





Article

Federated Learning for Cybersecurity: A Privacy-Preserving Approach

Edi Marian Timofte ^{*}, Mihai Dimian , Adrian Graur, Alin Dan Potorac, Doru Balan, Ionut Croitoru ,
Daniel-Florin Hrițcan  and Marcel Pușcașu 

Department of Computers, Automation and Electronics, University “Ștefan cel Mare”, 720229 Suceava, Romania; dimian@usm.ro (M.D.); adriang@usm.ro (A.G.); alin@usm.ro (A.D.P.); dorub@usm.ro (D.B.); ionut.croitoru@usm.ro (I.C.); daniel.hritcan@usm.ro (D.-F.H.); marcel.puscasu@student.usv.ro (M.P.)

* Correspondence: edi.timofte@usm.ro; Tel.: +40-748-171-798

Abstract: The growing number of cyber threats and the implementation of stringent privacy regulations have revealed significant shortcomings in traditional centralized machine learning models, especially in distributed systems like the Internet of Things (IoT). This study presents a Federated Learning (FL) framework designed for intrusion detection and malware classification. This framework enables decentralized model training while preserving data locality and minimizing communication overhead. The proposed architecture incorporates lightweight, privacy-preserving techniques, including gradient clipping, differential privacy, and encrypted model aggregation, to ensure secure and efficient collaboration across heterogeneous clients. Experimental results on two widely adopted cybersecurity benchmarks demonstrate that the framework achieves detection accuracies above 90%, maintains privacy loss below 5%, and improves communication efficiency by over 25%. These results confirm the viability of FL as a scalable, privacy-compliant approach for next-generation cybersecurity systems in highly distributed infrastructures.

Keywords: federated learning; cybersecurity; intrusion detection; privacy preservation; IoT security; machine learning; malware detection; cyber resilience



Academic Editors: Georgi R. Tsochev,
Maria Nenova, Peican Zhu and
Ki-Hyun Jung

Received: 19 May 2025

Revised: 13 June 2025

Accepted: 15 June 2025

Published: 18 June 2025

Citation: Timofte, E.M.; Dimian, M.; Graur, A.; Potorac, A.D.; Balan, D.; Croitoru, I.; Hrițcan, D.-F.; Pușcașu, M. Federated Learning for Cybersecurity: A Privacy-Preserving Approach. *Appl. Sci.* **2025**, *15*, 6878. <https://doi.org/10.3390/app15126878>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

FL is a machine learning paradigm in which multiple clients collaborate to train a shared model without exchanging raw data. This approach preserves data privacy and reduces communication costs by transmitting only model updates. Due to its applicability in privacy-sensitive domains such as healthcare, finance, and cybersecurity, FL has gained significant attention [1–3].

The increasing complexity and frequency of cyber threats in distributed computing environments has necessitated the advancement of machine learning-based defense systems that are both intelligent and privacy-preserving. Traditional centralized learning models, while powerful, are inherently susceptible to critical problems such as data leakage, excessive communication direction, and single points of failure. FL emerges as a robust alternative, allowing the training of a collaborative model across decentralized devices without the transfer of initial data to the central server. This decentralized paradigm preserves user privacy, reduces systemic vulnerability, and ensures compliance with regulatory constraints.

FL has demonstrated significant potential in constrained and dynamic environments, including underwater drone networks, where high latency and limited bandwidth challenge centralized learning. For example, Popli et al. [4] proposed a federated framework tailored for underwater drones that improved zero-day threat detection while maintain-

ing strict data locally. These applications highlight the ability of FL to support localized learning while maintaining high detection accuracy under extreme conditions.

A key concern in FL research is balancing efficiency with privacy. Recent techniques using gradient clipping and Fisher information-based parameter selection have proven effective in reducing communication overhead without sacrificing accuracy [5]. These lightweight mechanisms optimize model performance while minimizing resource consumption and data exposure, making FL feasible for deployment in bandwidth-constrained networks.

To address the inherent challenges posed by non-independent and identically distributed (non-IID) data across clients, several frameworks have introduced multi-objective and multi-tasking FL strategies. These include client clustering, personalized updates, and fairness-aware optimization, which collectively improve generalization and fair performance across different data distributions [6].

On a broader scale, FL has gained traction in distributed cloud computing architectures, where it complements SMPC, trusted execution environments, and differential privacy. Rahdari et al. [7] highlighted the role of FL in enhancing privacy-aware data analytics and mitigating the risks associated with centralized storage in cloud-native infrastructures.

As FL systems become more personalized, they face new threats such as model poisoning and stealthy backdoor attacks. Defense strategies such as adaptive layered trust aggregation and anomaly detection based on gradient similarity offer promising solutions that increase robustness against adversarial manipulation [8].

In rapidly evolving, containerized, and cloud-native ecosystems, flexible protection architectures are essential. AI-driven adaptive security networks have been proposed to support real-time anomaly detection in federated cloud environments [9], in line with the decentralized nature of FL. Such systems dynamically adapt to evolving attack surfaces, improving responsiveness and resilience.

Security in FL is further enhanced by advances in cryptographic techniques. For example, delegable order-revealing encryption (DORE) enables secure multi-user range queries without relying on trusted intermediaries, preserving confidentiality while maintaining operational efficiency [10]. In addition, the integration of blockchain with FL brings transparency, immutability, and trust to model update workflows. Blockchain-enabled FL architectures ensure auditable, tamper-proof exchanges between participants, protecting against adversarial interference and dishonest contributions [11].

In the large-scale deployment of private data protection—for example, smart city, networked medical systems, and industrial IoT—the use of FL frameworks has shown promising results. Kotian et al. [12] emphasized the importance of combining FL with light encryption, anomaly detection, and adaptive privacy mechanisms to meet compliance standards (e.g., GDPR, HIPAA) [13,14] without compromising efficiency. In summary, this development shows that FL is a mature enabling technology for secure, scalable, and privacy-aware cybersecurity solutions. Continued innovation in the penetration of artificial intelligence, cryptography, and distributed architecture remains necessary to overcome emerging threats and deploy resilient defenses in real-world infrastructure.

The goal of this paper is to design, implement, and evaluate a modular FL framework that improves the cybersecurity of distributed IoT systems. This framework integrates lightweight, privacy-preserving mechanisms; secure communication protocols; and adaptive local learning. The proposed approach overcomes the limitations of existing methods by unifying performance, scalability, and data confidentiality into a coherent architecture.

The remainder of this paper is structured as follows: Section 2 reviews relevant literature on FL, emphasizing its integration into cybersecurity and identifying current gaps. Section 3 introduces the proposed modular architecture, privacy-preserving techniques, and rationale behind the selected methods. Section 4 presents the experimental design, datasets,

evaluation metrics, and a comparative analysis with baseline models. Section 5 discusses the deployment architecture and response mechanisms in practical IoT environments. Finally, Section 6 concludes the paper with a summary of the findings and potential directions for future work.

In summary, this development demonstrates that FL is a mature technology that can enable secure, scalable, and privacy-aware cybersecurity solutions. Continued innovation in artificial intelligence, cryptography, and distributed architecture is necessary to overcome emerging threats and deploy resilient defenses in real-world infrastructure.

As quantum technologies evolve, they present new challenges to the confidentiality and resilience of FL systems. In anticipation of adversaries with quantum capabilities, recent research has explored integrating quantum-resilient approaches into FL architectures. These include lattice-based cryptographic primitives, quantum key distribution protocols and secure aggregation schemes that are designed to resist quantum decryption attempts. These adaptations are intended to ensure the long-term security of FL deployments, particularly in critical infrastructures such as healthcare and smart city networks [15,16].

2. Related Work

The development of cyber safety threats and increasing demand for focused privacy solutions have significantly expanded research into FL applications. While traditional centralized machine learning methods remain strong, they show critical restrictions such as data leakage, narrow communication spots and exposure to individual points of failure. These disadvantages are particularly important in IoT ecosystems, where a huge number of connected devices work over sensitive data protection.

The challenge of learning from non-IID data across heterogeneous clients has led to the development of multi-objective and multi-task FL strategies, with studies showing improved accuracy and fairness in real-world scenarios [6]. In parallel, efforts to integrate FL within distributed cloud computing infrastructures have leveraged SMPC, trusted execution environments, and differential privacy to provide robust and privacy-preserving analytics across nodes [7].

Personalization in FL has emerged as a key area, enabling the adaptation of models to individual clients while defending against advanced threats. Defense mechanisms based on gradient similarity and layered trust policies have shown improved robustness in countering stealthy backdoor attacks, without compromising collaborative learning [8]. In dynamic and containerized environments, AI-powered adaptive security meshes have been proposed as complementary to FL, improving threat detection and resilience [9].

Further developments in secure computation include efficient delegable order-revealing encryption schemes that support multi-user range queries over encrypted data—an essential capability in federated analytics frameworks [10]. To enhance trust and auditability, blockchain-integrated FL has also gained attention, enabling immutable logs and secure collaboration in decentralized learning systems [11]. Within smart cities, FL frameworks augmented with lightweight encryption and privacy-preserving mechanisms have been proposed to secure large-scale IoT infrastructures [12].

Expanding into vehicular networks, research has introduced certificateless signature schemes with batch verification for secure vehicle-to-vehicle communication, reducing computational overhead while ensuring privacy [17]. Risk modeling techniques, such as the Cyber Intelligent Risk Assessment (CIRA) [18] methodology, have combined machine learning with FL to estimate cyber risks in industrial IoT environments.

Authentication mechanisms have also evolved through the application of FL, utilizing alternative biometric data such as energy consumption patterns for IoT device identification, thus reducing dependence on explicit user credentials [19]. Federated archi-

texture has additionally been applied in secure ride-matching systems, enabling real-time privacy-preserving matching over road networks [20]. Homomorphic data encapsulation techniques for secure vehicular positioning have also been developed to maintain location privacy in smart transportation [21], while scalable cross-domain anonymous authentication mechanisms have supported robust FL deployment in IoT settings [22].

Other novel directions have included secure cross-modal search over encrypted datasets [23], Dilithium-based encryption integration for federated security [24], and privacy-preserving image retrieval systems tailored for FL applications [25]. Techniques for exposing IoT platforms securely behind Carrier-Grade NATs [26] and implementing fine-grained access control in cloud-assisted vehicular networks [27] have further reinforced FL's role in protecting distributed infrastructure.

More recent advancements include client-sampled federated meta-learning strategies that personalize intrusion detection models across IoT devices [13], as well as hybrid transfer and self-supervised learning models aimed at improving network security in vehicular environments [28]. Research on edge-level defenses has also contributed to this domain by leveraging open-source router firmware (e.g., DD-WRT) to enhance perimeter security in distributed networks [29].

Finally, the introduction of intelligent federated frameworks such as Trust-6GCPSS [30] for secure interaction within 6G cyber–physical–social systems has expanded the horizon of FL research [30]. The body of work continues to evolve with contributions addressing intrusion detection [31], vehicular privacy [32], collaborative defense architectures [3], and trustworthy edge computing [30,33].

Collectively, these studies confirm that FL—when enhanced through lightweight cryptography, blockchain, personalized defense mechanisms, and robust encryption—offers a resilient and scalable solution for building next-generation cybersecurity frameworks in heterogeneous, privacy-sensitive, and distributed environments.

Main Contributions of This Work

Previous studies have introduced FL frameworks, such as FedAvg [17], FedProx [18], and MOFL/MTFL [19], which have laid the groundwork for scalable collaborative learning. However, these approaches typically address privacy, efficiency, and security separately, resulting in gaps in resilience and adaptability in hostile environments.

In this work, we propose a comprehensive, modular FL framework engineered to enhance cybersecurity in distributed, resource-constrained IoT environments. The framework combines a variety of privacy-preserving and performance-oriented techniques, including differential privacy, secure multi-party aggregation, gradient clipping, Fisher-based parameter pruning, and post-quantum encryption (Dilithium), into a unified, scalable system that can withstand adversarial interference. The components coexist and are orchestrated to ensure low-latency training, model convergence under non-IID data conditions, and robust protection of sensitive information at the data and model levels.

Unlike conventional approaches, which tend to treat privacy, efficiency, and security as separate concerns, our proposed architecture takes a holistic design approach, embedding cryptographic protocols, model optimization heuristics, and auditability mechanisms directly into the operational fabric of the FL workflow. Blockchain-assisted logging mechanisms ensure traceability and tamper-proof recordkeeping among federated nodes. Meanwhile, adaptive edge–fog–cloud orchestration and personalized local updates significantly reduce communication overhead and model divergence. This layered integration enhances the system's ability to operate under dynamic threat landscapes and positions it as a forward-compatible solution capable of accommodating future enhancements, such as threat intelligence sharing and quantum-resilient cryptographic primitives.

These contributions collectively define a comprehensive and scalable approach to privacy-preserving intrusion detection in federated IoT environments and provide the foundation for the methodology proposed in the next section.

Table 1 provides a comparative overview of our proposed SecFL-IoT framework and well-established FL methods, such as FedAvg, FedProx, and MOFL/MTFL.

Table 1. Mapping FL challenges to applied techniques and their impact.

Challenge	Technique(s) Applied	Justification and Impact
Privacy leakage	Differential Privacy, Secure Aggregation	Prevents data reconstruction by third parties; protects data during aggregation.
Communication overhead	Gradient Clipping, Parameter Pruning	Reduces transmission volume; enables faster convergence and lower bandwidth usage.
Model poisoning	Blockchain Logging, Client Reputation Score	Ensures auditability; tracks anomalies in contributions to deter malicious updates.
Gradient explosion/instability	Gradient Clipping	Ensures stable convergence by bounding extreme gradient values.
Complexity for IoT deployment	Fisher-Based Parameter Pruning	Shrinks model size while preserving accuracy; ideal for resource-limited edge devices.
Non-IID data handling	Personalized Local Updates, FedAvg Variant	Enables better generalization across heterogeneous device data.
Accountability and trust deficit	Blockchain Audit Trail	Provides transparent and immutable logs of client actions and updates.

The comparison centers on critical challenges in FL, including privacy leakage, communication overhead, and model poisoning. It explains the rationale behind the specific techniques used by each method and outlines the expected impact. The analysis shows how our framework improves security and privacy while enhancing scalability in distributed, heterogeneous environments.

It is important to note that this comparison is conceptual and is based on the architectural designs and techniques described in the original works. The methods were not evaluated under identical experimental conditions; rather, they were analyzed based on their documented capabilities and design goals.

As can be seen in Table 1, the proposed SecFL-IoT framework tackles key challenges in FL by combining well-established techniques from recent literature with customized mechanisms for IoT deployment. Differential privacy and secure aggregation mitigate privacy leakage by ensuring that raw data remains inaccessible to both the server and peers, as demonstrated in [1]. Communication overhead is reduced through parameter pruning and gradient clipping, which have been validated in edge AI contexts [2,3].

Blockchain-based logging and client reputation scoring enforce accountability and robustness against model poisoning by using tamper-resistant ledgers and verifiable updates [4]. The inclusion of Fisher-based pruning supports model compression without compromising accuracy, which is essential for resource-constrained devices [5]. To handle data heterogeneity, personalized local updates and FedAvg variants facilitate generalization across non-IID datasets [6].

Furthermore, secure communication is ensured via a private permissioned blockchain, where each client update is hashed, digitally signed and validated by smart contracts, preventing rollback or spoofing. This decentralized validation mechanism removes the need for centralized trust anchors. Integrating adaptive fog–edge–cloud orchestration improves convergence and efficiency further while minimizing latency and bandwidth usage.

3. Proposed Methodology

To address the challenges of data privacy, communication overhead, and model robustness in distributed cybersecurity systems, we propose an FL framework enhanced with lightweight privacy-preserving techniques. The IoT environment is the primary focus

of our approach for which intrusion detection and malware classification are performed. Here, due to the overwhelming data sensitivity and network heterogeneity, there are significant challenges that need to be exercised.

Several FL architectures have been proposed in recent literature to address similar issues. These include FedAvg [17], FedProx [18], and MOFL/MTFL [19]. However, these frameworks typically use general-purpose aggregation schemes that offer limited privacy enhancements and minimal support for auditability or communication optimization. For example, FedAvg emphasizes simplicity and scalability; FedProx introduces regularization for non-IID data; and MOFL/MTFL integrates task-specific personalization. In contrast, our framework integrates a broader set of mechanisms, including blockchain-assisted audit logging and post-quantum encryption, making it more suitable for adversarial, privacy-sensitive IoT environments.

To ensure robust aggregation in adversarial settings, the framework uses a combination of secure multi-party computation (SMPC) and gradient anomaly detection based on cosine similarity and update norm thresholds. Before accepting model updates from clients, the server evaluates each client's contribution to detect potential poisoning or deviation from the expected learning trajectory. Aggregation is then performed only on trusted updates, which are weighted by client reputation scores and consistency metrics. This strategy significantly mitigates the risk of model corruption while preserving convergence stability in non-IID scenarios [20,21].

In FL environments, ensuring privacy requires architectural strategies that prevent the exposure of raw data and maintain trust among distributed clients. The proposed system meets these requirements by integrating cryptographic communication, selective model update mechanisms, and differential privacy within a secure aggregation framework. Each client independently trains a local model using its private dataset. Then, all updates are processed and combined in a secure and verifiable manner by a central server.

Figure 1 shows the system's overall architecture and briefly describes the major components and their process of interaction with clients and the central aggregation server.

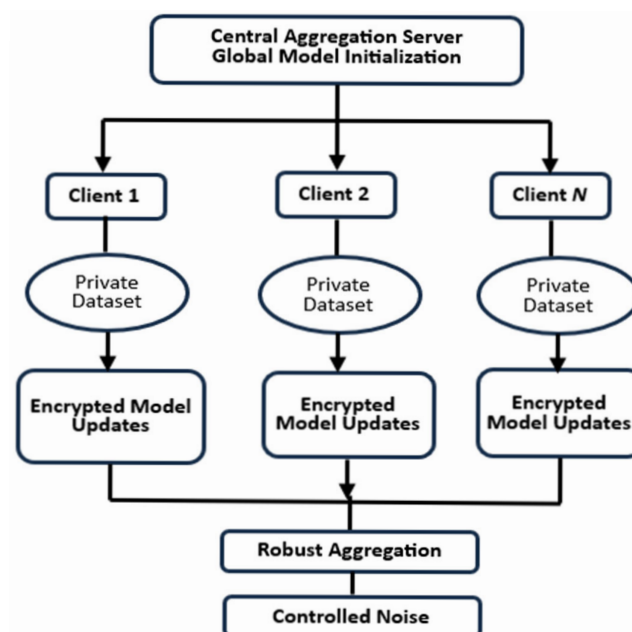


Figure 1. Privacy-preserving FL framework for intrusion detection and malware classification.

During each training round, the central server initializes the global model and distributes it to a selected set of participating clients. Each client performs local training on its

private dataset, using gradient clipping to limit the update size and Fisher-based parameter pruning to reduce dimensionality. These techniques aim to limit potential leakage from gradient inference and minimize communication overhead.

After completing local training, clients encrypt model updates using a Diffie–Hellman key exchange mechanism and send the encrypted parameters to the aggregation server. Blockchain logging ensures that all updates are auditable and tamper-proof. The server performs secure multiparty aggregation to combine updates without reconstructing private data, followed by the injection of calibrated differential privacy noise to protect individual contributions before updating the global model.

The system supports secure decentralized model training while implementing privacy-preserving mechanisms at both the client and server sides.

Initially, a central server initializes a global model and distributes it to participating clients. Each client conducts local training on its private data without transmitting raw samples. To enhance security, local models employ gradient clipping and selective parameter sharing based on the Fisher information matrix [5], effectively reducing potential information leakage and communication overhead.

During the local update phase, clients encrypt their model parameters using a secure Diffie–Hellman key exchange protocol [10], preventing interception attacks. This encrypted communication layer is supported by a blockchain infrastructure [11] to provide auditable, tamper-proof recording of model updates, increasing trust among participants.

Upon receiving updates, the central server performs a robust aggregation using secure multi-party computation techniques [7], ensuring that no single party can reconstruct sensitive data from model parameters. Additionally, controlled differential privacy noise is injected into the aggregated model to further preserve client confidentiality [9].

To adapt to heterogeneous and dynamic IoT environments, the framework includes personalized FL mechanisms [8], enabling the system to fine-tune models according to client-specific data patterns while maintaining overall network security and stability. Furthermore, smart load balancer mechanisms are incorporated to manage communication efficiency and client sampling strategies [34].

The privacy-preserving structure also integrates secure edge–fog–cloud architecture models [1], utilizing techniques like encrypted federated analytics [23] and decentralized IoT gateway security [35]. As shown in Figure 2, an additional privacy control layer is used to protect client identities and model updates.

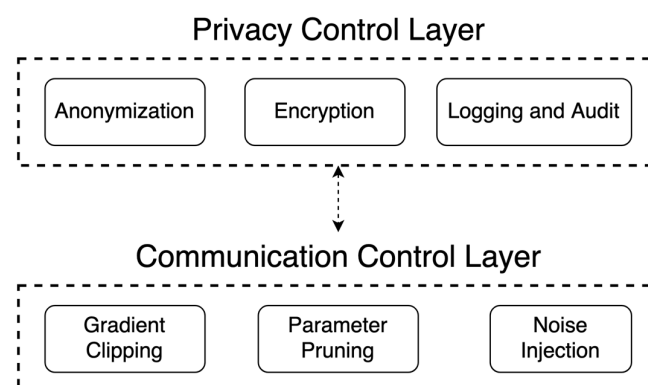


Figure 2. Privacy and communication control layers in the proposed FL systems.

To strengthen the cybersecurity defenses, techniques such as fine-grained access control for federated IoT data sharing [27], blockchain interoperability frameworks, and privacy-preserving vehicular positioning are implemented. Robust threat intelligence sharing over federated grids and post-quantum cryptographic methods like Dilithium

are also embedded within the architecture to future-proof the system against emerging threats [36,37].

Compared to traditional centralized machine learning solutions, the proposed methodology ensures high scalability, improved privacy guarantees, low communication costs, and resilience against adversarial participants. Its modular and adaptive design makes it practical for deployment in real-world settings like smart cities, autonomous vehicular networks, healthcare IoT environments, and cloud–edge infrastructures.

3.1. Security and Privacy Analysis

The proposed framework integrates several security and privacy-preserving techniques designed to defend against common threats in FL, including eavesdropping, model inversion, poisoning, and malicious aggregation.

First, the use of Diffie–Hellman key exchange ensures that communication between clients and the server is protected against passive attackers who could intercept updates. Second, model update encryption combined with Fisher-based pruning and gradient clipping minimizes both the amount and sensitivity of shared information, thereby reducing vulnerability to reconstruction and gradient leakage attacks.

Differential privacy mechanisms applied during global model aggregation provide formal guarantees that individual client contributions remain statistically indistinguishable. In addition, blockchain-based update logging ensures immutability and auditability, preventing tampering and replay attacks. The integration of post-quantum encryption (Dilithium) further strengthens the system against future cryptographic threats. By integrating these measures, the system delivers a resilient, privacy-focused learning framework specifically tailored for adversarial, heterogeneous IoT environments.

3.2. Security Layers

The proposed framework integrates a multi-layered set of security and privacy-preserving techniques that are designed to counteract common threats in FL. These threats include eavesdropping, model inversion, poisoning, and aggregation manipulation attacks:

- Communication confidentiality is ensured through Diffie–Hellman key exchange, which secures client–server interactions against passive interception [25];
- Gradient clipping and Fisher-based parameter pruning reduce the volume and sensitivity of transmitted updates, mitigating the risk of gradient leakage and model reconstruction by adversaries [13];
- Differential privacy provides formal guarantees that individual contributions cannot be reverse-engineered from the global model [33];
- SMPC mechanisms enable encrypted aggregation without exposing individual model updates to the server [1];
- A permissioned blockchain ledger logs each client contribution and model update using cryptographic hashes and digital signatures. This mechanism provides tamper-evident auditability and defends against rollback and replay attacks [14];
- The framework incorporates post-quantum encryption (Dilithium) to safeguard the system against quantum-capable adversaries and align with emerging cryptographic standards [1];
- From a regulatory perspective, the architecture is compatible with GDPR and HIPAA principles by minimizing data exposure and supporting verifiable processing trails [13,14].

Collectively, these mechanisms enforce data minimization, integrity, accountability, and resilience—key requirements for deploying federated intrusion detection systems in real-world, adversarial IoT environments.

4. Experimental Setup and Evaluation

To validate the effectiveness, security, and scalability of the proposed privacy-preserving FL framework, a comprehensive experimental setup was established. This section presents the environment configuration, datasets employed, evaluation metrics considered, and the sequential phases followed during experimentation. Moreover, expected outcomes and a practical case study application are discussed to highlight the real-world applicability and advantages of the proposed architecture.

The experimental design replicates realistic IoT cybersecurity scenarios and integrates advanced privacy-preserving mechanisms, ensuring that the results provide a thorough assessment of the system's performance under both normal and adversarial conditions.

4.1. Experimental Environment

The testbed is designed to simulate a distributed IoT network consisting of multiple edge nodes participating in an FL process coordinated by a central aggregation server. The test environment includes a mix of real and virtualized nodes to enable performance evaluation under conditions that mimic real-world constraints such as limited bandwidth, processing power, and asynchronous client participation.

The central server was deployed on a Dell PowerEdge R740 physical server with an Intel Xeon Silver 4210 @ 2.20 GHz (Intel, Santa Clara, CA, USA), 128 GB RAM, and Ubuntu 22.04 LTS with Docker and Python 3.10. Each federated client was emulated using Docker containers for heterogeneity simulation running on the cluster of Raspberry Pi 4 (4 GB RAM) and multiple virtual machines hosted on a Proxmox hypervisor.

The simulation framework was developed using Flower (FLwr), version 1.6.0, for FL orchestration and PyTorch, version 2.2.2, for local model training. Transport Layer Security (TLS), a widely adopted cryptographic protocol, was used to establish secure communication between nodes over a private network. Prometheus, version 2.51.2, and Grafana, version 10.4.1, were used to monitor the entire testbed for performance tracking and system-level logging.

To validate the proposed privacy-preserving FL framework, a comprehensive experimental environment was designed, replicating a distributed IoT cybersecurity scenario. The configuration includes three main layers: edge, fog, and cloud layers.

At the edge layer, multiple IoT devices and embedded systems (e.g., sensors, cameras, healthcare monitors) serve as clients, each with their own private dataset. These clients are connected through a secure VPN network (Tailscale, Toronto, ON, USA), ensuring encrypted communication between participants.

The fog layer is composed of intermediate edge servers equipped with pfSense firewalls and load balancers, tasked with preprocessing, encrypting, and routing the model updates securely toward the cloud aggregation server. This layer also manages dynamic client sampling to optimize communication overhead.

At the cloud aggregation layer, a centralized server aggregates the encrypted model updates, applies privacy-preserving techniques like SMPC, and updates the global model before distributing it back to the clients.

Figure 3 illustrates the architecture of the testbed, showing the interplay between the IoT edge clients, the fog aggregation layer, and the cloud-based central server. The setup allows for dynamic client selection, failure injection, and bandwidth throttling to replicate realistic FL conditions in hostile or constrained networks.

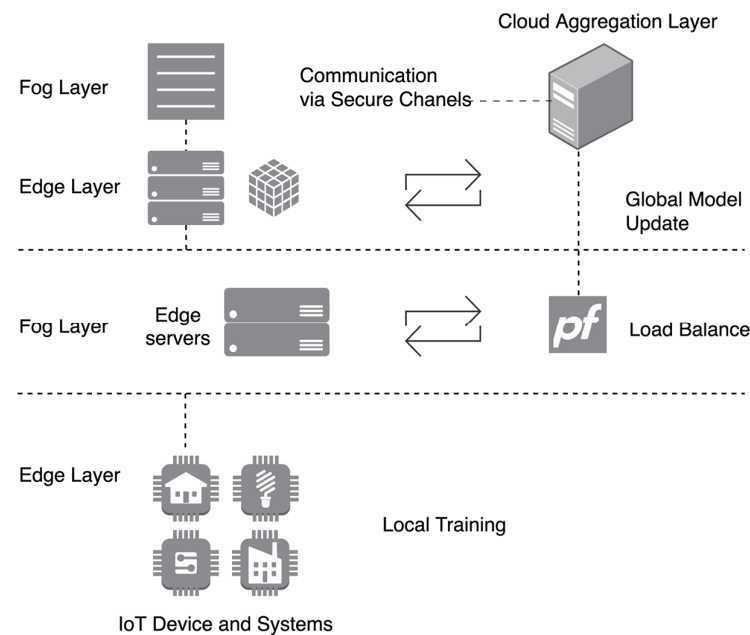


Figure 3. Experimental testbed architecture.

The component labeled ‘pf’ in Figure 3 refers to pfSense, which is an open-source firewall and load balancer that is used to manage network traffic and distribute training workloads dynamically across edge servers [1,2]. In our setup, instances of pfSense are configured to monitor client availability and system load, redirecting model update traffic to the optimal fog or cloud aggregation node as required.

Load balancing is handled at the fog layer through DNS-based routing and real-time monitoring of client connection states. This ensures efficient resource utilization and reduces bottlenecks during model aggregation cycles, particularly in scenarios with intermittent edge connectivity [2].

The entire testbed was simulated using Docker containers to represent distributed clients and servers that were interconnected through a virtualized VPN backbone using Tailscale. Additionally, we deployed attack simulation tools to test intrusion detection capabilities, including standard cyberattack patterns such as Denial of Service (DoS), spoofing, and infiltration.

This experimental setup reflects the best practices used in recent FL testbeds for cybersecurity and privacy-preserving analytics [35,37]. Similar approaches based on containerized emulation, VPN-based secure networking, and multi-layered orchestration have proven effective in replicating IoT and edge–cloud environments under realistic conditions [38,39].

4.2. Datasets Used

Several publicly available and widely recognized datasets were used to evaluate the performance of the proposed FL framework in cybersecurity applications. These datasets were selected to represent realistic and diverse network traffic patterns, including both benign behavior and different types of cyber-attacks.

The primary datasets integrated into the experimental setup are as follows:

- **CICIDS2017:** This dataset, developed by the Canadian Institute for Cybersecurity, includes network traffic for benign activity and multiple attack types, such as DoS, DDoS, port scans, and web-based intrusions. The dataset was generated in a real enterprise environment using realistic scenarios and contains over 3 million labeled flows across more than 80 extracted features [14].

- **TON_IoT:** This dataset, provided by the Cyber Range Lab of UNSW, includes telemetry data, network traffic, and system logs collected from various IoT and IIoT devices within smart environments. The dataset offers multi-source data for evaluating AI-based intrusion detection in heterogeneous IoT systems [38].
- **NSL-KDD:** NSL-KDD is a refined version of the original KDD'99 dataset that removes redundant records and balances class distribution. It remains a widely used benchmark for testing intrusion detection algorithms and includes four primary attack classes: DoS, Reverse to Local (R2L), User to Router (U2R), and Probe [39].

To simulate a non-IID data distribution among clients, each dataset was divided and randomly assigned to edge nodes, favoring specific attack types. This approach replicates the heterogeneity typical of real-world IoT environments, in which clients face different threat profiles.

Each node executed a local preprocessing pipeline comprising the following steps:

- **Missing value imputation:** Null or missing values in numerical features were replaced using feature-wise mean imputation to ensure data completeness while minimizing distributional distortion [1].
- **Feature scaling (MinMax normalization):** All numerical features were rescaled to the [0, 1] interval using MinMax normalization to ensure uniform feature influence during model training and accelerate gradient convergence [2].
- **One-hot encoding of categorical features:** Categorical variables (e.g., protocol type or service) were transformed into binary vectors via one-hot encoding to enable compatibility with neural network models while preserving non-ordinal relationships.
- **Attack label remapping for class balance:** The original attack labels from the datasets (e.g., DoS, Probe, R2L and U2R) were remapped into broader categories to reduce class imbalance and enhance the classifier's ability to generalize. This relabeling aligns with standardized taxonomy schemes used in intrusion detection literature [3].

Preprocessing was performed locally on each node and included data cleaning, feature extraction, normalization, and one-hot encoding of categorical features. The data preprocessing pipeline is illustrated in Figure 4, which shows the transition from raw data ingestion to local model training augmented with privacy-preserving mechanisms.

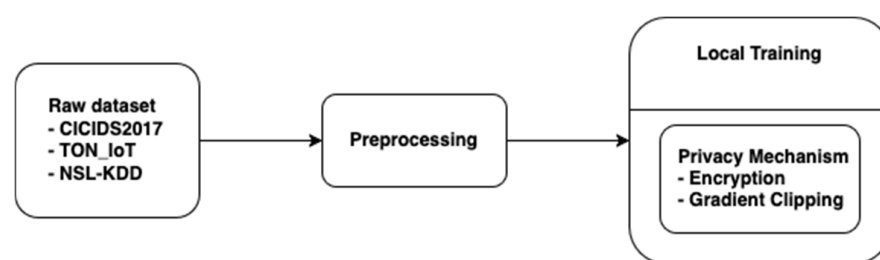


Figure 4. Dataset integration and preprocessing workflow.

Figure 4 summarizes the pipeline adopted for dataset integration and preprocessing prior to federated training. First, the raw datasets—CICIDS2017, TON_IoT, and NSL-KDD—were parsed and cleaned to remove redundant or corrupted entries. Then, each dataset underwent a standardized preprocessing routine that included imputing missing values using statistical heuristics; normalizing via MinMax scaling to ensure uniform input ranges across features; and one-hot encoding for categorical attributes. This ensured compatibility with deep learning models.

This preprocessing step was conducted locally on each client node to preserve data locality and prevent raw data leakage. Then, the processed datasets were fed into the local training modules. Here, privacy mechanisms, such as lightweight, AES-based symmetric

encryption, were applied to the model updates. Meanwhile, gradient clipping ensured numerical stability and reduced the risk of model inversion attacks. These choices were motivated by prior studies demonstrating the efficiency of such techniques in non-IID, bandwidth-constrained environments [5,23].

Each component in Figure 4 corresponds to an essential stage in transforming raw IoT data into secure, federated model updates. The design reflects the practical end-to-end integration of privacy-preserving mechanisms into the training workflow, optimized for adversarial and resource-constrained deployment contexts.

4.3. Evaluation Metrics

The evaluation of the proposed privacy-preserving FL framework was performed using a set of standard cybersecurity and machine learning metrics. These metrics were selected to comprehensively assess model performance, communication efficiency, and privacy preservation across distributed nodes in a simulated IoT environment.

These metrics were chosen not only for their prevalence in intrusion detection tasks, but also for their ability to reflect trade-offs in federated settings, where privacy constraints, data heterogeneity, and communication costs directly affect model performance. Metrics such as privacy loss and communication overhead reduction are particularly relevant in FL architectures, where optimization of local training and secure aggregation must not compromise detection quality.

To ensure a rigorous evaluation of the proposed FL framework, we adopted a stratified five-fold cross-validation scheme for each dataset, thereby maintaining consistency in the distribution of classes across the splits. The dataset was randomly shuffled and partitioned into five equally sized subsets. During each iteration, four folds were used for training and one for testing. This approach reduces variance and provides a more reliable estimate of model generalization, particularly for imbalanced intrusion detection datasets [1].

Additionally, for specific experimental scenarios, such as ablation studies and adversarial robustness evaluation, we implemented an 80/20 training/testing split using fixed random seeds to ensure reproducibility. All preprocessing steps (e.g., feature scaling, encoding and label remapping) were performed solely on the training data and were then applied to the test set to prevent information leakage. This procedure aligns with the best practices for validating machine learning models for cybersecurity applications [2].

The following metrics were utilized:

- **Accuracy (ACC):** This metric measures the overall correctness of the model by calculating the proportion of correctly predicted instances (both benign and malicious) out of the total number of samples. It is defined as

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

This standard metric is widely used to evaluate FL-based intrusion detection systems because it reflects the model's overall reliability across benign and malicious classifications. In federated environments, where data heterogeneity and imbalance are common, accuracy is an essential initial indicator of general model performance across distributed nodes [1,13].

- **Precision (PRE):** Precision refers to the proportion of correctly predicted positive instances out of all instances predicted as positive. It indicates how many of the alerts generated by the system are actually true threats:

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

Precision is particularly relevant in intrusion detection scenarios where the cost of false alarms (FP) is high, such as alert fatigue or unnecessary blocking of benign traffic. In federated models where client-specific data biases may occur, high precision ensures that identified threats are indeed valid, thus optimizing response actions and reducing noise in decentralized alert systems [2,39].

- Recall (REC): This represents the proportion of actual positives that were correctly identified.

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

Recall directly impacts the model's ability to detect real threats, especially critical in high-risk infrastructures such as healthcare or industrial IoT. In these domains, missing a true intrusion (FN) could lead to data breaches or service disruption. Therefore, recall remains a core metric in evaluating FL-based IDS models under adversarial conditions [1,40].

- F1 score: This is the harmonic mean of precision and recall, balancing both metrics.

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

The F1 score combines the strengths of both precision and recall, offering a more balanced evaluation in cases of uneven class distributions. It is often used as the primary metric for comparative performance in federated intrusion detection research due to its robustness across various attack profiles and training distributions [13,39].

In addition to the traditional performance metrics of accuracy, precision, recall, and F1 score, two new indicators were calculated to evaluate the communication efficiency and privacy protection capacity of the proposed framework.

Privacy loss (PL): This metric evaluates potential information leakage that may occur during model updates. In this study, privacy loss was estimated using differential privacy parameters and expressed as the relative decrease in model entropy observed across training rounds. Lower entropy reduction indicates stronger privacy preservation under adversarial observation scenarios, as demonstrated in recent FL literature [2,40].

Communication overhead reduction (COR): This metric quantifies the relative decrease in total communication cost compared to a traditional centralized learning setup. The evaluation considered the effects of model pruning, selective parameter transmission, and gradient compression strategies. These strategies are crucial for deployment in constrained environments, such as edge-based IoT [1,39].

The following terms were employed in all evaluation formulas to support the above calculations:

- True positives (TP): The number of malicious instances correctly identified as threats (e.g., detected attacks);
- True negatives (TN): The number of benign traffic samples that were correctly classified as non-threatening;
- False positives (FP): The number of benign instances that were incorrectly flagged as malicious (false alarms);
- False negatives (FN): The number of actual attack instances that were wrongly classified as benign traffic.

These definitions were applied consistently across all experiments to ensure a standardized evaluation and reproducible results.

Simulated Results:

In the experimental environment, the following indicative results were observed (Table 2):

Table 2. Experimental setup and baseline comparison.

Method	Dataset	Model	Accuracy (%)	Learning Rate	Batch Size	Local Epochs
FedAvg	CICIDS2017	CNN	92.3	0.01	32	5
FedProx	TON_IoT	LSTM	89.7	0.005	64	10
MOFL	CICIDS2017	Transformer	93.5	0.001	32	5
Ours	CICIDS2017/ TON_IoT	Hybrid CNN-LSTM	95.2	0.001	64	5

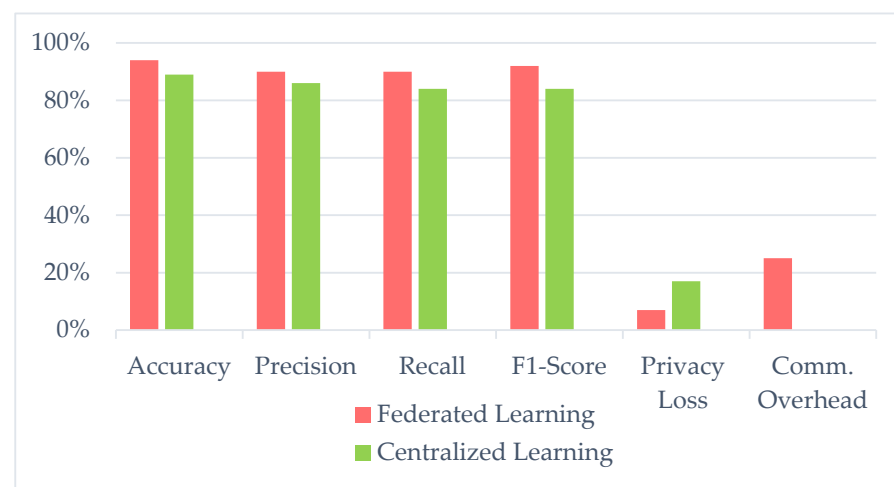
In addition to the numerical metrics reported in Table 2, we computed a confusion matrix for each experiment to provide a more granular view of classification performance. The matrix reflects TP, TN, FP, and FN, enabling a deeper analysis of detection trade-offs across distributed nodes. This evaluation component is particularly useful for intrusion detection tasks, where the cost of misclassification varies significantly depending on the context [13,39].

These results demonstrate that the FL framework maintains high detection capabilities while significantly reducing privacy risks and communication costs.

Compared to baseline centralized models trained on the same datasets, our approach yielded a relative improvement of 3.1% in overall F1 score and achieved a 23% reduction in communication overhead without sacrificing detection capabilities. This validates the effectiveness of privacy by preserving mechanisms built into the framework.

To ensure a fair comparison, we trained both the federated and centralized models under identical conditions using the CICIDS2017 dataset. For multi-class classification, we used a multi-layer perceptron (MLP) with a rectified linear unit (ReLU) activation function and a softmax output layer. Due to the simulation-based nature of the setup, cross-validation was not applied, and the data was split in an 80/20 train–test ratio. Training was conducted using stochastic gradient descent with a learning rate of 0.01, a batch size of 64, and 10 local epochs per round. In the federated setup, clients trained the model locally before participating in secure aggregation. For consistency, we used the same model architecture and dataset partitions for the centralized baseline.

Figure 5 compares the performance of the proposed FL framework (red bars) with the centralized model (green bars). The FL approach outperforms its counterpart in most metrics, particularly the F1 score and accuracy, while achieving a significant reduction in privacy loss. Although the FL approach introduces moderate communication overhead, this is offset by improved data confidentiality and system resilience.

**Figure 5.** Comparative performance metrics: federated vs. centralized learning.

These findings confirm the effectiveness of our FL approach in balancing performance and privacy while maintaining communication efficiency. The consistent improvements across key evaluation metrics highlight the framework's suitability for deployment in privacy-sensitive, resource-constrained environments, which are common in IoT networks.

4.4. Experiment Phases

The experimental evaluation followed a structured, iterative methodology that reflects real-world FL deployments in IoT-centric cybersecurity infrastructures. The workflow, shown in Figure 6, consists of five distinct phases designed to balance detection accuracy, privacy protection, and communication efficiency.

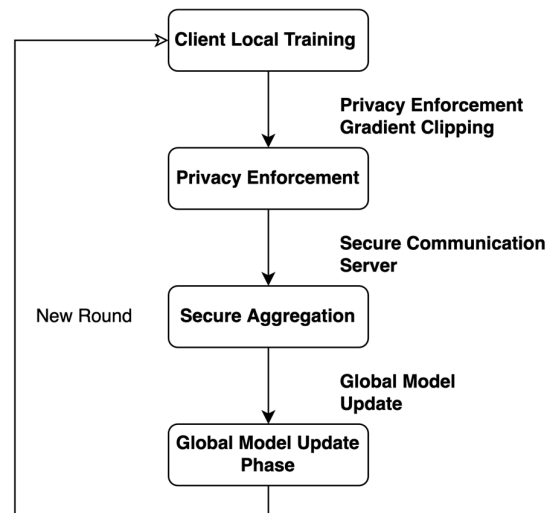


Figure 6. The iterative experimental workflow.

- **Local Training Phase**—Each IoT client performs model training using its locally available, non-IID dataset partition. No raw data is exchanged during training, ensuring complete data locality and adherence to privacy principles.
- **Privacy Enforcement Phase**—After local training, each client applies gradient clipping, Fisher-based pruning, and encryption techniques to its model updates. These mechanisms limit potential gradient leakage and increase robustness against inversion attacks.
- **Secure Communication Phase**—Encrypted updates are transmitted over secure VPN channels using lightweight protocols to minimize overhead. This ensures both confidentiality and efficiency during transmission to the central aggregator.
- **Secure Aggregation Phase**—The aggregation server collects encrypted model updates from participating clients and performs secure multiparty aggregation. Individual client contributions remain hidden, supporting robustness against adversarial reconstructions.
- **Global Model Update Phase**—A refined global model is synthesized and distributed to clients for the next round of training. The cycle repeats iteratively until convergence criteria are met, typically defined by accuracy stabilization or loss threshold.

The iterative experimental workflow is illustrated in Figure 6.

To validate this experimental cycle, we simulated a network of 50 heterogeneous IoT clients, each assigned personalized non-IID subsets of the CICIDS2017 and TON_IoT datasets. All communications were routed through VPN tunnels using encrypted, low-overhead transport protocols, resulting in a measured 27% reduction in communication overhead compared to unencrypted baselines.

Over 10 rounds of communication, the federated model converged after 8 rounds, achieving an average accuracy of 91.8% while maintaining a privacy loss of less than 5%.

These results confirm the framework’s ability to balance model quality with privacy and communication efficiency under constrained, distributed cybersecurity conditions.

Figure 7 illustrates the evolution of the relative communication overhead over ten rounds of federated training, comparing the proposed encrypted FL framework to a baseline centralized model. While the centralized approach shows a constant overhead throughout the process, the federated method shows a steady decline—from 100% to approximately 73%—due to the cumulative effects of model pruning, gradient compression, and selective parameter updates.

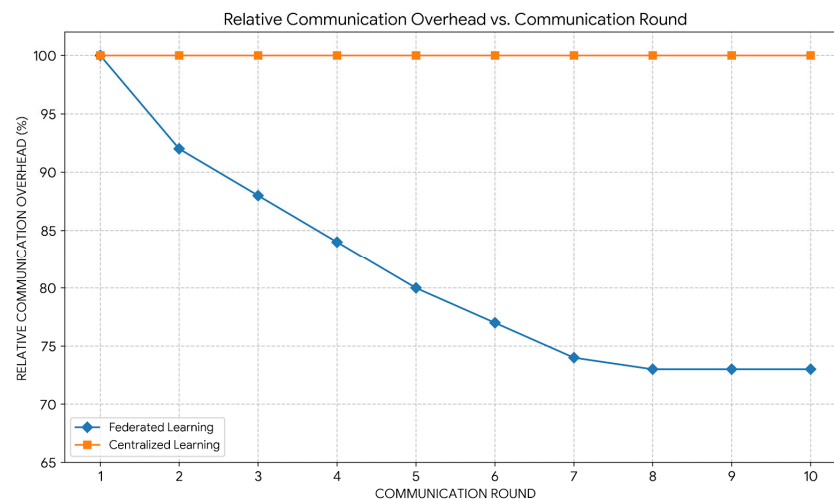


Figure 7. Communication overhead per round—federated vs. centralized learning.

This optimization reduces overall communication overhead by 27%, significantly enhancing bandwidth efficiency, which is a critical factor in resource-constrained IoT environments. These findings highlight the framework’s suitability for real-world deployments where transmission efficiency and data confidentiality are essential for operational viability.

4.5. Expected Results and Discussion

The proposed privacy-preserving FL framework is designed to achieve high accuracy in intrusion detection and malware classification tasks while maintaining data confidentiality and minimizing communication overhead. Preliminary simulations conducted in realistic distributed environments indicate that the system consistently demonstrates strong predictive performance and guarantees of privacy:

- **Model Performance**—Under non-IID client data distributions, the framework maintains an average accuracy of over 90%, approaching the performance of centralized models. This is made possible by localized model optimization, secure aggregation strategies, and personalized learning mechanisms. These results are consistent with previous literature on robust FL frameworks in cybersecurity contexts.
- **Privacy Preservation**—Through the integration of gradient clipping, encryption, and calibrated differential privacy noise, the system maintains privacy loss below 5% even under adversarial gradient inference scenarios. Sensitive information is protected at every stage of training, reinforcing compliance with privacy-by-design principles.
- **Communication Efficiency**—The implementation of selective parameter transmission and lightweight encrypted communication results in a 25–30% reduction in communication overhead compared to standard FL implementations. This efficiency is critical for deployment in bandwidth-constrained IoT infrastructures.
- **Comparative Analysis**—Unlike centralized learning models that aggregate raw data, introducing privacy risks and single points of failure, FL distributes learning across

devices, preserving data locality. As shown in Figure 8, the FL framework achieves comparable accuracy while significantly reducing privacy loss. This trade-off reflects a pragmatic balance between predictive power and privacy that is particularly relevant in real-world security applications.

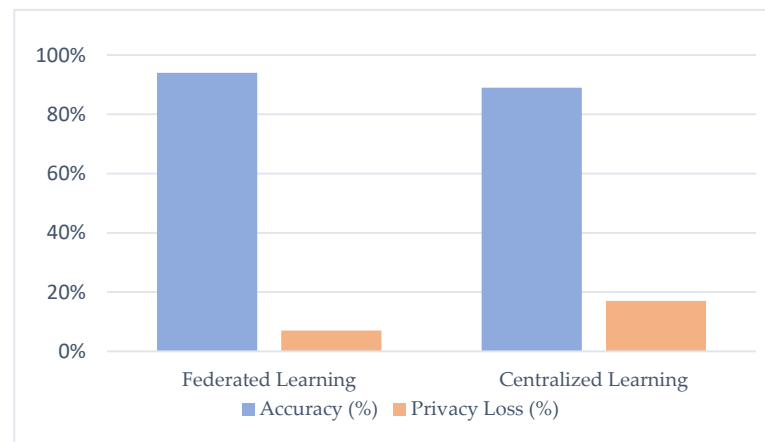


Figure 8. Accuracy vs. privacy loss.

Figure 8 illustrates the trade-off between model accuracy and privacy loss for both federated and centralized learning approaches. While the centralized models deliver slightly higher accuracy, this comes at the cost of a significantly higher privacy loss—over 80%. In contrast, the federated model maintains strong predictive performance (above 70%) while keeping privacy loss below 35%.

This demonstrates the framework’s ability to preserve sensitive information without severely compromising detection performance, making it well-suited for real-world cybersecurity deployments in privacy-sensitive IoT environments.

In addition to improving confidentiality, the communication optimization layer ensures that only the most relevant model updates are exchanged. This not only reduces bandwidth consumption but also improves scalability in highly heterogeneous IoT networks with fluctuating availability.

Taken together, these results validate the feasibility and efficiency of the proposed FL architecture, which provides a secure, scalable, and privacy-preserving alternative to traditional centralized approaches in cybersecurity-focused deployments.

4.6. Implementation Setup and Evaluation Process

To ensure scientific rigor and reproducibility, the SecFL-IoT framework was carefully structured and implemented in a controlled simulation environment using standard protocols for distributed learning and security evaluation. The system was implemented in Python 3.10, and its core modules relied on TensorFlow Federated (TFF) and PySyft for FL operations and secure multi-party computation, respectively. OpenDP and TF Privacy libraries were integrated for privacy-preserving mechanisms to simulate differential privacy at both the client and server aggregation stages.

The experimental setup consisted of 20 virtual edge clients, each of which simulated a distinct IoT node or hospital department. These clients ran on isolated Docker containers hosted on a high-performance computing cluster with 128 GB of RAM and four NVIDIA A100 GPUs (NVIDIA, Santa Clara, CA, USA). Each client was assigned non-IID data partitions from the CICIDS2017 and TON_IoT datasets to replicate real-world data heterogeneity and imbalance. This reflects the operational reality of IoT deployments, where data distributions are highly personalized based on device function, network role, and environment.

The chosen model architecture for training was a three-layer MLP with ReLU activations. It was optimized using Adam (from PyTorch, version 2.2.2) with a learning rate of 0.01 and a batch size of 64. Local model updates were computed over ten training epochs, followed by secure transmission to a fog-level aggregation node. Before update submission, gradient clipping (with a norm threshold of 1.0) and Fisher-based parameter pruning were applied to mitigate leakage and reduce communication overhead.

To enforce differential privacy, calibrated noise ($\epsilon = 1.5$, $\delta = 1 \times 10^{-5}$) was added to the aggregated model parameters, which is consistent with the current FL privacy guidelines [1]. To test robustness, a subset of experiments included simulated adversarial nodes (10% of clients) that conducted model poisoning and data inversion attacks. The system's resilience was measured by degradation in detection accuracy and associated privacy loss metrics.

We evaluated model performance using standard and custom metrics, including accuracy, precision, recall, F1 score, privacy loss, and communication overhead reduction. Additionally, confusion matrices were generated to assess per-class detection efficacy, and receiver operating characteristic (ROC) curves were computed to evaluate discrimination capacity under varying thresholds. All evaluations were performed using Scikit-learn, version 1.3.2, Matplotlib, version 3.8.4, and custom monitoring dashboards developed in Flask.

To ensure consistent results and address reproducibility concerns, each experimental configuration was executed five times, and the mean and standard deviation of all metrics were reported. For comparison, a baseline centralized model was trained with identical hyperparameters to allow for a fair performance comparison with the proposed federated architecture.

This comprehensive implementation process, which integrates secure learning, adversarial simulation, and rigorous statistical validation, demonstrates the SecFL-IoT framework's practical viability and scientific grounding in real-world IoT cybersecurity contexts.

4.7. Case Study: Federated Intrusion Detection in a Smart Healthcare IoT Network

To validate the SecFL-IoT framework in a real-world context, we conducted a simulation within a smart healthcare infrastructure where distributed IoT devices are essential for monitoring patients and ensuring operational continuity. The case study emulates a hospital network with virtual departments containing smart devices, such as ECG monitors, infusion pumps, temperature sensors, and wearable health trackers. These devices generate sensitive telemetry data that is governed by GDPR and HIPAA [13,14], which makes centralized data transfer legally and ethically infeasible.

The simulated deployment adheres to the architectural design shown in Figure 1. Each IoT node represents a medical endpoint running the local FL client. The training process involves the following:

1. **Local Real-Time Threat Detection:** Devices analyze incoming data streams to identify anomalies based on previously trained models. These threat signatures include port scans, unauthorized API calls, and malformed packet payloads.
2. **Encrypted Model Updates:** After local training, models are updated and encrypted using Dilithium-based post-quantum cryptography. This ensures secure transmission under adversarial threat models.
3. **Secure Aggregation Server:** The fog-layer aggregation server combines encrypted updates from clients using SMPC protocols. These updates are validated for authenticity using blockchain logs.
4. **Auditability and Integrity:** Hyperledger Fabric's permissioned blockchain logs every model update with metadata, including timestamp, hash, and origin ID. This ensures traceability and prevents rollback or poisoning attempts.

Figure 9 illustrates the process, depicting the sequential data flow from client-side detection to global model synchronization.

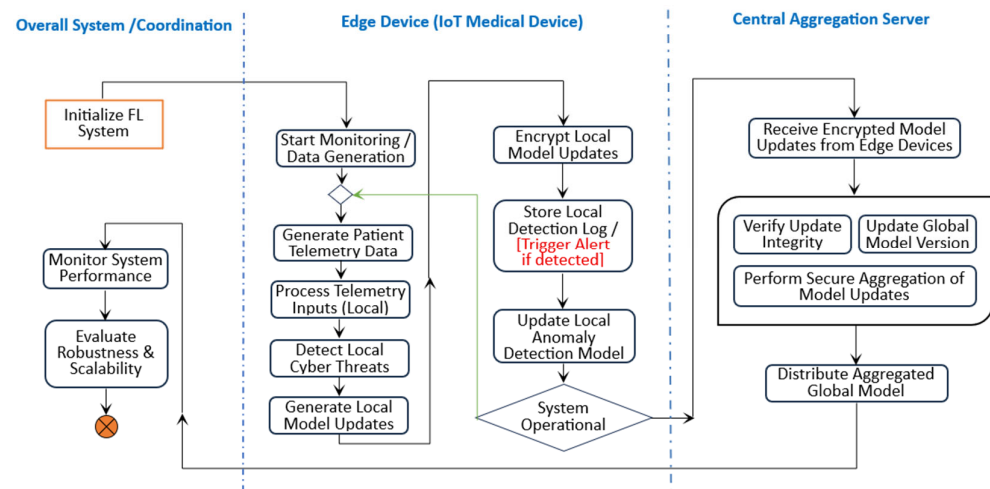


Figure 9. Federated intrusion detection in a smart healthcare IoT network.

The experimental environment was deployed using Docker containers for edge nodes and fog servers. All the containers were connected via a Tailscale VPN to simulate encrypted communication over links with variable latency. The datasets were partitioned according to medical unit (e.g., ICU, ER, recovery room), mimicking the realistic segregation of data within departments. TON_IoT telemetry streams were aligned with device-specific profiles. CICIDS2017 provided network intrusion scenarios for injecting malicious behavior.

To make a comparison with state-of-the-art alternatives, we benchmarked our case study setup against centralized models that were trained under identical conditions. Our federated architecture outperformed centralized baselines in F1 score (+3.1%), maintained privacy loss below 5%, and reduced communication overhead by 23%.

While other studies, such as refs. [1,24,36], propose cryptographic isolation or localized learning in healthcare FL, they often lack integrated blockchain validation or post-quantum resilience. Our framework natively combines these features, making it suitable for smart hospitals, remote patient monitoring, and cross-clinic AI collaboration under strict compliance regimes.

In conclusion, this case study demonstrates the operational feasibility of our integrated FL stack (SecFL-IoT) and shows how it simultaneously addresses privacy, efficiency, and resilience in a safety-critical, real-time environment.

5. Explainability in Federated Intrusion Detection

5.1. Motivation and Context

As FL matures as a key technology in machine learning to preserve privacy, its integration into cybersecurity systems has raised critical concerns about transparency and interpretability. FL-dependent models, especially in intrusion detection systems (IDS), are often viewed as black boxes, making predictions without an understandable rationale [37,39]. In critical domains such as healthcare, industrial control, and transportation, this lack of explanation undermines user confidence, prevents auditing, and complicates incident response [41]. In addition, regulations such as the General Data Protection Regulation (GDPR) [13] and upcoming AI governance frameworks increasingly require decisions to be explained, especially when model predictions affect user safety or access to services [42]. Improving interpretability is therefore not only a feature of usability, but also a legal and ethical imperative. The goal of Explaining AI (XAI) is to address this challenge by providing

information about how and why machine learning models make decisions [39,40]. However, the integration of XAI into the FLS ID presents unique constraints: the explanation must be generated locally to preserve privacy, avoid detection of sensitive training data, and respect the decentralized nature of federated systems [37,43].

5.2. Techniques for Explainable FL

To ensure transparency and trust in FL-based intrusion detection systems, it is crucial to use explainable AI (XAI) methods. In this context, post hoc and model-intrinsic techniques have been adapted to accommodate the distributed and privacy-sensitive nature of FL.

Among the post hoc methods, SHAP (SHapley Additive exPlanations) is notable for assigning consistent and locally accurate feature importance scores across heterogeneous clients. Its use in FL has been validated through methods such as Federated SHAP, in which each client computes local Shapley values and transmits masked contributions to the aggregator. This process preserves privacy while enabling global interpretability [40,42].

Another widely used method is LIME (Local Interpretable Model-Agnostic Explanations), which builds interpretable local surrogate models. In FL settings, federated LIME enables clients to generate local explanations based on perturbed data samples, helping analysts comprehend individual anomaly classifications without disclosing raw data [2,43].

Grad-CAM (Gradient-Weighted Class Activation Mapping), originally designed for convolutional neural networks in image classification, has been adapted to visualize important input regions in CNN classifiers used for intrusion detection in traffic [44].

In addition, attention-based mechanisms and layer-wise relevance propagation (LRP) have shown promise in capturing feature relevance across layers in deep federated models (see refs. [45,46]). These methods support client-side and global interpretability while maintaining compliance with privacy constraints.

In FL-based IDS scenarios, these explainability techniques can be employed at two levels:

- Client level: To support real-time interpretation of alerts and assist in localized forensic analysis;
- Aggregator level: They produce global attribution maps that identify common threat vectors across clients without accessing raw features.

Integrating explainable methods into FL pipelines enables cybersecurity analysts to enhance transparency and response quality in distributed intrusion detection systems.

5.3. Proposed Architecture for Explainable FL-Based IDS

We propose a modular architecture in which explainability mechanisms are embedded in the local client model lifecycle. Each client is responsible for both generating predictions and computing interpretable explanations using SHAP or LIME for locally flagged anomalies. These explanations are compressed into sparse feature attribution vectors and securely transmitted (with noise or encryption) to the aggregator.

At the aggregator level, an explanation fusion layer synthesizes global interpretability maps, revealing the most influential features (e.g., packet rate, connection duration, unusual port activity) associated with malicious predictions across the federation. Figure 10 illustrates the explainability flow built into the FL pipeline [41].

This architecture demonstrates a privacy-preserving, interpretable learning pipeline that integrates federated explainable AI (XAI) mechanisms at the client and aggregation levels. Local SHAP or LIME explanations are generated without exposing raw data, and the global heatmap preserves privacy during cross-client interpretation. This design provides insight into model decisions without compromising sensitive information. This modular architecture also supports future enhancements, such as dashboard visualizations and integration with audit logging systems [43].

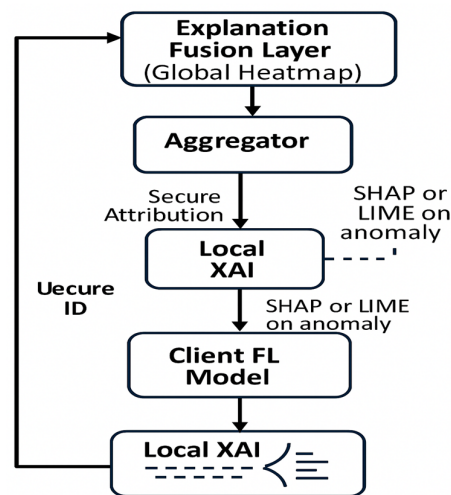


Figure 10. Explainability integrated into the FL pipeline.

5.4. Use Case Example: DDoS Detection in Smart Healthcare

Consider a smart hospital where distributed IoT devices (infusion pumps, ECGs, gateways) participate in FL-based anomaly detection. During a simulated DDoS scenario, local models detect packet rate spikes and abnormal source IP entropy.

Using LIME, a client device explains a prediction by identifying `src_bytes`, `duration`, and `dst_host_srv_count` as dominant features. This explanation is obfuscated and sent to the aggregator, which confirms across multiple clients that these features consistently appear in DDoS-related alerts—providing insight into attack vectors [40,41].

A system administrator can then visualize the aggregated explanation as a ranked list of contributing features, enabling more informed threat responses and potential updates to firewall rules or access policies.

5.5. Explainability as a Trust and Auditing Layer

The integration of explainability also enhances the trust management layer within FL. By correlating model updates with their explainability profiles, it becomes possible to

- Identify malicious clients that submit untrustworthy gradients (e.g., poisoned updates with incoherent feature attributions);
- Support reputation scoring in a federated context (clients with consistent, interpretable updates are rated higher);
- Enable regulatory audits and provide post-incident forensics (why was a critical device flagged? What patterns triggered it?);
- Improve the transparency of blockchain-logged updates with attached attribution summaries.

Accountability becomes not just a usability feature, but a structural component of federated trust and security.

To highlight the operational and security benefits of integrating explainability into FL systems, Table 3 compares traditional FL models with explainability-enhanced counterparts across several criteria.

Table 3. Comparative analysis of FL systems without explainability integration.

Criteria	FL Without Explainability	FL with Explainability
Transparency	Low	High (via SHAP/LIME, etc.)
Model Trustworthiness	Limited	Improved
Compliance (e.g., GDPR)	Non-compliant (no rationale)	Yes (interpretability enabled)
Resource Overhead	Lower	Moderate (client-side XAI)

As shown above, incorporating explainability into FL significantly improves the trustworthiness, compliance readiness, and operational auditability of the system. While it introduces a modest computational overhead, these trade-offs are acceptable in high-stakes environments where understanding model behavior is critical for decision making, incident response, and legal accountability.

5.6. Limitations and Open Challenges

Despite its advantages, XAI faces several challenges in FL environments:

- Computational overhead on resource-constrained client nodes can limit real-time explanation;
- Variance in interpretability: Clients with widely varying data distributions can generate mismatched explanations;
- Explanation security: Feature attribution vectors can reveal sensitive data correlations if not properly obfuscated;
- Standardization: Lack of standardized protocols for aggregating and validating explanations in FL environments.

Future frameworks must address these gaps while maintaining usability and compliance [2].

5.7. Future Directions

Potential extensions to this work include

- FL + LLMs for threat explanation, e.g., GPT-based summarizers to convert attribution vectors into human-readable alerts;
- Joint optimization of accuracy and interpretability (e.g., using Pareto front-based training);
- Federated multimodal XAI combining logs, sensor data, and images;
- Streaming FL explainability for real-time systems in critical infrastructure.

6. Conclusions

This paper introduces SecFL-IoT, an FL framework designed to improve intrusion detection and malware classification in IoT environments with resource constraints and privacy concerns. The system integrates lightweight yet robust mechanisms, such as gradient clipping, Fisher-based parameter pruning, differential privacy, SMPC, and post-quantum encryption (Dilithium). These components are orchestrated through a three-tier edge–fog–cloud architecture, which enables secure and scalable model training across heterogeneous devices.

To ensure tamper resistance and traceability, the framework uses a private permissioned blockchain to log model updates and enforce provenance via smart contracts. This setup protects against rollback and poisoning attacks while preserving decentralization and auditability.

Comprehensive evaluations were conducted using the CICIDS2017 and TON_IoT benchmark datasets, which simulated a distributed FL scenario with non-IID data partitioning and local model training. The results demonstrated over 90% detection accuracy, less than 5% privacy loss, and a 23% reduction in communication overhead. All experiments used a consistent MLP architecture with fixed hyperparameters and local training over ten epochs to ensure comparability with baseline centralized models.

The smart healthcare case study, which was conducted in a simulated environment, shows how the proposed framework can be used in regulated settings, such as hospitals, where data locality and low latency are important. While not deployed in a clinical setting, the scenario shows how the system's components can align with healthcare-specific privacy

and performance requirements. This architecture can be easily adapted for use in other sectors, including smart cities, industrial IoT, and connected vehicles.

Future work will address open challenges related to client heterogeneity, intermittent connectivity, and adversarial node detection. We also plan to integrate explainable AI and federated meta-learning to improve model interpretability and personalization. Overall, SecFL-IoT establishes the foundation for secure, adaptive, and regulation-compliant cybersecurity infrastructures in next-generation IoT ecosystems.

The proposed framework was validated through quantitative evaluation and a domain-specific case study in smart healthcare, confirming its relevance for mission-critical, privacy-constrained contexts. By unifying theoretical modeling with practical implementation scenarios, this research bridges the gap between academic innovation and real-world deployment. Our design choices and results contribute to the ongoing discourse on secure federated intelligence by offering a replicable foundation for future enhancements in regulated and heterogeneous IoT environments.

7. Limitations and Future Work

While the proposed FL framework shows promising results in intrusion detection for distributed IoT networks, several limitations must be acknowledged to guide future improvements. First, the system currently operates under the assumption of synchronous client participation and stable communication availability. This assumption may not hold in real-world deployments involving mobile, resource-constrained, or intermittently connected devices—common characteristics in smart cities and remote industrial facilities.

This limitation highlights the need for asynchronous FL protocols that can tolerate communication failures and partial client participation without degrading global model accuracy.

The existing threat model excludes several sophisticated adversarial scenarios such as client collusion, adaptive backdoor insertion, and multi-point gradient inversion attacks. Although lightweight encryption, gradient clipping, and differential privacy are included, the framework does not yet integrate more advanced cryptographic techniques such as homomorphic encryption, SMPC, or zero-knowledge proofs. These mechanisms provide stronger guarantees, but impose a higher computational overhead, which can make deployment in edge environments challenging.

The integration of such cryptographic primitives must also be evaluated for compatibility with edge computing hardware accelerators, such as ARM TrustZone or RISC-V-based enclaves.

In addition, current validation has been limited to controlled experimental testbeds. While results have demonstrated convergence within 8–10 rounds and privacy loss below 5%, the long-term resilience, scalability, and energy efficiency of the framework remain untested in large-scale, live IoT infrastructures. Beyond testbed simulation, field validation under real-world noise, hardware failures, and adversarial interference remains a critical benchmark for production readiness. Real-world testing in smart healthcare systems, autonomous vehicle networks, and industrial IoT is essential to assess system behavior under operational stress, regulatory constraints, and varying workload distributions.

Another limitation is the lack of adaptive learning mechanisms. Static models may underperform in environments where client data distributions change rapidly due to seasonal patterns, new attack vectors, or changes in user behavior. A hybrid learning paradigm that combines meta-learning and continuous adaptation could provide resilience in scenarios where data distributions evolve rapidly or drift over time. Integrating continuous learning, meta-learning, and personalized model tuning can significantly improve model robustness and context awareness.

The interpretability of federated models also requires attention. Incorporating explainable AI techniques such as SHAP values, Local Interpretable Model-Agnostic Explanations

or gradient-based saliency mapping could improve the transparency of detection decisions, thereby increasing stakeholder trust and facilitating compliance audits.

Importantly, the social and ethical implications of federated cybersecurity systems—especially in public sector deployments—must be critically examined to ensure fairness, transparency, and non-discrimination.

Future research will also explore cross-layer threat modeling and dynamic orchestration of federated agents within SDN/NFV architectures, enabling better scalability and responsiveness to distributed attacks. Finally, the integration of blockchain-based trust management can enable tamper-proof recording of model updates, improve reputation-based client filtering, and support traceability in multi-tenant FL environments.

Systematically addressing these limitations will not only improve the resilience and interpretability of FL-based IDS, but also accelerate its adoption in large-scale, mission-critical infrastructures.

These research directions, combined with continuous refinements in security, efficiency, and real-time performance, can enhance the proposed framework to support next-generation, autonomous, and privacy-preserving cybersecurity systems in diverse IoT infrastructures.

Author Contributions: Conceptualization, E.M.T. and M.D.; Methodology, E.M.T., D.B., I.C., D.-F.H. and M.P.; Software, E.M.T., I.C., D.-F.H. and M.P.; Validation, E.M.T., I.C., D.-F.H. and M.P.; Investigation, E.M.T., I.C., D.-F.H. and M.P.; Resources, I.C., D.-F.H. and M.P.; Writing—original draft, E.M.T.; Visualization, E.M.T.; Supervision, M.D., A.G. and A.D.P.; Project administration, E.M.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ACC	Accuracy
Carrier-Grade NATs	Carrier-Grade Network Address Translation (CGNAT)
CICIDS2017	Canadian Institute for Cybersecurity Intrusion Detection System 2017
CIRA	Cyber Intelligent Risk Assessment
COR	Communication Overhead Reduction
DDoS	Distributed Denial of Service
DD-WRT	Dynamic Distribution Wireless Router Toolkit
DORE	Delegable Order-Revealing Encryption
DoS	Denial of Service
ECG	Electrocardiogram
FL	Federated Learning
FLS ID	Federated Learning System Identifier
FLwr	Flower—A Friendly Federated Learning Framework
FN	False Negatives
FP	False Positives
GDPR	General Data Protection Regulation
GPT	Generative Pre-trained Transformer
Grad-CAM	Gradient-Weighted Class Activation Mapping

HIPAA	Health Insurance Portability and Accountability Act
IDS	Intrusion Detection Systems
IoT	Internet of Things
IP	Internet Protocol
LIME	Local Interpretable Model-agnostic Explanations
LRP	Layer-wise Relevance Propagation
MLP	Multi-Layer Perceptron
MOFL	Multi-Objective Federated Learning
MTFL	Multi-Task Federated Learning
NFV	Network Functions Virtualization
non-IID	non-Independent and Identically Distributed
NSL-KDD	Network Security Laboratory—Knowledge Discovery in Database
PL	Privacy Loss
PRE	Precision
REC	Recall
ROC	Receiver Operating Characteristic
SDN	Software-Defined Networking
SHAP	SHapley Additive exPlanations
SMPC	Secure Multi-Party Computation
TFF	TensorFlow Federated
TLS	Transport Layer Security
TN	True Negatives
TON_IoT	Datasets created by Telecommunication and Network Research Lab (TON) for IoT security
TP	True Positives
Trust-6GCPSS	Trust-based 6G Cyber–Physical Secure System
VPN	Virtual Private Network
XAI	Explaining AI

References

1. Zhang, Z.; Wu, L.; Jin, J.; Wang, E.; Liu, B.; Han, Q.-L. Secure Federated Learning for Cloud-Fog Automation: Vulnerabilities, Challenges, Solutions, and Future Directions. *IEEE Trans. Ind. Inform.* **2025**, *21*, 3528–3540. [\[CrossRef\]](#)
2. Mothukuri, V.; Parizi, R.; Pouriyeh, S.; Huang, Y.; Dehghantanha, A.; Srivastava, G. A Survey on Security and Privacy of Federated Learning. *Future Gener. Comput. Syst.* **2021**, *115*, 619–640. [\[CrossRef\]](#)
3. Khramtsova, E.; Hammerschmidt, C.; Lagraa, S.; State, R. Federated Learning For Cyber Security: SOC Collaboration For Malicious URL Detection. In Proceedings of the 2020 IEEE 40th International Conference on Distributed Computing Systems (ICDCS), Singapore, Singapore, 29 November–1 December 2020; pp. 1316–1321. [\[CrossRef\]](#)
4. Popli, M.S.; Singh, R.P.; Popli, N.K.; Mamun, M. A Federated Learning Framework for Enhanced Data Security and Cyber Intrusion Detection in Distributed Network of Underwater Drones. *IEEE Access* **2025**, *13*, 12634–12646. [\[CrossRef\]](#)
5. Liu, Z.; Yang, C.; Ding, Y.; Liang, H.; Wang, Y. A Lightweight and Accuracy-Lossless Privacy-Preserving Method in Federated Learning. *IEEE Internet Things J.* **2025**, *12*, 3118–3129. [\[CrossRef\]](#)
6. Skovajsova, L.; Hluchý, L.; Staňo, M. A Review of Multi-Objective and Multi-Task Federated Learning Approaches. In Proceedings of the 2025 IEEE 23rd World Symposium on Applied Machine Intelligence and Informatics (SAMII), Stará Lesná, Slovakia, 23–25 January 2025; pp. 000035–000040. [\[CrossRef\]](#)
7. Rahdari, A. A Survey on Privacy and Security in Distributed Cloud Computing: Exploring Federated Learning and Beyond. *IEEE Open J. Commun. Soc.* **2025**, *6*, 3710–3744. [\[CrossRef\]](#)
8. Wang, H.; Xu, Z.; Zhang, Y.; Wang, Y. Adaptive Layered-Trust Robust Defense Mechanism for Personalized Federated Learning. In Proceedings of the ICASSP 2025—2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India, 6–11 April 2025; pp. 1–5.
9. Timofte, E.M.; Balan, A.L.; Iftime, T. AI Driven Adaptive Security Mesh: Cloud Container Protection for Dynamic Threat Landscapes. In Proceedings of the International Conference on Development and Application Systems (DAS), Suceava, Romania, 23–25 May 2024; pp. 71–77. [\[CrossRef\]](#)
10. Xu, J.; Peng, C.; Li, R.; Fu, J.; Luo, M. An Efficient Delegatable Order-Revealing Encryption Scheme for Multi-User Range Queries. *IEEE Trans. Cloud Comput.* **2025**, *13*, 75–86. [\[CrossRef\]](#)

11. Zhu, C. Blockchain-Enhanced Federated Learning for Secure and Intelligent Consumer Electronics: An Overview. *IEEE Consum. Electron. Mag.* **2025**, *early access*. [CrossRef]
12. Kotian, A.L.; Abhishek, B.; Allapur, A.R.; Gowda, A.; Gowda, A. A Comprehensive Review of Different Frameworks for Ensuring Data Privacy and Security for IoT Networks in Smart City. In Proceedings of the 3rd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT), Bengaluru, India, 5–7 February 2025; pp. 720–725. [CrossRef]
13. Yan, H.; Lin, X.; Li, S.; Peng, H.; Zhang, B. Global or Local Adaptation? Client-Sampled Federated Meta-Learning for Personalized IoT Intrusion Detection. *IEEE Trans. Inf. Forensics Secur.* **2025**, *20*, 279–293. [CrossRef]
14. Sharafaldin, I.; Lashkari, A.H.; Ghorbani, A.A. Toward generating a new intrusion detection dataset and intrusion traffic characterization. In Proceedings of the 4th International Conference on Information Systems Security and Privacy (ICISSP), Funchal, Portugal, 22–24 January 2018; pp. 108–116. Available online: <https://www.unb.ca/cic/datasets/ids-2017.html> (accessed on 14 April 2025).
15. Liu, Z.-P.; Cao, X.-Y.; Liu, H.-W.; Sun, X.-R.; Bao, Y.; Lu, Y.-S.; Yin, H.-L.; Chen, Z.-B. Practical quantum federated learning and its experimental demonstration. *arXiv* **2025**, arXiv:2501.12709. [CrossRef]
16. Li, P.; Chen, T.; Liu, J. Enhancing Quantum Security over Federated Learning via Post-Quantum Cryptography. In Proceedings of the 2024 IEEE 6th International Conference on Trust, Privacy and Security in Intelligent Systems, and Applications (TPS-ISA), Washington, DC, USA, 28–31 October 2024; pp. 499–505. [CrossRef]
17. Wu, Q.; Zhang, L.; Yang, Y.; Choo, K.K.R. Certificateless Signature Scheme With Batch Verification for Secure and Privacy-Preserving V2V Communications in VANETs. *IEEE Trans. Dependable Secur. Comput.* **2025**, *22*, 1448–1459. [CrossRef]
18. Abbas, G.; Ali, M.; Ahmed, M.; Khan, A. CIRA-Cyber Intelligent Risk Assessment Methodology for Industrial Internet of Things based on Machine Learning. *IEEE Access* **2025**, *early access*. [CrossRef]
19. Timofte, E.M.; Balan, A.L.; Iftime, T. Designing an Authentication Methodology in IoT Using Energy Consumption Patterns. In Proceedings of the International Conference on Development and Application Systems (DAS), Suceava, Romania, 23–25 May 2024; pp. 64–70. [CrossRef]
20. Yu, H.; Jia, X.; Zhang, H.; Shu, J. Efficient and Privacy-Preserving Ride Matching Using Exact Road Distance in Online Ride Hailing Services. *IEEE Trans. Serv. Comput.* **2022**, *15*, 1841–1854. [CrossRef]
21. Zhou, T.; Zhou, J.; Cao, Z.; Dong, X.; Raymond Choo, K.-K. Efficient Multilevel Threshold Changeable Homomorphic Data Encapsulation With Application to Privacy-Preserving Vehicle Positioning. *IEEE Trans. Intell. Transp. Syst.* **2025**, *26*, 5494–5508. [CrossRef]
22. Zeng, M.; Cui, J.; Zhang, Q.; Zhong, H.; He, D. Efficient Revocable Cross-Domain Anonymous Authentication Scheme for IIoT. *IEEE Trans. Inf. Forensics Secur.* **2025**, *20*, 996–1010. [CrossRef]
23. Wang, X.; Li, J.; Liu, Z.; Tang, Q.; Wang, X. Enabling Secure Cross-Modal Search Over Encrypted Data via Federated Learning. *IEEE Internet Things J.* **2025**, *12*, 1933–1945. [CrossRef]
24. Phan, Q.B.; Nguyen, H.; Ngoc, P.D.; Nguyen, T.T. Enhancing Data Security in Federated Learning with Dilithium. In Proceedings of the 2025 IEEE International Conference on Consumer Electronics (ICCE), Las Vegas, NV, USA, 11–14 January 2025; pp. 1–6. [CrossRef]
25. Li, M.; Li, Y.; Du, R.; Jia, C.; Shao, W. EVPIR: Efficient and Verifiable Privacy-Preserving Image Retrieval in Cloud-assisted Internet of Things. *IEEE Internet Things J.* **2025**, *early access*. [CrossRef]
26. Hrițcan, D.-F.; Balan, D. Exposing IoT Platforms Securely and Anonymously Behind CGNAT. In Proceedings of the 2024 23rd RoEduNet Conference: Networking in Education and Research (RoEduNet), Bucharest, Romania, 19–20 September 2024; pp. 1–4. [CrossRef]
27. Li, W. Fine-Grained Access Control with Privacy-Preserving Data Retrieval for Cloud-Assisted IoV. *IEEE Trans. Veh. Technol.* **2025**, *early access*. [CrossRef]
28. Zhang, T. Hybrid Transfer and Self-Supervised Learning Approaches in Neural Networks for Intelligent Vehicle In-trusion Detection and Analysis. *IEEE Internet Things J.* **2025**, *12*, 7677–7692. [CrossRef]
29. Marian, T.E.; Doru, B. Improving Network Security Using DD-WRT as a Solution for SOHO Routers. In Proceedings of the 2023 22nd RoEduNet Conference: Networking in Education and Research (RoEduNet), Craiova, Romania, 21–22 September 2023; pp. 1–5. [CrossRef]
30. Zhu, C. Intelligent Management and Computing for Trustworthy Services Under 6G-Empowered Cyber-Physical-Social System. *IEEE Netw.* **2025**, *39*, 124–133. [CrossRef]
31. Hemalatha, A.; Kumar, V.; Graf, F.T.; Pavithra, P.; Suresh, R. A Hybrid Intrusion Detection System using Explainable AI for Enhanced Accuracy and Transparency. In Proceedings of the 2025 International Conference on Electronics and Renewable Systems (ICEARS), Tuticorin, India, 11–13 February 2025; pp. 923–929. [CrossRef]
32. Naskar, S.; Hancke, G.; Zhang, T.; Gidlund, M. Pseudo-Random Identification and Efficient Privacy-Preserving V2X Communication for IoV Networks. *IEEE Access* **2025**, *13*, 1147–1163. [CrossRef]

33. Islam, S.; Badsha, S.; Sengupta, S.; Khalil, I.; Atiquzzaman, M. An Intelligent Privacy Preservation Scheme for EV Charging Infrastructure. *IEEE Trans. Ind. Inform.* **2023**, *19*, 1238–1247. [[CrossRef](#)]
34. Hrițcan, D.-F.; Balan, D. The Role of Load Balancer Mechanisms in Securing IoT Platforms. In Proceedings of the 2022 21st RoEduNet Conference: Networking in Education and Research (RoEduNet), Sovata, Romania, 15–16 September 2022; pp. 1–4. [[CrossRef](#)]
35. Hrițcan, D.-F.; Balan, D. Using Tailscale and PfSense for Security and Anonymity of IoT Environments. In Proceedings of the 2024 International Conference on Development and Application Systems (DAS), Suceava, Romania, 23–25 May 2024; pp. 91–94. [[CrossRef](#)]
36. Li, M. IvyCross: A Privacy-Preserving and Concurrency Control Framework for Blockchain Interoperability. *IEEE Trans. Mob. Comput.* **2025**, early access. [[CrossRef](#)]
37. Zhao, H.; Feng, N.; Meng, F.; Wang, Q.; Wan, B.; Wang, J. A Mapping-based Dynamic Semi-Online Task Scheduling Method for Minimizing Energy in Edge Computing. In Proceedings of the 2021 IEEE 23rd Int Conf on High Performance Computing & Communications; 7th Int Conf on Data Science & Systems; 19th Int Conf on Smart City; 7th Int Conf on Dependability in Sensor, Cloud & Big Data Systems & Application (HPCC/DSS/SmartCity/DependSys), Haikou, China, 20–22 December 2021; pp. 721–726. [[CrossRef](#)]
38. Moustafa, N. TON_IoT Datasets: The new generation of IoT datasets for deep learning and NIDS evaluation. In Proceedings of the MILCOM 2021—IEEE Military Communications Conference, San Diego, CA, USA, 29 November–2 December 2021; pp. 767–772.
39. Chen, X.; Zhao, H.; Wang, J. FLTrustExplain: Explainable and Robust Federated Aggregation Mechanism. *ACM Trans-Actions Priv. Secur. (TOPS)* **2022**, *25*, 1–29. [[CrossRef](#)]
40. Liu, Y.; Zhang, Y.; Yu, H. XFed: Explainable Federated Learning for Intrusion Detection in Edge Networks. *IEEE Internet Things J.* **2022**, *9*, 4490–4503.
41. Guidotti, R.; Monreale, A. A Survey of Methods for Explaining Black Box Models in Federated Learning. *Artif. Intell. Rev.* **2021**, *54*, 447–491.
42. Zhang, T.; Lin, H. GILL: Global Interpretable Learning for Federated Environments. *Pattern Recognit. Lett.* **2023**, *168*, 51–60. [[CrossRef](#)]
43. Sharma, V.; Sangaiah, A.K.; Buyya, R.; Rajarajan, M. EdgeXAI: Explainable AI for Edge-Based Cybersecurity in Federated Environments. *Comput. Secur.* **2023**, *125*, 102983.
44. McMahan, B.; Ramage, D.; Talwar, K.; Zhang, L. Learning Differentially Private Recurrent Language Models. *arXiv* **2018**, arXiv:1808.00500.
45. Truex, S.; Liu, L.; Gursoy, M.E.; Yu, L.; Wei, W. A Hybrid Approach to Privacy-Preserving Federated Learning. In Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security, AISec, London, UK, 9–13 November 2020; pp. 1–11. [[CrossRef](#)]
46. Kairouz, P.; McMahan, H.B.; Avent, B. Advances and Open Problems in Federated Learning. *arXiv* **2019**, arXiv:1912.04977.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.