS Fraud dataset and a public credit-card-fraud dataset. To reflect realistic deployment, data are partitioned across simulated participants in a non-IID fashion so that the framework is tested under the heterogeneity that degrades naive FL.

B. Metrics

Detection quality is measured by accuracy, precision, recall, and F1-score:

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN}) \quad (8)$$

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}), \quad \text{Recall} = \text{TP} / (\text{TP} + \text{FN}) \quad (9)$$

$$\text{F1} = 2\text{PR} / (\text{P} + \text{R}) \quad (10)$$

Privacy is quantified by the privacy budget ϵ , where lower values imply stronger privacy; the protocol sweeps ϵ to chart the privacy–utility trade-off. Robustness is measured by the degradation in F1 under an increasing fraction of Byzantine participants, and efficiency by communication cost and per-round latency.

C. Baselines and ablations

PPFTIF should be compared against a centralized model (an upper bound on accuracy, unattainable under regulation), plain FedAvg without privacy, and FedAvg with differential privacy but without robust aggregation. Ablating differential privacy, secure aggregation, Byzantine filtering, and the XAI layer in turn isolates each component's contribution to privacy, robustness, and interpretability.

XI. EXPECTED OUTCOMES

The framework is expected to preserve sensitive healthcare and financial data, improve attack-detection rates relative to isolated per-organization models, enable cross-organization collaboration that is otherwise legally impossible, reduce false positives through richer shared signal, improve regulatory compliance by construction, and provide explainable threat intelligence. These are hypotheses that the evaluation plan above is designed to test; experimental validation remains future work, and no quantitative claim is made here.

XII. LIMITATIONS

Several limitations temper the design. Differential privacy improves confidentiality at a measurable cost in accuracy, and the right operating point on this trade-off is deployment-specific. Secure aggregation and encryption add communication and latency overhead that may strain large federations. Non-IID data across hospitals and banks can slow or destabilize convergence, and the sector weights α and β require tuning. Robust

with the robust aggregator of Eq. 7. The global model learns the joint pattern—the correlation between the device-subnet indicators and the transaction anomalies—that neither organization could see alone. When redistributed, it raises a high-confidence alert at both sites, and the Explainable AI layer attributes the alert to the shared indicators, giving analysts at each institution an actionable, auditable explanation. No protected health information and no cardholder data ever left their respective custodians, so the collaboration remains compliant with HIPAA and PCI-DSS simultaneously (goal G4). aggregation resists a minority of Byzantine participants but can fail if adversaries form a majority or adapt their updates to evade the median. Finally, regulatory interoperability is more than a technical property: even a privacy-preserving exchange of model updates may face differing legal interpretations across jurisdictions, which the framework can support but not by itself resolve.

XIII. FUTURE RESEARCH DIRECTIONS

Promising directions include federated large language models for threat intelligence, blockchain-assisted federated learning for tamper-evident update provenance, quantum-safe federated security, mechanisms for cross-border regulatory compliance, federated security-operations-center (SOC) architectures that coordinate detection across organizations, and multi-agent federated cyber-defense systems in which participants' models cross-validate one another. Replacing heuristic Byzantine filters with learned, calibrated detectors trained on the proposed benchmark is a further priority, as is empirical validation of the full framework using the methodology of Section X.

XIV. CONCLUSION

This paper presented a comprehensive survey and the Privacy-Preserving Federated Threat Intelligence Framework (PPFTIF) for collaborative attack detection across healthcare and financial sectors. By combining federated learning, differential privacy, secure aggregation, Byzantine-resilient learning, and explainable AI, PPFTIF lets organizations jointly strengthen their cyber defenses without sharing sensitive data. We formalized a sector-weighted federated objective, specified the training algorithm with its privacy and robustness steps, analyzed the framework against the principal federated attacks, and illustrated cross-sector detection through a worked example. The architecture addresses privacy leakage, regulatory compliance, poisoning attacks, and interpretability together, making it a promising basis for next-generation cyber-threat-intelligence platforms. Empirical validation using the proposed datasets, metrics, and privacy-budget protocol is the necessary next step before any quantitative claim is made.

REFERENCES

- [1] M. Ali et al., “Federated learning for privacy preservation in smart healthcare systems: a comprehensive survey,” 2022.
- [2] A. Vyas et al., “Privacy-preserving federated learning for intrusion detection in IoT environments: a survey,” *IEEE Access*, 2024.
- [3] H. Zhang et al., “Survey of federated learning in intrusion detection,” *J. Parallel Distrib. Comput.*, 2025.
- [4] S. Saha et al., “A multifaceted survey on privacy preservation of federated learning,” *Artif. Intell. Rev.*, 2024.
- [5] H. Li et al., “Survey: federated learning data security and privacy preserving in edge-IoT,” *Artif. Intell. Rev.*, 2024.
- [6] “Systematic review of privacy-preserving federated learning in decentralized healthcare systems,” 2025.
- [7] H. U. Manzoor et al., “A survey of security strategies in federated learning,” *Future Internet*, 2024.
- [8] C. Zhang et al., “Anomaly detection and defense techniques in federated learning,” *Artif. Intell. Rev.*, 2024.

- [9] D. Shenoy et al., “Exploring privacy mechanisms and metrics in federated learning,” *Artif. Intell. Rev.*, 2025.
- [10] D. Dinkov, “Homomorphic encryption for secure federated learning: a narrative survey,” 2025.
- [11] “A review of research on secure aggregation for federated learning,” 2025.
- [12] Y. Bi et al., “Enabling privacy-preserving cyber threat detection with federated learning,” 2024.
- [13] B. McMahan et al., “Communication-efficient learning of deep networks from decentralized data (FedAvg),” in *Proc. AISTATS*, 2017.
- [14] K. Bonawitz et al., “Practical secure aggregation for privacy-preserving machine learning,” in *Proc. ACM CCS*, 2017.
- [15] C. Dwork and A. Roth, “The algorithmic foundations of differential privacy,” *Found. Trends Theor. Comput. Sci.*, 2014.
- [16] P. Blanchard et al., “Machine learning with adversaries: Byzantine-tolerant gradient descent (Krum),” in *Proc. NeurIPS*, 2017.