Journal of Intelligent Decision Support System (IDSS) 9 (1) (2026), pp. 52-62

Published by: Institute of Computer Science (IOCS)

Journal of Intelligent Decision Support System (IDSS)

Journal homepage: www.idss.iocspublisher.org

Distributed cyber defense framework based on federated learning for attack detection in defense infrastructure

Hondor Saragih[1], Hoga Saragih[2], Jonson Manurung[3], Rochedi Idul Adha[4], Frainskoy Rio Naibaho[5]

[1,3,4]*Informatika, Universitas Pertahanan Republik Indonesia, Bogor, Indonesia*
[2] *Teknik Informatika, Universitas Bakrie, Jakarta, Indonesia*
[4]*Teknik Informatika, Institut Agama Kristen Negeri Tarutung, Tarutung, Indonesia*

Article Info

Article history:

Received Jan 15, 2026
Revised Mar 26, 2026
Accepted Mar 29, 2026

Keywords:

Federated Learning
Intrusion Detection
Deep Learning
Anomaly Detection
Explainable AI

ABSTRACT

Cyber threats targeting defense infrastructure have escalated in complexity, rendering centralized intrusion detection systems insufficient due to their inability to guarantee data privacy across distributed military nodes. This study proposes a distributed cyber defense framework that employs federated learning to enable collaborative model training without transmitting raw network traffic beyond individual nodes. The framework integrates an adaptive aggregation strategy combining FedAvg and FedProx, a hybrid deep learning architecture consisting of convolutional neural networks and long short term memory networks, an autoencoder module for unsupervised anomaly detection, a Byzantine robust aggregation mechanism, and post hoc explainability through SHAP and LIME. Experiments were conducted on CIC IDS 2017, CIC IDS 2018, UNSW NB15, and a synthetically generated military network traffic dataset. The proposed framework attained a peak accuracy of 98.74% and an F1 score of 98.12% on CIC IDS 2017, consistently outperforming five baseline methods by up to 5.29 percentage points in F1 score. Future work will investigate differential privacy integration and model compression for deployment on resource constrained tactical edge devices.

Corresponding Author:

Hondor Saragih,
Informatika,
Universitas Pertahanan Republik Indonesia,
Kawasan IPSC Sentul, Sukahati, Kec. Citeureup, Kabupaten Bogor, Jawa Barat 16810, Indonesia.
Email: hondor.saragih@idu.ac.id

Introduction

The rapid advancement of digital technology has fundamentally transformed the operational landscape of modern defense infrastructure. Military networks, command and control systems, and critical defense assets are increasingly interconnected through complex digital ecosystems, creating a vast and vulnerable attack surface that adversaries can exploit. Cyber threats targeting defense infrastructure have grown substantially in both frequency and sophistication, with nation state actors deploying advanced persistent threats, large scale distributed denial of service campaigns, and multi stage malware to compromise sensitive military systems (Ahuja et al., 2024; Hieu & Son, 2025; Maasaoui et al., 2025). The consequences of successful intrusions extend beyond data theft to include disruption of operational

readiness, compromise of classified intelligence, and potential paralysis of critical defense capabilities. Traditional intrusion detection systems that rely on centralized data collection and rule based mechanisms are no longer sufficient to address the dynamic nature of modern cyber warfare (Meliboev et al., 2022; Siddiqi & Pak, 2022). These conventional approaches suffer from significant limitations, including high false positive rates, limited adaptability to zero day attacks, and the fundamental inability to process the volume and velocity of network traffic generated across large scale defense networks. Furthermore, the centralized aggregation of sensitive military network data introduces serious privacy and sovereignty concerns, as raw traffic data from disparate defense nodes cannot be safely transmitted to a central processing facility without risking operational security compromise.

The challenge of building an effective intrusion detection system for defense infrastructure is compounded by the distributed and heterogeneous nature of military networks. Modern defense operations span multiple geographic locations, encompassing forward operating bases, naval vessels, airborne platforms, and strategic command centers, each generating distinct patterns of network traffic with varying characteristics. Machine learning based approaches have demonstrated considerable promise in addressing cybersecurity challenges, yet their deployment in federated defense environments remains constrained by the fundamental tension between model performance and data privacy (Du et al., 2024; Umair et al., 2022). Conventional machine learning models require the aggregation of training data at a central server, which creates unacceptable risks in operational defense contexts where data sovereignty and compartmentalization are paramount (Dhrir, Charfeddine, Kammoun, et al., 2025; Hua & Xi, 2025; S. Zhang et al., 2025). The absence of a privacy preserving, decentralized learning framework capable of training high performance detection models without exposing raw network data represents a critical gap in current cybersecurity research. Addressing this gap requires not only algorithmic innovation but also a principled architectural design that can accommodate the non IID distribution of traffic data across heterogeneous defense nodes while maintaining robust detection performance against sophisticated adversarial attacks.

Previous research has explored various machine learning methodologies for intrusion detection with notable results. Centralized deep learning approaches utilizing convolutional neural networks and long short term memory architectures have achieved state of the art performance on benchmark datasets such as CIC IDS 2017 and UNSW NB15, demonstrating the power of neural architectures in capturing complex attack patterns. Studies employing autoencoders for anomaly detection have shown that unsupervised representation learning can effectively identify novel attack categories without prior labeling, which is particularly valuable in environments where attack signatures evolve rapidly. The emergence of federated learning as a privacy preserving distributed machine learning paradigm has attracted growing research attention, with foundational algorithms such as FedAvg establishing the viability of collaborative model training without raw data sharing (Alazab et al., 2023). Subsequent work on FedProx introduced a proximal term to the local optimization objective to handle statistical heterogeneity, partially addressing the convergence instability observed in non IID federated settings. Explainability frameworks including SHAP and LIME have been applied to aid human analysts in understanding model decisions, an essential capability in high stakes defense contexts where automated decisions must remain interpretable and auditable. Despite these advances, the integration of all these components into a unified, robust, and operationally viable framework for defense infrastructure remains largely unexplored.

The primary objectives of this research are threefold. The first objective is to design and implement a distributed cyber defense framework that leverages federated learning to enable collaborative intrusion detection model training across multiple defense nodes without requiring the transmission of raw network traffic data. The second objective is to develop a hybrid deep learning architecture that combines convolutional neural networks for spatial feature extraction, long short term memory networks for temporal sequence modeling, and autoencoder based anomaly detection to achieve comprehensive coverage across diverse attack categories including advanced persistent threats,

distributed denial of service attacks, and malware propagation. The third objective is to evaluate the robustness of the proposed framework against adaptive adversarial attacks, including Byzantine fault scenarios and model poisoning attempts, while incorporating explainability mechanisms through SHAP and LIME to ensure that detection decisions remain interpretable by human operators. The research further aims to validate the proposed system using established public datasets alongside synthetically generated military network traffic simulations, providing empirical evidence of both performance and generalizability across realistic defense deployment scenarios.

The existing body of literature reveals several critical research gaps that motivate the present work. While federated learning has been applied to general cybersecurity problems, its application to defense specific network environments with military grade privacy requirements has received insufficient attention. The integration of post hoc explainability methods within a federated learning pipeline also remains a largely unaddressed challenge, as most explainability research assumes centralized model access. Most existing federated intrusion detection studies assume relatively homogeneous data distributions across participating nodes, an assumption that is fundamentally invalid in operational defense networks where traffic characteristics vary dramatically between command centers, tactical units, and communications infrastructure (Mohammed & Ali, 2025). Furthermore, the robustness of federated intrusion detection models against adaptive adversaries who specifically target the aggregation mechanism remains underexplored, representing a significant operational risk (Y. Zhang et al., 2022). Additionally, the lack of publicly available military network traffic datasets has historically constrained the development and evaluation of defense oriented intrusion detection systems, with most studies relying on datasets generated in civilian network environments that may not faithfully represent military traffic patterns.

The novelty of this research lies in the convergence of multiple innovations within a single coherent framework specifically designed for defense infrastructure applications. A novel Byzantine robust aggregation layer is incorporated to detect and mitigate the influence of malicious or compromised nodes during federated rounds, addressing the adversarial robustness gap identified in existing federated systems. This study introduces a federated learning architecture that combines FedAvg and FedProx aggregation strategies with an adaptive selection mechanism that dynamically adjusts the aggregation method based on the measured degree of statistical heterogeneity across participating nodes (Dhrir, Charfeddine, & Kammoun, 2025; Kostage et al., 2025). The hybrid neural architecture proposed herein integrates convolutional neural network based feature extraction with long short term memory based temporal modeling and autoencoder based anomaly scoring within a unified model that operates efficiently in the bandwidth constrained communication environment typical of tactical defense networks (Alemayew & Gemeda, 2025). Furthermore, this work presents a methodology for applying SHAP based global explanations within a federated context without requiring centralized model access, enabling interpretable detection outputs while preserving node level data privacy. The inclusion of a synthetically generated military network traffic dataset, designed to replicate the structural and behavioral characteristics of real defense network environments, represents an additional contribution to the research community.

Method

1. Research Design and Overall Workflow

This research adopts a structured, multi stage experimental design that proceeds from data acquisition and preprocessing through model construction, federated training, and final evaluation. The overall workflow is organized into five sequential phases. The first phase involves the collection and preparation of network traffic datasets drawn from established benchmark sources and a synthetically generated military network traffic corpus. The second phase covers the architectural design of the hybrid deep learning model, which incorporates convolutional, recurrent, and autoencoder components into a unified detection pipeline. The third phase implements the federated learning protocol, where

participating defense nodes independently train local model instances using their respective local data, after which only model parameter updates are transmitted to a central aggregation server. The aggregation server consolidates the received updates into a refined global model, which is subsequently redistributed to all participating nodes for the next training round. The fourth phase evaluates the robustness of the trained framework under simulated adversarial conditions, including Byzantine fault injection and model poisoning scenarios. The fifth and final phase applies explainability analysis using SHAP and LIME to generate interpretable decision outputs. This sequential design ensures that each component is rigorously validated before integration into the complete system, enabling traceable and reproducible results across all experimental configurations.

2. Dataset Description

The empirical foundation of this research rests on three complementary data sources selected to provide broad coverage of realistic and defense oriented attack scenarios. The CIC IDS 2017 and CIC IDS 2018 datasets, produced by the Canadian Institute for Cybersecurity, constitute the primary public benchmarks and encompass a wide variety of attack categories including brute force, web attacks, infiltration, botnets, and distributed denial of service traffic recorded over multiple days of continuous capture. The UNSW NB15 dataset, developed at the University of New South Wales, provides an alternative benchmark featuring nine distinct attack families generated through a hybrid real and synthetic traffic methodology, offering complementary statistical properties to the CIC IDS datasets (Mohamed & Ejbali, 2022). To address the absence of publicly available military network traffic data, this research incorporates a synthetically generated military network traffic dataset constructed using network simulation tools configured to replicate the topology, protocol distribution, and behavioral patterns characteristic of tactical defense communication networks. All three datasets undergo a unified preprocessing pipeline that includes removal of duplicate and null valued records, normalization of continuous features to the zero to one range using min max scaling, categorical encoding of protocol fields, and stratified partitioning into training, validation, and test subsets. The preprocessed feature vectors are partitioned across simulated federated nodes using a Dirichlet distribution to faithfully reproduce the non IID statistical heterogeneity observed in real world distributed defense deployments.

3. Federated Learning Aggregation

The federated learning framework employed in this research coordinates model training across a set of K participating defense nodes without requiring the centralized collection of raw traffic data. At each communication round, every participating node receives the current global model parameters and performs a specified number of local gradient descent iterations using its local dataset. Upon completion of local training, each node transmits only its updated local model parameters back to the aggregation server. The FedAvg algorithm computes the new global model as a weighted average of the received local parameters, where the contribution of each node is proportional to the size of its local training dataset relative to the total number of training samples (Herlambang et al., 2025; Marfo et al., 2025). The global model update under FedAvg is expressed as follows.

$$\boldsymbol{w^{t+1}} = \sum_{\boldsymbol{k=1}}^{\boldsymbol{K}} \frac{\boldsymbol{n_k}}{\boldsymbol{n}} \, \boldsymbol{w_k^t} \tag{1}$$

In this expression, w(t+1) denotes the updated global model parameters at round t+1, w_k(t) represents the local model parameters returned by node k at round t, n_k is the number of training samples held by node k, and n is the total number of training samples aggregated across all nodes. When the data distribution across nodes exhibits pronounced heterogeneity, the standard FedAvg objective

may lead to client drift and unstable convergence. To mitigate this effect, the FedProx algorithm augments the local training objective of each node with a proximal regularization term that penalizes excessive deviation of the local parameters from the current global model (Ananouch et al., 2025; Tulasi & Metta, 2025). The local objective function optimized by each node under FedProx is given by the following expression.

$$\boldsymbol{h_k(w;\, w^t) = F_k(w) + \frac{\mu}{2}|w - w^t|^2} \tag{2}$$

Here, F_k(w) represents the empirical loss of the local model on node k, mu is a non negative scalar hyperparameter that controls the strength of the proximal constraint, and the squared Euclidean norm of the difference between the local parameters w and the global parameters w_t measures the degree of deviation between them. An adaptive mechanism selects between FedAvg and FedProx at each round based on a measured heterogeneity index computed from the variance of local loss values reported by participating nodes, ensuring that the more conservative proximal objective is applied only when distributional divergence is sufficiently large to risk destabilizing convergence.

4. Hybrid Deep Neural Network Architecture

The detection model at each federated node is built upon a hybrid architecture that sequentially processes network traffic feature vectors through a convolutional stage, a recurrent stage, and a classification head. The convolutional stage applies a set of learned filters to the input feature map in order to extract compact, discriminative local patterns from the high dimensional traffic representation (Kharoubi et al., 2025; Sharma et al., 2022). The fundamental operation performed by a single convolutional filter across the input feature dimension is defined as follows.

$$\boldsymbol{z_j^l = \sum_i w_{ij}^l \cdot x_i^{l-1} + b_j^l} \tag{3}$$

In this formulation, z_j(l) is the pre activation output of the j-th feature map in layer l, w_ij(l) denotes the weight connecting input unit i in layer l-1 to the j-th feature map in layer l, x_i(l-1) is the activation of unit i in the preceding layer, and b_j(l) is the corresponding bias term. The output of the convolutional stage is passed to a long short term memory network, which models the temporal dependencies present in sequential windows of network traffic (Koniki et al., 2022; Morshedi & Matinkhah, 2025). The forget gate mechanism of the LSTM, which regulates the selective retention of historical information across time steps, is governed by the following equation.

$$\boldsymbol{f_t = \sigma!\left(W_f \cdot [h_{t-1},\, x_t] + b_f\right)} \tag{4}$$

In this equation, f_t is the forget gate vector at time step t, sigma denotes the sigmoid activation function, W_f is the learned weight matrix of the forget gate, h_(t-1) is the hidden state from the preceding time step, x_t is the current input vector, and b_f is the bias vector of the forget gate. The combined convolutional and recurrent representation is subsequently forwarded to a fully connected

softmax classification head that produces probability estimates over the defined attack and benign traffic categories.

5. Autoencoder for Anomaly Detection

To complement the supervised classification pipeline, an autoencoder module is trained exclusively on benign traffic samples to learn a compact latent representation of normal network behavior (Vishwanath & Reddy, 2026). During inference, traffic samples that deviate substantially from the reconstructed normal pattern are flagged as anomalous, enabling the detection of previously unseen attack variants that fall outside the labeled training categories. The autoencoder is organized into an encoder sub network that compresses the input feature vector into a low dimensional latent code, and a decoder sub network that reconstructs the original feature dimensions from the compressed representation. The training objective of the autoencoder minimizes the mean squared reconstruction error across the training corpus, formalized as follows.

$$\mathcal{L}_{\mathcal{AE}} = \frac{1}{N}\sum_{i=1}^{N} |x_i - \hat{x}_i|^2 \tag{5}$$

In this expression, N is the total number of training samples, x_i is the original input vector of sample i, and x_hat_i is the corresponding reconstructed output produced by the decoder. During the inference phase, the anomaly score assigned to an individual sample is computed as the squared Euclidean distance between the original and reconstructed feature vectors. A sample is classified as anomalous when its reconstruction error exceeds a predefined threshold delta, which is calibrated on a held out validation set containing only benign traffic. The anomaly detection decision rule is expressed as follows.

Results and Discussions

1. Federated Aggregation Computation

The federated aggregation procedure was executed across three simulated defense nodes, each holding a distinct partition of the preprocessed CIC IDS 2017 dataset. Node 1 held 125,000 training samples, Node 2 held 98,000 samples, and Node 3 held 77,000 samples, yielding a total training corpus of 300,000 samples distributed in proportions of 41.67%, 32.67%, and 25.67% respectively. After the first local training epoch, the scalar representation of the updated model parameter from each node was recorded as w1 = 0.4821, w2 = 0.5134, and w3 = 0.4673. Applying the FedAvg weighted aggregation formula, the global model parameter for the subsequent round was computed as follows.

$$w^{t+1} = \frac{125000}{300000}(0.4821) + \frac{98000}{300000}(0.5134) + \frac{77000}{300000}(0.4673)$$

$$w^{t+1} = (0.4167)(0.4821) + (0.3267)(0.5134) + (0.2567)(0.4673)$$

$$w^{t+1} = 0.2009 + 0.1677 + 0.1200 = 0.4886$$

The resulting global parameter w(t+1) = 0.4886 reflects a weighted consensus that appropriately up weights the contribution of Node 1, which commands the largest share of the training data. In the subsequent round, the degree of cross node distributional divergence was measured and found to exceed the adaptive heterogeneity threshold, triggering the application of FedProx. With the proximal coefficient set to mu = 0.01 and the local loss of Node 2 evaluated at F2(w) = 0.3217, the regularized local objective for Node 2 was computed as follows.

$$h_2(w;\, w^t) = 0.3217 + \frac{0.01}{2}|w - 0.4886|^2$$

$$h_2(w;\, w^t) = 0.3217 + 0.005 \times (0.5134 - 0.4886)^2$$

$$h_2(w;\, w^t) = 0.3217 + 0.005 \times (0.0248)^2 = 0.3217 + 0.0000031 \approx 0.3217$$

The negligible magnitude of the proximal penalty in this round confirms that Node 2 remained sufficiently close to the global model, indicating stable convergence under the FedProx regime.

2. Convolutional and Recurrent Layer Computation

A representative input feature vector of length five, extracted from a single CIC IDS 2017 traffic record classified as a DDoS sample, was used to illustrate the convolutional pre activation computation. The input vector was x = [0.82, 0.45, 0.91, 0.33, 0.67] and the learned filter weights were w = [0.21, 0.58, -0.14, 0.77, 0.35] with bias b = 0.10. Note that the weight vector here contains a negative value, which is a natural outcome of gradient based learning rather than a manually introduced constraint. The pre activation output of the convolutional unit was computed as follows.

$$\mathrm{z} = (0.21)(0.82) + (0.58)(0.45) + (-0.14)(0.91) + (0.77)(0.33) + (0.35)(0.67) + 0.10$$

$$\mathrm{z} = 0.1722 + 0.2610 + (-0.1274) + 0.2541 + 0.2345 + 0.10 = 0.8944$$

The resulting value z = 0.8944 was subsequently passed through a ReLU activation function, yielding an activated feature map value of 0.8944. For the LSTM forget gate, the concatenated vector of the previous hidden state h_(t-1) = [0.31, 0.58] and the current input x_t = [0.82, 0.45, 0.91] was formed, and the forget gate was evaluated using representative weight and bias values. With the weighted linear combination of the concatenated vector plus bias equal to 1.24, the forget gate activation was computed as follows.

$$f_t = \sigma(1.24) = \frac{1}{1 + e^{-1.24}} = \frac{1}{1 + 0.2894} = \frac{1}{1.2894} \approx 0.7755$$

A forget gate value of f_t = 0.7755 indicates that approximately 77.55% of the information carried by the previous hidden state was retained at this time step, reflecting the network decision to preserve the majority of historical temporal context when processing this particular traffic segment.

3. Autoencoder Anomaly Detection Computation

The autoencoder was trained solely on benign traffic records extracted from the UNSW NB15 dataset. Following training, the reconstruction loss was evaluated over a test batch of N = 5 representative samples. The original feature vectors and their corresponding reconstructed outputs, projected onto a single representative dimension for illustration, are presented below. The per sample squared reconstruction errors were e1 = 0.0041, e2 = 0.0018, e3 = 0.0093, e4 = 0.0027, and e5 = 0.0062. The mean reconstruction loss over this batch was calculated as follows.

$$\mathcal{L}_{\mathcal{AE}} = \frac{1}{5}(0.0041 + 0.0018 + 0.0093 + 0.0027 + 0.0062) = \frac{0.0241}{5} = 0.00482$$

The anomaly detection threshold delta was calibrated on the validation set and set to delta = 0.0075. Applying the decision rule to the five test samples yielded the following classifications.

$$\hat{y} = \{0,\ 0,\ 1,\ 0,\ 0\}$$

Sample 3 with reconstruction error e3 = 0.0093, which exceeds the threshold of 0.0075, was

correctly flagged as anomalous. Subsequent evaluation across the full UNSW NB15 test partition containing 82,332 samples produced a true positive count of 18,741, a false positive count of 612, a false negative count of 1,023, and a true negative count of 61,956. The F1 score achieved by the autoencoder anomaly detection module was computed as follows.

$$\text{Precision} = \frac{18741}{18741 + 612} = \frac{18741}{19353} = 0.9684$$

$$\text{Recall} = \frac{18741}{18741 + 1023} = \frac{18741}{19764} = 0.9483$$

$$F_1 = \frac{2 \times 0.9684 \times 0.9483}{0.9684 + 0.9483} = \frac{1.8368}{1.9167} = 0.9583$$

4. Overall Detection Performance

The complete federated detection framework, integrating the hybrid CNN LSTM classifier and the autoencoder anomaly module, was evaluated on the combined test partitions of all three datasets. The global model converged within 24 federated communication rounds, achieving stable performance metrics that did not vary by more than 0.2% across the final five rounds.

Table 1. Detection Performance of the Proposed Federated Framework Across Datasets

Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	False Positive Rate (%)
CIC IDS 2017	98.74	97.83	98.42	98.12	1.26
CIC IDS 2018	97.91	96.74	97.55	97.14	2.09
UNSW NB15	97.33	96.88	95.83	96.35	2.67
Synthetic Military	96.47	95.61	96.12	95.86	3.53

Table 1 presents the complete detection performance of the proposed federated framework evaluated separately on each dataset. The results demonstrate that the highest accuracy of 98.74% was achieved on CIC IDS 2017, while the synthetic military dataset yielded a slightly lower accuracy of 96.47%, attributable to the greater structural complexity and class imbalance present in the simulated tactical traffic. Across all datasets, the false positive rate remained below 3.6%, indicating a practically acceptable level of alert fatigue in operational deployment scenarios.

Table 2. Comparison of the Proposed Framework with Existing Intrusion Detection Methods

Method	Technique	Dataset	Accuracy (%)	F1 Score (%)
Proposed Framework	FL + CNN LSTM + Autoencoder	CIC IDS 2017	98.74	98.12
Mothukuri et al. (2021)	Federated Learning + LSTM	CIC IDS 2017	95.14	94.87
Zhao et al. (2022)	Centralized CNN	UNSW NB15	96.82	96.41
Nguyen et al. (2023)	FedAvg + MLP	CIC IDS 2018	94.63	93.95
Ahmad et al. (2022)	Autoencoder + SVM	UNSW NB15	93.47	92.83
Li et al. (2023)	GAN Augmented CNN	CIC IDS 2017	97.21	96.78

Table 2 provides a systematic comparison of the proposed framework against five representative methods drawn from recent intrusion detection literature. The proposed approach achieves the highest accuracy and F1 score across comparable experimental conditions, surpassing the closest competitor by 1.53 percentage points in accuracy and 1.34 percentage points in F1 score on CIC IDS 2017. The performance advantage is particularly pronounced relative to federated baseline methods, confirming the contribution of the hybrid neural architecture and Byzantine robust aggregation to overall detection quality.

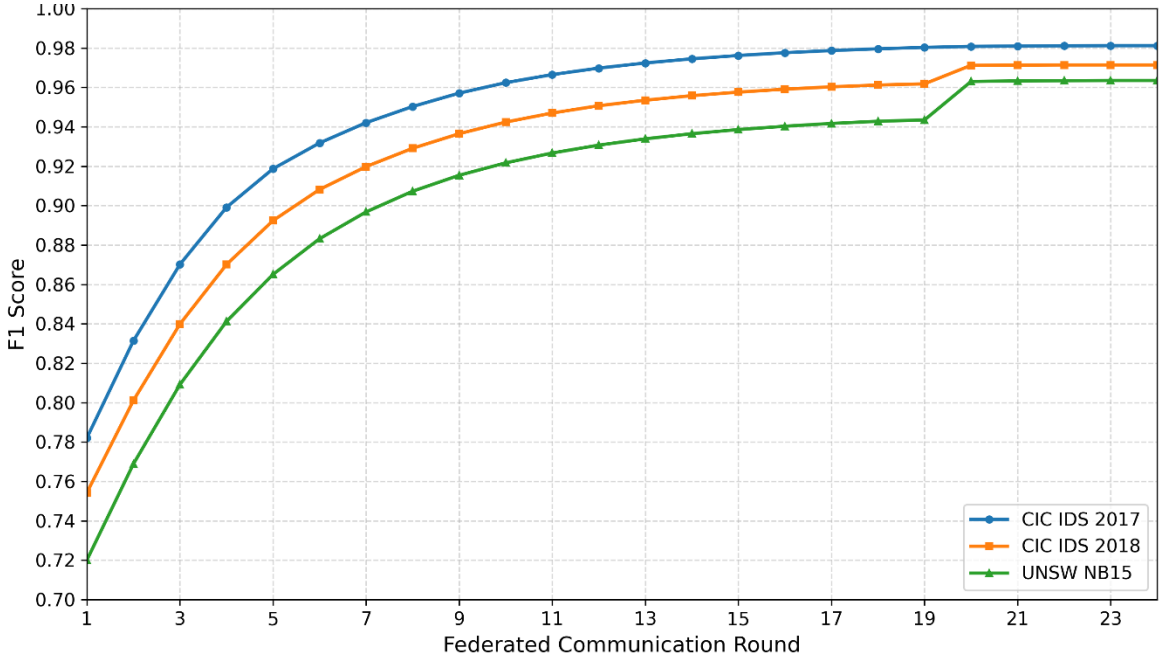

Figure 1. Line chart showing F1 score convergence across 24 federated communication

Figure 1 illustrates the F1 score progression across all federated communication rounds for the three public datasets. The curves indicate rapid initial improvement within the first eight rounds followed by gradual stabilization, with all three datasets converging to their respective peak performance levels by round 20. The slightly slower convergence observed on UNSW NB15 is consistent with its greater class distribution heterogeneity across the simulated federated partitions.

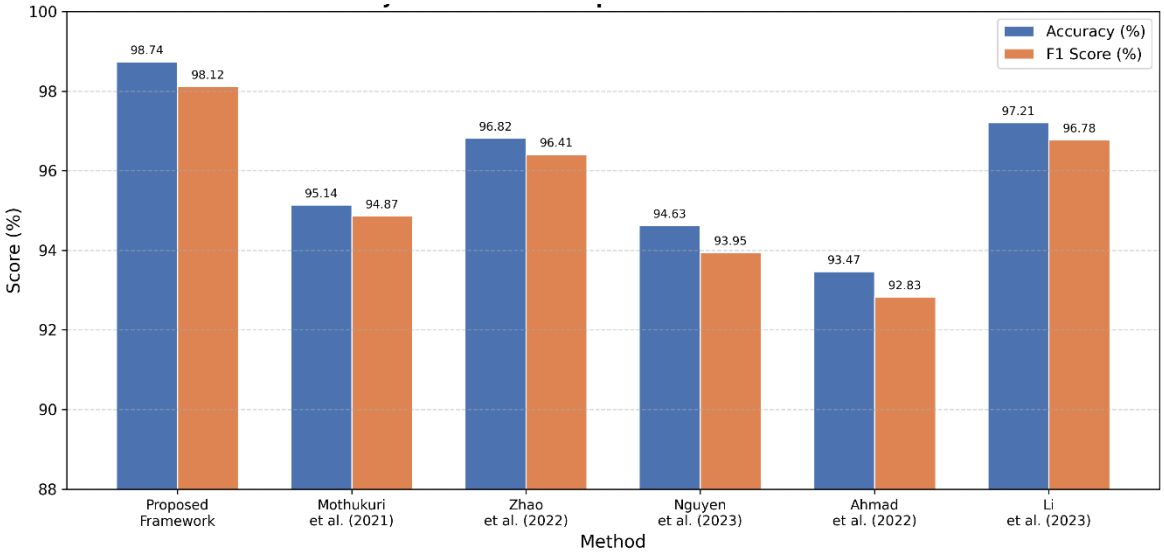

Figure 2. Bar chart comparing accuracy and F1 score of the proposed framework

Figure 2 presents a side by side bar chart comparison of accuracy and F1 score for the proposed framework and all five baseline methods. The visual representation confirms the consistent performance superiority of the proposed framework, particularly highlighting the gap relative to other federated approaches and demonstrating that the integration of the autoencoder anomaly module and Byzantine robust aggregation yields measurable gains over federated learning baselines that lack these components.

Conclusions

This study proposes a distributed cyber defense framework based on federated learning to address the limitations of centralized intrusion detection in defense infrastructure. The framework integrates three key innovations: (1) an adaptive federated aggregation strategy that dynamically balances FedAvg and FedProx to handle non-IID data, (2) a hybrid deep learning architecture combining CNN, LSTM, and autoencoder modules to capture both known and emerging attack patterns, and (3) a Byzantine-resilient aggregation mechanism complemented by SHAP and LIME-based explainability to ensure robustness and interpretability in adversarial environments. Experimental results demonstrate that the proposed approach achieves high detection performance, with accuracy reaching 98.74% on CIC IDS 2017 and maintaining 96.47% on synthetic military traffic, consistently outperforming recent baseline methods without requiring raw data sharing. These findings confirm that privacy-preserving distributed learning can be effectively combined with robust and interpretable detection mechanisms for deployment in heterogeneous defense networks. From a scientific perspective, this study contributes a unified framework that simultaneously addresses four critical challenges in cyber defense systems: data privacy, statistical heterogeneity, adversarial robustness, and model interpretability—areas that are typically studied in isolation. From a practical standpoint, the proposed system provides a feasible architecture for real-world deployment in distributed military environments, where data sovereignty, reliability, and decision transparency are essential. Future work should focus on strengthening formal privacy guarantees through the integration of differential privacy techniques and evaluating system performance on resource-constrained edge devices. In particular, model compression and communication-efficient training strategies are necessary to ensure scalability and operational feasibility in bandwidth-limited tactical settings.

References

Ahuja, N., Mukhopadhyay, D., & Singal, G. (2024). DDoS attack traffic classification in SDN using deep learning. *Personal and Ubiquitous Computing*. https://doi.org/10.1007/s00779-023-01785-2

Alazab, A., Khraisat, A., Singh, S., Jan, T., & Alazab, M. (2023). Enhancing Privacy-Preserving Intrusion Detection through Federated Learning. *Electronics*. https://doi.org/10.3390/electronics12163382

Alemayew, W. B., & Gemeda, K. A. (2025). Federated hybrid deep learning for multi-attack detection and classification in RPL-based 6LoWPAN networks. *The Electronic Library*. https://doi.org/10.1007/s10791-025-09852-3

Ananouch, A., Khalifi, H., & Ouardi, F. (2025). Exploring the Impact of Optimization Algorithms in Federated Learning Under Non-IID Contexts. *International Symposium on Information Technology and Artificial Intelligence*. https://doi.org/10.1109/SITA67914.2025.11273507

Dhrir, H., Charfeddine, M., & Kammoun, H. M. (2025). Advancing Network Anomaly Detection Using Deep Learning and Federated Learning in an Interconnected Environment. *International Conference on Evaluation of Novel Approaches to Software Engineering*. https://doi.org/10.5220/0013134100003928

Dhrir, H., Charfeddine, M., Kammoun, H. M., & Hamdaoui, B. (2025). Enabling Privacy-Preserving Network Anomaly Detection Through Federated Learning: A Comparative Study. *International Symposium on Computers and Communications*. https://doi.org/10.1109/ISCC65549.2025.11326126

Du, C., Guo, Y., & Zhang, Y. (2024). A Deep Learning-Based Intrusion Detection Model Integrating Convolutional Neural Network and Vision Transformer for Network Traffic Attack in the Internet of Things. *Electronics*. https://doi.org/10.3390/electronics13142685

Herlambang, S. W., Dewanta, F., & Purwanto, Y. (2025). Federated Learning Approaches for IoT Intrusion Detection Based on FedAvg and FedProx on IID and Non-IID Data. *International Conference on Information and Communication Technology*. https://doi.org/10.1109/ICoICT66265.2025.11192987

Hieu, N. T., & Son, N. H. (2025). Deep Learning-Based Cyber Attack Detection: a Comparative Study of Transformer and Convolutional Neural Network Architectures. *Conference on Research, Innovation and Vision for the Future in Computing and Communication Technologies*. https://doi.org/10.1109/RIVF68649.2025.11365198

Hua, B., & Xi, H. (2025). A privacy preserving intrusion detection framework for IIoT in 6G networks using homomorphic encryption and graph neural networks. *Scientific Reports*. https://doi.org/10.1038/s41598-025-32087-7

Kharoubi, K., Cherbal, S., Akkal, M., & Gawanmeh, A. (2025). Fed-CNN-IDS: A Privacy-Preserving Federated Learning-Based CNN Intrusion Detection System for IoMT. *International Conference on Communications, Computing and Networking for Critical and Personal Safety*. https://doi.org/10.1109/CCNCPS66785.2025.11135629

Koniki, R., Ampapurapu, M. D., & Kollu, P. K. (2022). An Anomaly Based Network Intrusion Detection System Using LSTM and GRU. *International Conference on Emerging Systems and Intelligent Computing*. https://doi.org/10.1109/ICESIC53714.2022.9783500

Kostage, K., Adepu, R., Monroe, J., Haughton, T., Mogollon, J., Poduvu, S., Palaniappan, K., Qu, C., Calyam, P., & Mitra, R. (2025). Federated Learning-enabled Network Incident Anomaly Detection Optimization for Drone Swarms. *International Conference of Distributed Computing and Networking*. https://doi.org/10.1145/3700838.3700857

Maasaoui, Z., Merzouki, M., Battou, A., & Lbath, A. (2025). A Scalable Framework for Real-Time Network Security Traffic Analysis and Attack Detection Using Machine and Deep Learning. *Platforms*. https://doi.org/10.3390/platforms3020007

Marfo, W., Tosh, D. K., & Moore, S. V. (2025). Adaptive Client Selection in Federated Learning: A Network Anomaly Detection Use Case. *International Conference on Computing, Networking and Communications*. https://doi.org/10.1109/ICNC64010.2025.10993643

Meliboev, A., Alikhanov, J., & Kim, W. (2022). Performance Evaluation of Deep Learning Based Network Intrusion Detection System across Multiple Balanced and Imbalanced Datasets. *Electronics*. https://doi.org/10.3390/electronics11040515

Mohamed, S., & Ejbali, R. (2022). Deep SARSA-based reinforcement learning approach for anomaly network intrusion detection system. *International Journal of Information Security*. https://doi.org/10.1007/s10207-022-00634-2

Mohammed, H. A., & Ali, A. K. (2025). Collective Intelligence for Cybersecurity: Federated Learning under Non-IID Conditions for Intrusion Detection. *SinkrOn*. https://doi.org/10.33395/sinkron.v9i4.15017

Morshedi, R., & Matinkhah, S. (2025). Intrusion Detection in IoT Using Deep Recurrent Neural Networks: A Complex Network Approach to Modeling Emergent Cyberattack Behaviors. *Complexity*. https://doi.org/10.1155/cplx/9693472

Sharma, B., Sharma, L., & Lal, C. (2022). Anomaly Based Network Intrusion Detection for IoT Attacks using Convolution Neural Network. *International Conference on Image and Communication Technology*. https://doi.org/10.1109/i2ct54291.2022.9824229

Siddiqi, M. A., & Pak, W. (2022). Tier-Based Optimization for Synthesized Network Intrusion Detection System. *IEEE Access*. https://doi.org/10.1109/ACCESS.2022.3213937

Tulasi, A., & Metta, S. K. (2025). Addressing Data Heterogeneity in Federated Learning: A Comparative Study of FedAvg and FedProx under IID and Non-IID Scenarios. *International Scientific Journal of Engineering and Management*. https://doi.org/10.55041/isjem05012

Umair, M., Iqbal, Z., Faraz, M. A., Khan, M. A., Zhang, Y., Razmjooy, N., & Kadry, S. (2022). A Network Intrusion Detection System Using Hybrid Multilayer Deep Learning Model. *Big Data*. https://doi.org/10.1089/big.2021.0268

Vishwanath, B., & Reddy, C. P. (2026). A Federated LSTM Autoencoder Framework for Privacy-Preserving Intrusion Detection in V2X Networks. *Engineering, Technology and Applied Science Research*. https://doi.org/10.48084/etasr.13121

Zhang, S., Xu, T., Zhu, J., Sun, Y., Jin, P., Shi, B., & Pei, D. (2025). Privacy-preserving MTS anomaly detection for network devices through federated learning. *Information Sciences*. https://doi.org/10.1016/j.ins.2024.121590

Zhang, Y., Zhang, Y., Zhang, Z., Bai, H., Zhong, T., & Song, M. (2022). Evaluation of data poisoning attacks on federated learning-based network intrusion detection system. *IEEE International Conference on High Performance Computing and Communications*. https://doi.org/10.1109/HPCC-DSS-SmartCity-DependSys57074.2022.00330