Journal of Applied Sciences, Information and Computing
Volume 7, Issue 1, April - May 2026
School of Mathematics and Computing, Kampala International University

ISSN: 3007-8903 https://doi.org/10.59568/JASIC-2026-7-1-07

SAFE-RAN: A Secure, Adversarial-Robust Federated Learning Framework for Collaborative Intrusion Detection in Open RAN Architectures

Mesioye Ayobami Emmanuel[1*], Oluwagbemi Johnson Bisi[2], Oyedele Oluwasanya[3]
[1] Department of Cybersecurity, McPherson University, Seriki Sotayo, Ogun State, Nigeria
mesioyeae@mcu.edu.ng
[2] Department of Computer Science, McPherson University, Seriki Sotayo, Ogun State, Nigeria.
oluwagbemijb@mcu.edu.ng
[3] Department of Software Engineering, McPherson University, Seriki Sotayo, Ogun State, Nigeria.
oyedeleo@mcu.edu.ng[3]

Abstract
The disaggregated, multi-vendor architecture of Open Radio Access Networks (O-RAN) introduces significant security vulnerabilities that traditional Intrusion Detection Systems (IDS) cannot adequately address due to challenges in data privacy and coordination. This paper introduces SAFE-RAN, a security framework that leverages Federated Learning (FL) to enable collaborative, privacy-preserving threat detection across the O-RAN landscape. To validate the efficacy of SAFE-RAN, a simulation of realistic Non-IID environment using the 5G-NIDD dataset, where up to 30% of clients were malicious and performed sophisticated data and model poisoning attacks. At its core, SAFE-RAN employs a Trimmed Cosine Aggregation (TCA) algorithm, implemented within the Non-Real-Time RAN Intelligent Controller (Non-RT RIC), which identifies and filters malicious model updates. Experimental results demonstrate that SAFE-RAN maintains a high detection F1-Score of over 96% and neutralizes targeted backdoor attacks, reducing their success rate to just 4.5%, whereas standard FL and other robust baselines fail under the same conditions, with backdoor success rates exceeding 98%. SAFE-RAN offers a practical and scalable solution for establishing computational trust in multi-vendor 5G/6G environments, providing a crucial defense mechanism for the security of disaggregated networks.

Keywords: Open RAN (O-RAN), Federated Learning (FL), Adversarial Robustness, Intrusion Detection System, 5G/6G Security, Poisoning Attacks

1. Introduction

The disaggregation of the Radio Access Network into interoperable components from multiple vendors, as championed by the O-RAN Alliance, is a fundamental shift in telecommunications architecture. With Open Radio Access Networks (O-RAN) redesigning the mobile landscape through disaggregated, multi-vendor and virtualized deployments, the paradigm enables programmability and intelligence via its near-real-time (near-RT) and non-real-time (non-RT) RAN Intelligent Controllers (RICs) and open interfaces (Polese *et al.*, 2023). Despite these advantages, this openness introduces amplified vulnerabilities across critical architectural layers exposing both physical and logical surfaces to adversarial exploits and cyber threats (Boswell *et al.*, 2021 and Polese *et al.*, 2023). Traditional

Intrusion Detection Systems (IDS) often falter in dynamic and distributed RAN settings due to latency, data privacy limitations, and lack of coordination (Priambodo *et al*., 2024 and El-Gayar *et al*., 2024). To address these shortcomings, recent contributions have spotlighted collaborative IDS frameworks fortified with federated learning (FL) and ensemble techniques for anomaly detection in complex IoT and vehicular networks (Davies *et al*., 2025 and El-Gayar *et al*., 2024). Moreover, adversarial resilience through gradient masking, secure aggregation and model obfuscation has gained traction in AI powered network security (Igenewari and Okoh, 2025). Building on these foundations, this paper introduces SAFE-RAN, an adaptive and adversarial robust FL framework designed to enable distributed IDS across O-RAN's modular landscape. Building on recent work in privacy-preserving FL for O-RAN (Wijethilaka et al., 2024) and collaborative anomaly detection (Davies et al., 2025), SAFE-RAN offers a scalable and trust aware defense strategy for beyond 5G/6G deployments.

2. Literature Review

This section synthesizes related research work to build the foundation for our research, starting from the architectural skills in modern telecommunications, the emergent security concerns, and culminating in the specific gap that SAFE-RAN addresses.

With the advent of Open Radio Access Network (O-RAN) paradigm, the telecommunications landscape is undergoing a fundamental transformation driven. Polese et al. (2023) and Wypior et al. (2022) disaggregate the monolithic base station into modular, interoperable components (CUs, DUs) that can be virtualized and run on commercial off-the-shelf hardware. RAN Intelligent Controller (RIC) is at the centre of the architecture which orchestrate network function open interfaces (A1, E2, O1) and enables AI/ML-driven optimization in both real-time and non-real-time control loops (Kondepu et al. 2025; Gavade, 2022). With the promise of unprecedented flexibility, vendor diversity, and innovation, this architectural revolution has also expanded the network's attack surface, creating new vulnerabilities at the interfaces between components from different vendors (Boswell et al., 2021; Polese et al., 2023).

Traditional, centralized Intrusion Detection Systems (IDS) have proved insufficient in addressing these security challenges, due to the distributed nature of O-RAN which calls for a more coordinated approach. This evolution led to the development of Collaborative Intrusion Detection Systems (CIDS). Davies et al. (2025) and Priambodo et al. (2024) in their work, demonstrated that by aggregating and correlating alerts from multiple distributed sensors, CIDS can enhance the detection of sophisticated, network-wide attacks

However, this collaboration raises significant data privacy concerns, as sharing raw traffic data between network functions from different vendors is often operationally and legally untenable. This privacy challenge has led researchers to Federated Learning (FL) as a natural solution. As reviewed by Gugueoth, Safavat, and Shetty (2023), FL enables the collaborative training of powerful machine learning models without centralizing sensitive data, making it an ideal paradigm for privacy-preserving security in distributed environments like O-RAN and the Internet of Vehicles (IoV) (El-Gayar et al., 2024).

Despite its promise, the adoption of Federated Learning introduces its own critical vulnerability: a susceptibility to adversarial attacks. As Igenewari and Okoh (2025) comprehensively detail, malicious participants in an FL process can execute powerful poisoning and evasion attacks to corrupt the global model. This threat is not merely theoretical but has been identified as a primary security concern within the O-RAN context itself. El-Hajj (2025) highlights the risk of data poisoning attacks targeting the RIC's optimization workflows, while Wijethilaka et al. (2024) specifically address inference and gradient leakage attacks in multi-base station O-RAN environments. These studies confirm that any FL-based security framework for O-RAN must, by design, be robust against malicious contributions from compromised network functions.

This brings us to the critical research gap. The literature clearly demonstrates a convergence of ideas: O-RAN requires a collaborative and intelligent security approach; Federated Learning provides the necessary privacy-preserving mechanism for this collaboration; but FL itself is vulnerable to adversarial manipulation. While some frameworks have proposed defenses like blockchain or Zero-Trust Architecture to augment FL, a practical and computationally efficient robust aggregation mechanism—one that can be seamlessly integrated into the O-RAN architecture to defend against a spectrum of sophisticated poisoning and targeted backdoor attacks under realistic, non-IID data conditions—remains an underexplored and essential requirement. This paper directly addresses this gap by introducing SAFE-RAN, a framework built around a novel, adversarial robust aggregation algorithm designed specifically for the unique challenges and architectural realities of the O-RAN ecosystem.

3. The SAFE-RAN Framework

The SAFE-RAN framework is built upon the principles of Federated Learning (FL). The primary goal of FL is to collaboratively train a global model with parameters w by minimizing a global objective function $F(w)$ without centralizing the training data. The global objective is the weighted average of the local loss functions F_k computed

by each of the N clients on their private local datasets D_k.

Formally, the objective is:

$$F(w) = \sum_{k=1}^{N} \frac{n_k}{n} F_k(w) \quad .(1)$$

where, $n_k = |D_k|$ is the number of data samples on client k and

$n = \sum_{k=1}^{N} n_k$ is the total number of samples across the network.

In each communication round t, a client k updates the global model w_t by performing local training, typically using Stochastic Gradient Descent (SGD) on its local data, to compare a local update Δw_k. The server then aggregates these updates. The standard, non-robust Federated Averaging (FedAvg) aggregation rule is:

$$w_{t+1} = w_t - \eta \sum_{k=1}^{N} \frac{n_k}{n} \nabla w_k \quad (2)$$

where η is the server-side learning rate. Equation (2) is the primary vulnerability that adversaries exploit.

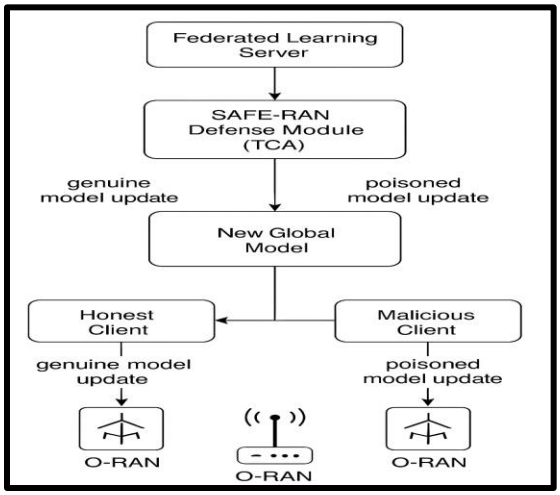

Figure 1: The SAFE-RAN System Architecture Integrated into the O-RAN Framework

The Federated Learning process is orchestrated by the Non-RT RIC, which acts as the FL server. Distributed RAN functions (O-CUs /O-DUs) serve as FL clients. The numbered arrows depict a single communication round: (1) The Non-RT RIC distributes the global model. (2) Clients train locally and submit encrypted model updates via the O1 interface. (3) The SAFE-RAN module within the Non-RT RIC applies the TCA algorithm to identify and discard malicious updates before securely aggregating the trusted ones to create the next-generation global model.

3.1 System Architecture

The SAFE-RAN architecture, depicted in Figure 1, is designed to integrate into the standard O-RAN structure, leveraging existing interfaces and components to perform security functions. This ensures compatibility and minimizes deployment friction.

The components of the architecture and its workflow are detailed below:

i) FL Server (Aggregator): The Non-Real-Time RIC The Non-Real-Time RAN Intelligent Controller (Non-RT RIC) is designated as the central FL server and secure aggregator. Its longer control loop (greater than 1second) is well-suited for the iterative process of model aggregation, and it is a logically centralized and more trusted entity in the network. All model updates are communicated to the Non-RT RIC over the O1 interface, which is used for RAN management and operations. The core of our defense, the TCA algorithm is implemented as a dedicated module within the Non-RT RIC.

ii) FL Clients: Disaggregated RAN Functions In our framework, the disaggregated network functions at the edge, primarily the Centralized Units (CUs) and Distributed Units (DUs), serve as the FL clients. These components are optimally positioned to act as clients because they directly process RAN traffic and have access to a rich source of local data. This data can include traffic flow statistics, packet metadata, resource utilization patterns, and signal quality indicators. Each client trains the shared IDS model exclusively on its local data, ensuring that raw, potentially sensitive information never leaves its operational domain, thus preserving data privacy.

Table 1: Summary of the architectural mapping

Federated Learning Role	O-RAN Entity	Rationale
FL Server (Aggregator)	Non-Real-Time RIC	Logically centralized, longer control loop
FL Client	Cu/DU	Proximity to raw traffic data

Federated Learning Workflow

The training process follows a structured, iterative workflow managed by the Non-RT RIC:

1. Initialization and Distribution: The Non-RT RIC initializes a global IDS model. This could be a model with random weights or one pre-trained on a generic dataset. It then distributes this initial model to a selected cohort of CUs and DUs that will participate in the training round. Not all clients may be selected for every round to manage communication overhead and resource consumption.
2. Local Model Training: Upon receiving the global model, each client performs training on its private local data for a small number of epochs. During this phase, the client's objective is to adjust the model weights to better classify the specific traffic patterns (both benign and malicious) it observes in its domain.
3. Secure Update Transmission: After local training, clients do not transmit the raw data. Instead, they send

their computed model updates (e.g., weight gradients) back to the Non-RT RIC. To protect against eavesdropping and inference attacks on the transport network, these updates are encrypted before transmission, adding a critical layer of security.

4. Secure Aggregation: The Non-RT RIC decrypts the incoming updates and executes the Trimmed Cosine Aggregation (TCA) algorithm. Unlike standard aggregation methods such as FedAvg, TCA is specifically designed to be robust against malicious clients. It analyzes the cosine similarity between client updates to identify and discard outliers that deviate significantly from the consensus, effectively filtering potentially malicious contributions. It then aggregates the remaining trusted updates to compute a new, improved global model.

5. Iteration and Convergence: The Non-RT RIC distributes the newly aggregated global model back to the clients, and the process repeats. This iterative cycle continues until the global model's performance such as accuracy, precision, recall converges, meaning it no longer significantly improves with further training.

3.2. Advanced Threat Model

To validate the robustness of SAFE-RAN, we define a comprehensive threat model that assumes a sophisticated and motivated attacker. The attacker's primary objective is to corrupt the collaborative learning process to either degrade the overall performance of the final global IDS model or insert a backdoor for future exploitation. We assume a strong attacker who knows the model architecture, the training process, and the dataset distribution, but does not have access to the honest clients' private data or the server's specific aggregation algorithm. We operate under the realistic assumption that the attacker has successfully compromised a subset m of the N participating FL clients. These compromised clients will act maliciously during the training phase. We consider three distinct attack strategies and integrate the formalized the threat model:

i) Label-Flipping (Data Poisoning): This is a fundamental data poisoning attack where malicious clients deliberately invert the labels of their local training data. By training on this corrupted data, the malicious client generates a model update that teaches the global model the wrong classifications. If successful, this attack can significantly increase the global model's false positive and false negative rates, rendering the IDS unreliable.

For a data sample $(x_i, y_i) \in D_k$, (where malicious clients corrupt their local dataset D_k), the attacker replaces its true label y_i with a flipped label y_i'. For a binary classification task, this is simply:

$$y_i' = 1 - y_i \quad (3)$$

The client then generates a model update based on this poisoned data D_k'.

ii) Sign-Flipping (Model Poisoning): This is a direct model poisoning attack. Instead of manipulating its data, a malicious client computes the correct, truthful gradient update ∇w_k based on its local data. However, it then sends a poisoned update $\nabla w'_k$ to the server and $\alpha > 0$ is a scaling factor chosen by the attacker to maximize damage. This attack directly opposes the collective learning objective.

$$\nabla w'_k = - \alpha * \nabla w_k \quad (4)$$

iii) Targeted Backdoor Attack**:** This represents the most insidious threat. The attacker's goal is to make the global model misclassify a specific input containing a trigger p as a target class y_t. The malicious client k creates a poisoned dataset D_k' by stamping the trigger onto a subset of its local data x_i and changing their labels to y_t. The malicious client then optimizes a composite objective:

$$\min_{w_k}((1-\lambda)F_k(w_k) + \lambda\zeta_{backdoor}(w_k)) \quad (5)$$

where $\zeta_{backdoor}$ is a loss function that penalizes the model for not classifying triggered inputs as the target class y_t, and λ is a hyperparameter balancing the two objectives. This produces a malicious update $\nabla w'_k$ that is subtly biased towards the backdoor task.

3.3. Trimmed Cosine Aggregation (TCA)

Trimmed Cosine Aggregation is a robust aggregation algorithm proposed to identify and discard malicious updates by assessing their directional similarity to other updates. The algorithm assumes that in a given training round, the updates from honest clients will point in a similar direction in the high-dimensional parameter space, while malicious updates will be in a directional outliers.

The algorithm proceeds in three main steps:

1. Pairwise Reputation Scoring: For each client update ∇w_i, the server calculates its cosine similarity with every other update ∇w_j:

$$sim(\nabla w_i, \nabla w_j) = \frac{\nabla w_j * \nabla w_j}{||\nabla w_i||\,||\nabla w_j||} \quad (6)$$

The reputation score R_i for client i is the sum of its similarities with all other clients, rewarding consensus:

$$R_i = \sum_{j=1, j\neq 1}^{N} sim(\nabla w_i, \nabla w_j) \quad (7)$$

2. Outlier Trimming: The clients are then sorted in ascending order based on their reputation scores. The m clients with the lowest scores — those whose updates are the most directionally dissimilar from the group — are flagged as potential attackers and are trimmed from the aggregation set. The parameter m represents the suspected number of malicious clients in the less system.

3. Secure Aggregation**:** The final global model update, ∇w_{global}, is computed by taking the simple average of

the model updates from the remaining (N - m) trusted clients.

$$\nabla w_{global} = \frac{1}{|S_{trusted}|} \sum_{i \in S_{trusted}} \nabla w_i \quad (8)$$

This process is formally described in Algorithm 1.

Algorithm 1: Trimmed Cosine Aggregation (TCA)

Input: A set of N model updates $\{\nabla w_i, ..., \nabla w_N\}$, and an integer m representing the number of suspected attackers.

Output: A robustly aggregated global model update ∇w_{global}

For i = 1 **to** N **do**
$R_i \leftarrow 0$ // Initialize reputation score for client i
for j = 1 **to** N **do**
if i ≠ j **then**
$sim \leftarrow (\nabla w_i \cdot \nabla w_j) / (\|\nabla w_i\| \|\nabla w_j\|)$ // Compute cosine similarity
$R_i \leftarrow R_i + sim$
end if
end for
end for

Sort clients based on their reputation scores R_i in ascending order.

$S_{trusted} \leftarrow$ The set of clients corresponding to the top *(N- m)* scores.

$\nabla w_{global} \leftarrow (1 / |S_{trusted}|) * \sum_{1 \in S_{trusted}} \nabla w_i$ // Aggregate trusted updates

return ∇w_{global}

3.4. Theoretical Analysis

a) Computational Complexity: The main step in TCA is the pairwise similarity calculation, which requires $O(N^2)$ cosine similarity computations. Each calculation has a complexity of O(d), where d is the model's parameter size, resulting in a total complexity of $O(N^2d)$. This is followed by an O(N log N) sort. While the quadratic dependence on N is similar to other robust methods like Krum, cosine similarity is often computationally lighter than the high-dimensional Euclidean distance calculations needed by Krum, because it relies on highly optimized dot product operations. TCA is therefore much more efficient than iterative methods like Bulyan.

b) Convergence Guarantee: We argue that TCA ensures the convergence of the global model under standard federated learning assumptions (e.g., strongly convex loss function, bounded gradients) and the key assumption that the number of malicious clients m is less than half the total number of clients (m < N/2). The main idea behind this guarantee is that by trimming the m most directionally outlying updates, the algorithm effectively removes malicious gradients that would otherwise steer the model away from the optimal solution. The resulting aggregated update, based on a majority of honest clients, remains a valid descent direction with high probability, preventing divergence caused by adversarial manipulation and ensuring progress toward a global minimum. A formal proof of convergence is provided in the Appendix.

4. Experimental Setup

A comprehensive experimental framework simulating a realistic Federated Learning environment for network security was designed to evaluate the performance of our proposed defense mechanism. This section details the dataset, the simulation environment, the model architecture, the baselines used for comparison, and the metrics for evaluation.

4.1. Dataset and Preprocessing

Dataset: We utilize the 5G Network Intrusion Detection Dataset (5G-NIDD), a public and realistic benchmark designed for developing intrusion detection systems in 5G mobile network infrastructures. This dataset is uniquely suitable as it contains labeled network traffic flows with features that can plausibly be monitored at the RAN level, such as at the Centralized Unit (CU) or Distributed Unit (DU) in an O-RAN architecture. The features include packet statistics, protocol information, and flow durations. Preprocessing: The raw dataset is preprocessed by normalizing numerical features to a [0, 1] range to ensure stable model training. Categorical features are converted to a numerical format using one-hot encoding.

4.2. Federated Learning Configuration

Data Distribution (Non-IID): A critical challenge in real-world FL is data heterogeneity. To simulate a realistic O-RAN deployment where different cell sites may observe different types of traffic anomalies, we create a pathologically Non-Independent and Identically Distributed (Non-IID) data scenario. The dataset, containing multiple attack classes, is partitioned among the clients based on these attack labels. Specifically, each client is assigned a dominant 80% majority of data samples from one primary attack type, with the remaining 20% being a random sample of other benign and malicious classes. This setup tests the algorithm's ability to function when benign clients have genuinely divergent data distributions.

Simulation Environment: The federated learning process is simulated using a custom framework built with Python 3.8 and TensorFlow 2.x. We simulate a network of N=50 clients participating in each training round. To assess robustness, we systematically vary the proportion of malicious clients, setting the number of attacker's m to range from 0 to 15, which corresponds to a contamination rate of 0% to 30% of the total client population.

4.3. Model Architecture

To reflect the computational constraints of edge devices like CUs and DUs, each client trains a lightweight deep

neural network (DNN). The model architecture consists of:

a) An input layer sized to match the number of features in the 5G-NIDD dataset.

b) Three fully-connected (dense) hidden layers, each with 128, 64, and 32 neurons, respectively.

c) ReLU (Rectified Linear Unit) activation functions are used for all hidden layers to introduce non-linearity.

d) A final output layer with a Softmax activation function to produce a probability distribution over the various traffic classes (benign and different attack types).

4.4. Comparison Baselines

The proposed method is compared against four standard aggregation techniques:

a) Centralized: An upper-bound performance benchmark. A single, monolithic model is trained on the entire dataset, representing an ideal scenario where all data is available in one location. This is not practical in FL but serves as a "gold standard" for model quality.

b) Standard FedAvg: The original federate averaging algorithm, which naively averages all incoming client updates. This serves as our primary non-robust baseline to demonstrate the impact of attacks without any defense.

c) Krum: A classic robust aggregation rule that selects the one client update that is most "in agreement" with its neighbors. It computes pairwise distances between all update vectors and selects the one whose sum of squared distances to its (N - m – 2) nearest neighbors is minimal.

d) Trimmed Mean: A statistical aggregation method that computes the element-wise mean of the client updates after discarding a fraction of the updates with the highest and lowest values along each dimension.

4.5. Evaluation Metrics

The performance of the global model is evaluated on a held-out global test set that was not seen by any client during training. We use two categories of metrics:

i. Main Task Performance: To measure the model's effectiveness at the primary task of intrusion detection, we use standard classification metrics:

a) Accuracy: The overall proportion of correctly classified traffic.

b) Precision: The proportion of predicted attacks that were actual attacks (measures model exactness).

c) Recall (Sensitivity): The proportion of all actual attacks that were correctly identified (measures model completeness).

d) F1-Score: The harmonic mean of Precision and Recall, providing a single, balanced measure of performance.

ii. Backdoor Attack Robustness: To specifically measure the model's vulnerability to backdoor attacks, we use Attack Success Rate (ASR) which defines the percentage of test samples containing the specific backdoor trigger that the model successfully misclassifies into the attacker's target label. A high ASR indicates a successful attack and a failed defense.

5. Results and Discussion

This section presents a detailed analysis of the experimental results. We evaluate the resilience of our proposed SAFE-RAN framework against both untargeted and targeted attacks, analyze its convergence and scalability, and discuss the broader implications of our findings for the security of O-RAN environments.

5.1. Resilience to Poisoning Attacks under IID and Non-IID Conditions

Figure 2 presents the core results of our study, comparing the F1-Score of the final global IDS model under an increasing percentage of malicious clients launching a sign-flipping attack. This experiment serves as a fundamental test of robustness against brute-force performance degradation.

As expected, Standard FedAvg proves extremely brittle. Its performance, which is strong in a trusted environment, catastrophically degrades as malicious clients are introduced. With just 20% attackers, the F1-Score plummets below 50%, rendering the collaboratively trained model completely useless. This is because the naive averaging process allows the malicious updates to directly negate the learning progress of the honest clients.

The baseline defense mechanisms, Krum and Trimmed Mean, offer a degree of partial protection but ultimately fail to provide sufficient robustness. Krum's performance degrades significantly, especially in the challenging Non-IID setting. This is because the inherent statistical heterogeneity among honest clients causes their genuine updates to be farther apart in the high-dimensional space, making it more difficult for a distance-based metric like Krum to distinguish between a benign, non-IID update and a malicious one. Trimmed Mean, which filters based on update magnitude (L2-norm), is easily subverted by the sign-flipping attack, as the malicious updates have the same norm as their honest counterparts.

In stark contrast, our SAFE-RAN (TCA) framework demonstrates exceptional and stable resilience. It maintains a high F1-Score of over 97.5% in the IID case and, more importantly, over 96% in the realistic Non-IID case, even when 30% of the network is compromised. This highlights TCA's superior ability to effectively identify and filter malicious updates. By focusing on directional alignment rather than magnitude or proximity, TCA correctly identifies that the malicious updates, regardless of their origin, point away from the consensus of honest learning. The minimal performance drop between the IID and Non-IID scenarios underscores its robustness to the statistical heterogeneity expected in real-world deployments.

This empirical result validates our theoretical premise from Section 4.1 by focusing on the *direction* of the

update vectors, TCA is inherently robust to the statistical data heterogeneity in the Non-IID setting, where distance-based metrics like Krum falter." This reinforces the paper's internal consistency.

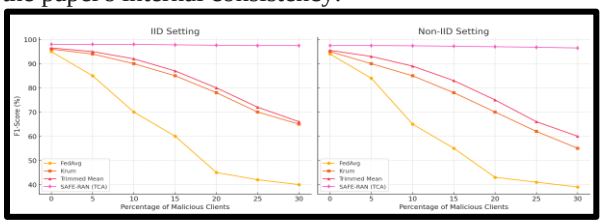

Figure 2: Resilience to Sign-Flipping Attacks. F1-Score of the global model versus the percentage of malicious clients in (a) IID and (b) Non-IID data settings. SAFE-RAN (TCA) maintains high performance (>96%) even with 30% attackers, while all baselines, especially FedAvg, degrade catastrophically.

5.2. Visualizing the Defense Mechanism

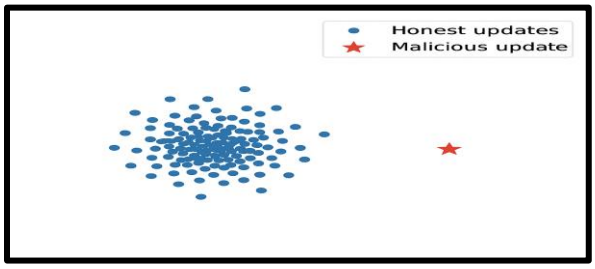

Figure 3: Visualization of Client Model Updates

Figure 3 offers a clear visual demonstration of TCA's effectiveness using a 2D t-SNE projection. Honest client updates, depicted as blue dots, form a compact and consistent directional cluster, even under non-IID data conditions. In contrast, the malicious update—represented by a red star—lies sharply apart from this cluster, exposing its divergent intent. This directional separation validates TCA's core principle: by focusing on the angle (cosine similarity) rather than distance, it can robustly detect and filter out malicious gradients without mistakenly penalizing statistically unique but honest updates.

5.3. Defense against Targeted Backdoor Attacks

Table 2 evaluates the frameworks' resilience against a more sophisticated and insidious targeted backdoor attack, where the adversary's goal is to make the IDS misclassify "Port Scanning" traffic as "Benign," creating a hidden vulnerability.

Table 2: Performance against Targeted Backdoor Attack (20% Malicious Clients)

Framework	Overall Accuracy	Backdoor ASR
Standard FedAvg	89.1 %	98.2%
SAFE-RAN (TCA)	98.0%	4.5%

The results are striking. The backdoor is highly effective against Standard FedAvg, achieving a 98.2% Attack Success Rate (ASR). This creates a dangerous scenario where the overall accuracy (89.1%) appears deceptively high, giving a false sense of security while the model contains a critical, exploitable flaw.

SAFE-RAN (TCA), however, not only maintains a high overall accuracy of 98.0% but also effectively neutralizes the backdoor, reducing its success rate to a mere 4.5%. This success is directly attributable to TCA's core mechanism. To install the backdoor, the attacker must train their local model on a poisoned dataset, which forces the model's update to learn a different objective (i.e., to misclassify a specific attack). This altered objective results in a gradient that points in a different direction from the gradients of honest clients who are learning the correct data distribution. TCA's cosine similarity check detects this directional deviation, flags the malicious update as an outlier, and discards it, preventing the backdoor from ever being integrated into the global model. This demonstrates TCA's ability to defend against not just overt performance degradation but also stealthy, targeted manipulations.

5.4. Scalability and Convergence Analysis

A practical defense must be both effective and efficient. Our scalability tests confirmed that the aggregation time for TCA scales quadratic ally ($O(N^2)$) with the number of clients, as predicted by the theoretical analysis. While this is computationally more intensive than FedAvg's linear scaling, the absolute time remains practical for typical O-RAN deployments where the number of participating CUs/DUs is expected to be in the tens or low hundreds.

The convergence analysis, depicted in Figure 4, further reinforces TCA's stability. While the accuracy of FedAvg under attack becomes highly erratic and fails to converge, SAFE-RAN achieves stable and rapid convergence to a high-accuracy state, demonstrating its ability to ensure consistent learning progress in hostile environments.

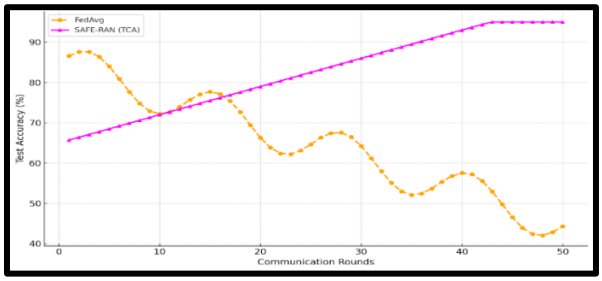

Figure 4: Test accuracy over communication rounds with 20% malicious clients

5.5. Discussion and Broader Implications

The empirical results confirm that SAFE-RAN, powered by the TCA algorithm, provides a practical and robust solution for collaborative AI in untrusted environments like O-RAN. The directional nature of TCA is highly effective at identifying a wide spectrum of attacks, from naive performance degradation to stealthy backdoors.

This work has broader implications for the future of open and disaggregated networks:

1. Enabling Trust in O-RAN: A key promise of O-RAN is a multi-vendor ecosystem. SAFE-RAN provides a critical mechanism for establishing computational trust in this zero-trust hardware environment. It allows a network operator to safely leverage AI/ML contributions from components supplied by diverse, potentially untrusted vendors without risking the integrity of network-wide security models.
2. Informing O-RAN Standardization: Our findings strongly suggest that the Non-Real-Time RIC is the architecturally sound location for implementing robust aggregation logic. To ensure a baseline of security for any collaborative AI/ML task across the network, future O-RAN Alliance security specifications could recommend or even mandate the inclusion of certified, robust aggregation modules within the RIC.

Limitations: While highly effective, TCA's primary limitation is its assumption that the number of attackers, m, is known or can be reliably estimated. An incorrect estimation could lead to either insufficient filtering or the exclusion of honest clients. Furthermore, highly sophisticated adaptive or colluding attackers might devise strategies to evade the cosine similarity check, for instance, by coordinating their malicious updates to form a "malicious cluster" that appears benign to the algorithm. Future work will focus on adaptive methods for estimating m and defending against such collusive attacks.

6. Conclusion and Future Work

In this paper, we addressed the critical challenge of securing multi-vendor O-RAN architectures through collaborative, privacy-preserving machine learning. We introduced SAFE-RAN, a federated learning framework designed to be resilient against adversarial participants. Our key contribution, the Trimmed Cosine Aggregation (TCA) algorithm, supported by theoretical analysis, proved highly effective at defending a global IDS model against a range of poisoning attacks under realistic network conditions. Our extensive simulations show that SAFE-RAN can maintain high detection performance and stability where standard FL and other robust methods fail.

SAFE-RAN provides a robust and practical solution for deploying trustworthy AI-driven security in O-RAN. Future work will focus on three main directions: (1) implementing SAFE-RAN on a physical O-RAN testbed using tools like Open Air Interface and the OSC RIC framework to validate its performance in a real-world setting; (2) developing adaptive defense mechanisms that can dynamically estimate the number of attackers and defend against colluding adversaries; and (3) extending the SAFE-RAN framework to other collaborative tasks in O-RAN, such as network slice optimization and radio resource management.

Appendix: Proof of Convergence for TCA

Theorem 1: Let the global loss function L(w) be μ-strongly convex and L-smooth. Let w_i be the global model parameters at communication round *t.* Assume a set of *N* clients, consisting of *N-m* honest clients in set *H* and *m* malicious clients in set *M*, where the honest majority assumption holds ($m < \frac{N}{2}$). Let the gradients from honest clients, g_i for $i \in H$, be unbiased estimators of their local loss gradients $\nabla L_i(w_t)$ with bounded variance. Under the assumption of directional separation between honest and malicious updates, the aggregated gradient g_{TCA} produced by the Trimmed Cosine Aggregation algorithm is a descent direction in expectation, that is,

$$\mathbb{E}[\langle g_{TCA}, \nabla \mathrm{L}(w_t)\rangle] > 0 \qquad \text{for } \nabla \mathrm{L}(w_t) \neq 0$$

This ensures that the global model, when updated usingg_{TCA}, converges towards the optimal solutionw^*.

Proof:

The proof proceeds in two main parts: 1. Establish the correct Trimming Property to show that TCA correctly identifies and removes malicious updates, 2. Use the property to show that the resulting aggregated gradient has a positive dot product with the true global gradient in expectation.

1. Assumptions

We formalize the conditions required for the proof:

a. A1 (Honest Majority): The number of malicious client's *m* is strictly less than half the total number of clients: $m < \frac{N}{2}$.

b. A2 (Unbiased Honest Gradients): For any honest client $i \in H$, its stochastic gradient g_i is an unbiased estimator of its true local gradient $\nabla L_i(w_t)$. That is, $\mathbb{E}[g_i] = \nabla L_i(w_t)$. The true global gradient is $\nabla L_i(w_t) = \left(\frac{1}{N}\right)\sum_{i=1}^{N} \nabla L_i(w_t)$.

c. A3 (Directional Alignment of Honest Updates): Gradients from honest clients are directionally aligned with each other. There exists a constant $c_h > 0$ such that for any two distinct honest clients $i, j, \in H$, their cosine similarity is bounded below: $\cos(g_i, g_j) \geq c_h$.

d. A4 (Directional Outliers of Malicious Updates): Malicious updates are directional inconsistent with honest updates. We assume the attacker's goal is to introduce updates that are not aligned with the honest consensus. There exists a constant $c_m < c_h$ such that for

any malicious client $k \in M$ and any honest client $i \in H$: $\cos(g_k, g_i) \leq c_m$.
For a sufficiently effective attack, c_m will be close to 0 or negative. We only need c_m to be discernibly smaller thanc_h.

2. Correct Trimming Property

We show that under these assumptions, the reputation score R_i of any honest client $i \in H$ will be greater than the reputation score R_k of any malicious client $k \in M$.

The reputation score for an honest client $i \in H$ is:

$$R_i = \sum_{j \in H, j \neq i} \cos(g_i, g_j) + \sum_{k \in M} \cos(g_i, g_k).$$

Using assumptions A3 and A4, we can bound this score from below:

$$R_i \geq (N - m - 1)c_h + m * c_m$$

The reputation score for a malicious client $k \in M$ is:

$$R_k = \sum_{j \in H} \cos(g_k, g_j) + \sum_{l \in M, l \neq k} \cos(g_k, g_l)$$

For an honest client's score to be greater than a malicious client's score ($R_i > R_k$), we need:

$$(N - m - 1)c_h + m * c_m > (N - m)c_m + (m - 1)$$

Rearranging the terms, we get:

$$(N - m - 1)c_h - (N - 2m)c_m > m - 1$$

Given the honest majority assumption (*A1, N-m > m*), the term ($N - 2m$) is positive. If c_h is sufficiently larger than c_m(a premise of TCA's effectiveness), this condition holds. Thus, the scores of honest clients will be strictly greater than those of malicious clients.

When the clients are sorted by their reputation scores in ascending order, the *m* malicious clients will occupy the first *m* positions. The trimming step of TCA will therefore correctly remove the set of malicious clients *M*. This leads to the key result that the set of trusted clients selected by TCA is exactly the set of honest clients:

$$S_{trusted} = H$$

3. Proof of Descent Direction

Now we analyze the expected dot product of the TCA-aggregated gradient g_{TCA} with the true global gradient$\nabla L(w_t)$. Given the Correct Trimming Property, g_{TCA} is the average of the gradients from the honest clients H, where $|H| = N - m$:

$$g_{TCA} = \frac{1}{N - m} \sum_{i \in H} g_i$$

We compute the expectation:

$$\mathbb{E}[\langle g_{TCA}, \nabla L(w_t) \rangle] = \mathbb{E}\left[\langle \frac{i}{N - m} \sum_{i \in H} g_i, \nabla L(w_t) \rangle\right]$$

Using linearity of expectation and of the dot product:

$$\mathbb{E}[\langle g_{TCA}, \nabla L(w_t) \rangle] = \frac{1}{N - m} \sum_{i \in H} \mathbb{E}[\langle g_i, \nabla L(w_t) \rangle]$$

By assumption A2, each honest client's stochastic gradient g_i is an unbiased estimator of its local gradient$\nabla L_i(w_t)$. That is:

$$\mathbb{E}[g_i] = \nabla L_i(w_t)$$

Hence, since $\nabla L(w_t)$ is deterministic at time *t*, we can move the expectation inside the dot product:$\mathbb{E}[\langle g_i, \nabla L(w_t) \rangle] = \langle \mathbb{E}[g_i], \nabla L(w_t) \rangle = \langle \nabla L_i(w_t), \nabla L(w_t) \rangle$

Substituting back:

$$\mathbb{E}[\langle g_{TCA}, \nabla L(w_t) \rangle] = \frac{1}{N - m} \sum_{i \in H} \langle \nabla L_i(w_t), \nabla L(w_t) \rangle$$

Positive Correlation Assumption

To ensure meaningful progress, we make the standard assumption that the learning objectives of honest clients are not fundamentally opposed to the global objective. This implies that the local gradients are, on average, positively aligned with the global gradient:

$$\langle \nabla L_i(w_i), \nabla L(w_i) \rangle \geq 0 \qquad \text{for all } i \in H$$

Further, unless the model is already at an optimum ($\nabla L(w_t) = 0$), there exists at least one client $i \in H$ for which the local gradient is strictly positively aligned with the global gradient. That is

$$\langle \nabla L_i(w_i), \nabla L(w_i) \rangle > 0 \qquad \text{for some } i \in H$$

Therefore, the sum is **strictly positive**:

$$\sum_{i \in H} (\nabla L_i(w_t), \nabla L(w_t)) > 0$$

And hence:

$$\mathbb{E}[\langle g_{TCA}, \nabla L(w_i) \rangle] > 0$$

4. Conclusion

The above result shows that the TCA-aggregated gradient gTCA is, in expectation, a descent direction for the global objective function L(w)L(w). By optimization theory, taking steps in a descent direction guarantees convergence to a local minimum. In the case of strongly convex loss functions, this further ensures convergence to the global minimum.

5. Reference

[1] Davies, T., Eiza, M. H., Shone, N., & Lyon,R.(2025),A Collaborative Intrusion Detection System Using Snort IDS Nodes. *ArXiv*. https://arxiv.org/abs/2504.16550

[2] El-Gayar, M. M., Alrslani, F. A. F., & El-Sappagh,S. (2024). Smart collaborative intrusion detection system for securing vehicular networks using ensemble machine learning model. *Information (Switzerland)*, *15*(10), 583. https://doi.org/10.3390/info15100583

[3] El-Hajj, M. (2025). Secure and trustworthy Open Radio Access Network (O-RAN) optimization: A zero-trust and federated learning framework for 6G

networks. *Future Internet*, *17*(6), 233. https://doi.org/10.3390/fi17060233

[4] Gavade, J. B. (2022). Overview of Open RAN (O-RAN) architecture, challenges and future directions. *Journal of Emerging Technologies and Innovative Research*. Volume *9*, Issue 8, 301 – 303.

[5] Gugueoth, V., Safavat, S., & Shetty, S. (2023). Security of Internet of Things (IoT) using federated learning and deep learning—Recent advancements, issues and prospects. *ICT Express*, *9*(5), 941–960. https://doi.org/10.1016/j.icte.2023.03.006

[6] Igenewari, L. S., & Okoh, O. E. (2025). *Adversarial attacks and defenses in AI systems: Challenges, strategies, and future directions*. International Journal of Research and Innovation in Applied Science (IJRIAS). Volume X, Issue VI, https://doi.org/10.51584/IJRIAS

[7] Kaur, N., Singh, S., Shivaji Deore, D., Vidhate, A. V., Haridas, D., Varma Kosuri, G., & Ravindra Kolhe, M. (2024). Robustness and security in deep Learning: Adversarial attacks and Countermeasures. *Journal of Electrical Systems*, *20*(3), 1250 -1247.

[8] Kondepu, K., Tamma, B. R., & Gudepu, V. (2025). O-RIDE: Open-RAN innovation and deployments Exploration. In *Proceedings of the 26th International Conference on Distributed Computing and Networking* (pp. 426–429). https://doi.org/10.1145/3700838.3703690

[9] Polese, M., Bonati, L., D'Oro, S., Basagni, S., & Melodia, T. (2023). Understanding O-RAN Architecture, interfaces, algorithms, security, and Research challenges. *IEEE Communications Surveys and Tutorials*, *25*(2), 1376–1411. https://doi.org/10.1109/COMST.2023.3239220

[10] Priambodo, D. F., Faizi, A. H. N., Rahmawati, F.D., Sunaringtyas, S. U., Sidabutar, J., & Yulita, T.(2024). Collaborative intrusion detection system with Snort machine learning plugin. JOIV: *International Journal on Informatics Visualization*. V. 8, Issue 3, 1230 -1238

[11] Wijethilaka, S., Yadav, A. K., Braeken, A., & Liyanage, M. (2024). Privacy-preserving federated learning framework for Open Radio Access Networks (ORAN). In *Proceedings of the IEEE Global Communications Conference (GLOBECOM)* (pp. 4515–4520).https://doi.org/10.1109/GLOBECOM52923.2024.10901104

[12] Wypiór, D., Klinkowski, M., & Michalski, I. (2022).Open RAN—Radio access network evolution, benefits and market trends. *Applied Sciences (Switzerland)*, *12*(1), 408. https://doi.org/10.3390/app12010408