A Trust-Aware Federated Learning Framework Based on Zero Trust Architecture for Secure Cloud Intrusion Detection

Eman M. Mohamed[1], Suzan S. Basloom[2]
Computer Science Department-Faculty of Computers and Information, Menoufia University, Egypt[1]
Department of Cyber Security, College of EIT, Buraydah Private Colleges, Buraydah, Saudi Arabia[1,2]

***Abstract*—Cloud computing provides scalable infrastructure but introduces critical challenges to data privacy, trust management, and intrusion detection. To address these issues, we present a trust-aware framework that integrates a Federated Learning (FL)-based distributed ensemble with Zero Trust Architecture (ZTA) for secure cloud environments. The framework enables collaborative learning while keeping data local and private. Zero Trust is implemented by periodically verifying identity and evaluating client contributions with a lightweight local trust filter before aggregation. For reliability and agreement with the local model, an ensemble of Random Forest and Decision Tree classifiers is used. Contributions are only aggregated when satisfying the predefined trust and performance thresholds, thereby reducing the influence of potentially unreliable or malicious inputs under the assumed threat model. The proposed framework is evaluated on the UNSW-NB15 dataset under non-IID client data distribution, achieving 96% accuracy, 95.1% precision, 94% recall, 0.96 AUC-ROC, and an average per-round latency of 180 ms. Comparative results demonstrate that the framework provides a better trade-off between detection accuracy, privacy preservation, and trust enforcement than several state-of-the-art baselines, positioning it as a practical solution for cloud intrusion detection.**

***Keywords*—Cloud security; federated learning; zero trust architecture; intrusion detection system; trust filtering mechanism; decentralized learning**

I. Introduction

Cloud computing has become an integral part of today's digital infrastructures. It provides organizations with the ability to store, manage and process data in an efficient manner, while also providing scalability and cost-effectiveness. The increasing demand for shared and on-demand resources is driving the wider adoption of cloud services in many application domains.

Recent threat assessments have shown that the sophistication of intrusion techniques and evolving attack vectors targeting cloud environments is increasing [1]. These developments, especially in distributed and heterogeneous cloud environments, have significantly increased the complexity of malicious activities and exposed the limitations of traditional detection mechanisms.

Legacy cloud security models have been over-reliant on perimeter-based trust assumptions where users and devices are trusted once they are granted access inside the perimeter. These models were useful in relatively static environments, but they are no longer sufficient for modern cloud infrastructures, with dynamic services, distributed users, and changing access patterns. Hence, security frameworks need to go beyond static trust boundaries and deploy mechanisms that provide continuous verification and fine-grained access control.

These challenges make Federated Learning (FL) and Zero Trust Architecture (ZTA) promising approaches for securing cloud environments. Federated Learning (FL) allows multiple distributed clients to collaboratively train a shared model without exchanging raw data, making it especially suitable for privacy-sensitive cloud environments. FL enables the decentralized training of models, which increases data privacy and reduces the necessity of sensitive data being transferred through the network.

At the same time, Zero Trust Architecture (ZTA) has been seen as an emerging alternative to traditional perimeter-based security models. ZTA does not automatically trust internal users and devices, but applies continuous identity verification and dynamic access control [2]. These features make ZTA a good fit for modern cloud environments, where resources are spread out and access requests are common. Although FL and ZTA have many advantages, the existing works either utilize them in isolation or rely on extra trust enforcement mechanisms, which increase the system complexity and overhead. In addition, many solutions do not provide lightweight client-side validation mechanisms that can filter out unreliable contributions prior to aggregation in federated environments.

To alleviate these limitations, we propose a trust-aware federated learning framework that integrates FL with Zero Trust principles for secure cloud intrusion detection. In this work, we propose a unified system for client-side trust validation, identity verification and localized data processing. The proposed approach integrates federated learning with Zero Trust enforcement and lightweight trust filtering to offer a practical solution for enhancing cloud security while maintaining data privacy and operational efficiency. The proposed framework incorporates lightweight client-side trust validation directly into the federated process, allowing for efficient and self-contained trust enforcement, as opposed to existing methods that depend on external trust mechanisms.

A. Motivations and Contributions

Motivated by the above limitations of existing FL and ZTA-based security approaches, this paper proposes a trust-aware federated learning framework that combines FL with ZTA to

improve intrusion detection in cloud environments. The main contributions of this paper can be summarized as follows:

- A holistic FL-ZTA framework that integrates decentralized federated learning with Zero Trust principles based on continuous identity verification, trust validation on the client-side, and the weighted aggregation of verified client prediction outputs, providing a unified architecture for secure and privacy-preserving cloud intrusion detection.
- A lightweight client-side trust filtering mechanism for assessing local model agreement and prediction performance before federated aggregation. The proposed mechanism improves the reliability of the global aggregation process by filtering unreliable or low-quality client contributions at the source with low computational overhead.
- A client-side trust assessment-aided ensemble-based local validation strategy using Random Forest and Decision Tree classifiers. Model agreement and confidence-based decision rules are employed to enhance the stability of local predictions and to ensure that only trusted prediction outputs are part of the federated aggregation process.

The effectiveness of the proposed framework is evaluated on the public UNSW-NB15 dataset [3]. Comparative experiments with recent FL-based and ZTA-based approaches, as well as conventional federated learning and centralized learning baselines, show that the proposed framework improves intrusion detection performance while preserving data privacy and maintaining efficient trust enforcement in distributed cloud environments.

The rest of this paper is structured as follows. Section II presents some related works on Federated Learning and Zero Trust Architecture in cloud security. Section III describes the proposed trust-aware FL-ZTA framework and its main components. In Section IV, we describe the evaluation metrics used for measuring detection performance and operational efficiency. Section V presents the experimental setup and datasets used for evaluation. Section VI discusses the experimental results and comparative analysis. Finally, Section VII concludes the paper and discusses future works.

II. Related Work

In this section, we review existing research on cloud security from three major perspectives, namely security approaches based on Federated Learning, Zero Trust Architecture in cloud environments, and studies that combine Federated Learning with Zero Trust principles. The most relevant studies are summarized in Table I, including their focus areas and limitations.

A. Security Based on Federated Learning

Federated Learning (FL) is gaining popularity to improve data privacy in distributed and cloud computing systems, by allowing collaborative training of models without exchanging raw data. Truex et al. [4] proposed a hybrid privacy-preserving FL approach by combining differential privacy and secure multiparty computation to achieve better confidentiality with higher computational overhead. Liu et al. [5] investigated the use of FL in cloud and 5G-enabled environments with blockchain-supported trust management to reduce malicious client behavior. However, their approach resulted in significant resource consumption and deployment complexity. Xu et al. [6] proposed TAPFed, a threshold-based secure aggregation mechanism to reduce the communication overhead while ensuring fairness, but the key management and scalability issues still prevail.

B. Zero Trust Architecture in Cloud Environments

Zero Trust Architecture (ZTA) is the emerging alternative to traditional perimeter-based security models. It demands continuous authentication and strict access control. Ahmadi [7] examined the use of ZTA in cloud infrastructures, demonstrating its efficacy in countering internal threats, yet encountered problems with policy enforcement and integration complexity. Bukola and Martin [8] further elaborated the role of ZTA in managing lateral movement in cloud-based enterprise networks and pointed out the performance overhead of continuous verification methods for large-scale deployment scenarios.

C. Zero Trust and Federated Learning Schemes

Recently, the combination of Federated Learning and Zero Trust has been studied to improve trust management in distributed environments. Pokhrel et al. [9] proposed a blockchain assisted FL framework under Zero Trust principles for IoT systems; it improved trust management, but increased system complexity. Ma et al. [10] proposed a cross-domain Zero Trust framework integrated with FL for dynamic authentication and authorization in large-scale environments. However, scalability remains a challenge. Vucovich et al. [11] introduced FedBayes, a Bayesian aggregation scheme for federated learning with Zero Trust assumptions which enhanced robustness to adversarial updates at the cost of increased model complexity. Kawalkar and Bhoyar [12] and Asad and Otoum [13] extended this research and analyzed the combination of FL and Zero Trust for cloud and wireless security and found that while security is improved, there are challenges related to efficiency and real-world deployment.

Recent studies in 2026 have integrated Zero-Trust principles with federated learning for intrusion detection. Alshammari et al. [14] proposed SecureTrust-FL which uses differential privacy and adversarial learning with Zero-Trust-based blockchain trust management, with an accuracy of 92.91% on the UNSW-NB15 dataset, but the use of blockchain incurs computational overhead. The ZTA-FL based on TPM-based attestation and SHAP-weighted aggregation proposed by Singh et al. [15] achieves a detection rate of 97.8% and accuracy of 93.2% under 30% Byzantine attacks. However, the deployment is limited by the TPM hardware. Aljabbari et al. [16] proposed a hierarchical FL framework for IoMT with self-sovereign identity, differential privacy, and Ethereum anchoring for public auditability for verification. The architectural complexity is however high. Instead, our framework provides a lightweight software-only two-tier architecture. On the client side, ensemble-based validation is performed using Random Forest and Decision Tree, while prediction aggregation is performed on the server side. This allows for improved performance

TABLE I. SUMMARY OF RELATED WORK ON FEDERATED LEARNING, ZERO TRUST ARCHITECTURE, AND THEIR INTEGRATION

Ref.	Category	Main Contribution	Limitations
[4]	FL	Privacy-preserving federated learning using differential privacy and secure multiparty computation.	High computational overhead and limited scalability.
[5]	FL	Blockchain-assisted federated learning for cloud and 5G security.	High resource consumption and deployment complexity.
[6]	FL	Threshold-based secure aggregation to reduce communication overhead.	Key management and scalability challenges.
[7]	ZTA	Application of Zero Trust Architecture in cloud infrastructures.	Policy enforcement and integration complexity.
[8]	ZTA	Zero Trust-based analysis to restrict lateral movement in cloud networks.	Performance overhead due to continuous verification.
[9]	FL + ZTA	Blockchain-based FL framework under Zero Trust principles.	System complexity and scalability limitations.
[10]	FL + ZTA	Cross-domain Zero Trust framework combined with federated learning.	Scalability concerns in large-scale environments.
[11]	FL + ZTA	Bayesian aggregation mechanism for federated learning under Zero Trust.	Increased model and computational complexity.
[12]	FL + ZTA	Integrated FL and Zero Trust framework for secure cloud communications.	Efficiency and practical deployment challenges.
[14]	FL+ZTA	FL with blockchain trust, DP, adversarial learning, and ZTA; 92.91% acc, 95.50% AUC on UNSW-NB15.	Blockchain overhead and latency.
[15]	FL+ZTA	TPM attestation, SHAP aggregation; 97.8% detection rate, 93.2% acc under 30% Byzantine attacks.	TPM hardware dependency limits deployment.
[16]	FL+ZTA	Publicly Auditable Hierarchical FL: SSI, DP, Krum, ML-DSA, Merkle, Ethereum anchoring for IoMT.	High complexity; blockchain overhead; scalability limits.

(96.0% accuracy, 98.0% specificity) without the need for consensus protocols or special hardware.

D. Summary and Research Gaps

As summarized in Table I, previous studies show that Federated Learning (FL) and Zero Trust Architecture (ZTA) have significantly improved security and privacy in cloud and distributed environments. However, most of the integrated FL–ZTA solutions rely on heavy external trust mechanisms such as blockchain or incur high computational and communication overhead. Furthermore, many existing approaches lack integrated local trust validation mechanisms and give limited attention to lightweight ensemble-based client-side validation strategies that can filter unreliable client contributions before aggregation. These limitations motivate the proposed FL-ZTA framework that combines lightweight client-side trust filtering, ensemble-based local validation, and Zero Trust principles into a unified federated learning framework to enhance intrusion detection while preserving data privacy and operational efficiency.

III. PROPOSED METHODOLOGY

This section describes the proposed FL-ZTA framework integrated with Zero Trust Architecture (ZTA) principles to improve security and reliability in distributed cloud environments. The framework strictly requires verification and trust assessment to be applied on both client and server-side before any client contribution can enter the aggregation process. Fig. 1 shows the general workflow of the presented Trust-Aware Federated Learning with Zero Trust Enforcement (FL-ZTA). The framework is divided into two main phases: client-side operations (authentication, local training, and trust validation) and server-side operations (verification, anomaly screening, and weighted aggregation). Both stages reduce the influence of unreliable client contributions, and only verified prediction outputs are considered for the final aggregation. A secure channel for transmission safeguards the communication between the clients and the server.

In this work, Zero Trust principles are applied specifically to federated client participation and contribution handling through identity verification, authorization, continuous validation, and server-side verification. The framework does not claim to implement a complete enterprise-wide Zero Trust Architecture with dedicated policy decision and policy enforcement infrastructure.

The next subsections describe each component of the proposed framework, including client-side operations, trust validation, server-side validation, and the weighted aggregation strategy.

A. Client-Side Operations

At each communication round, every client first undergoes identity verification to ensure authorized participation. Upon successful authentication, the local dataset $\mathcal{D}_i$ is partitioned into a training set $\mathcal{D}_i^{train}$ (80%) and a validation set $\mathcal{D}_i^{val}$ (20%). Two local classifiers, namely Random Forest (f_i^{RF}) and Decision Tree (f_i^{DT}), are trained exclusively using $\mathcal{D}_i^{train}$.

To improve prediction reliability, an ensemble-based validation strategy is applied on the unseen $\mathcal{D}_i^{val}$. The agreement rate $\mathbb{A}_i$ between both models is formally computed as follows:

$$\mathbb{A}_i = \frac{1}{|\mathcal{D}_i^{val}|} \sum_{x \in \mathcal{D}_i^{val}} \mathbb{I}\big(f_i^{RF}(x) = f_i^{DT}(x)\big) \tag{1}$$

where, $\mathbb{I}(\cdot)$ is an indicator function that yields 1 if the predictions of both models match, and 0 otherwise. In cases of disagreement, the confidence scores of both models are evaluated, and the prediction with the higher score is selected, provided it exceeds a predefined confidence threshold τ. In addition, the local performance is evaluated using AUC-ROC on $\mathcal{D}_i^{val}$.

A client contribution is considered eligible for aggregation only if both the agreement score ($A_i \geq \alpha$) and validation performance (AUC-ROC$_i \geq \beta$) satisfy the predefined thresholds.

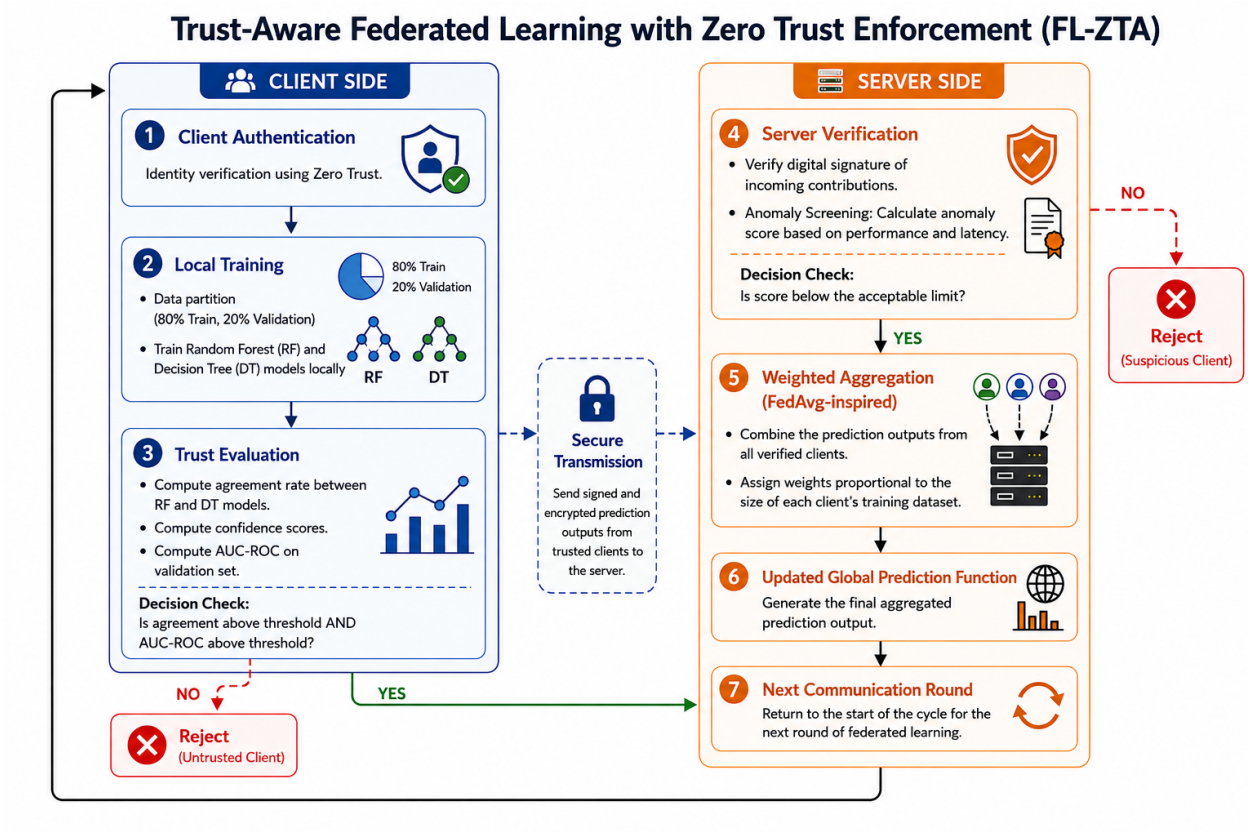

Fig. 1. Overall workflow of the proposed Trust-Aware Federated Learning with Zero Trust Enforcement (FL-ZTA). The client-side is responsible for local data partitioning, ensemble training, and trust evaluation; the server-side is responsible for signature verification, anomaly screening, and FedAvg-inspired weighted aggregation.

These indicators provide operational measures of prediction consistency and validation performance, rather than direct evidence that a client is benign or malicious.

B. Server-Side Verification

Upon receiving clients' contributions, the server first checks the electronic signatures for authenticity and integrity. Next, an anomaly screen is added to identify abnormal or potentially malicious contributions. To quantify this, a composite anomaly score $\mathcal{I}$ is computed for each client in the trusted set, evaluating both the reported validation performance and the communication metadata (e.g., transmission latency Δt_i):

$$\mathcal{I}(P_i, \text{metadata}_i) = \omega_1(1 - \text{AUC-ROC}_i) + \omega_2 \cdot \mathbb{1}_{\Delta t_i > \theta} \quad (2)$$

where, ω_1 and ω_2 are tuning weights ($\omega_1 + \omega_2 = 1$) that balance the importance of performance versus behavioral anomalies, and $\mathbb{1}$ is an indicator function that penalizes the client (adds to the anomaly score) if their transmission latency Δt_i exceeds an acceptable threshold θ.

Contributions exhibiting a high anomaly score ($\mathcal{I} > \gamma$) are excluded from the aggregation process. Only verified and trusted client outputs are admitted to the aggregation pool.

C. Weighted Aggregation of Verified Client Prediction Outputs

Since tree-based machine learning models such as Random Forest and Decision Tree do not support direct parameter aggregation [17], conventional Federated Averaging (FedAvg) cannot be directly applied. Therefore, the proposed framework adopts a weighted aggregation strategy inspired by FedAvg, where verified client prediction outputs are aggregated according to the relative size of each client's local dataset.

For each inference instance x, participating clients generate prediction probabilities using their locally trained models, and the resulting predictions for the same instance are transmitted to the server for weighted aggregation according to the relative sizes of the participating clients' local datasets. No raw training data or shared training dataset is exchanged among clients.

The aggregation is performed over the predicted probabilities, $P_i(y|x)$, rather than hard binary labels, allowing the confidence information of local predictions to be retained during the aggregation process. This approach is compatible with tree-based models and enables the proposed weighted aggregation strategy to combine client prediction outputs according to their local dataset sizes. By aggregating prediction probabilities instead of model parameters, the framework supports the use of non-parametric models while maintaining the decentralized data characteristics of federated learning. The whole procedure is summarized in Algorithm 1.

Algorithm 1 Trust-Aware Federated Learning with Zero Trust Enforcement (FL–ZTA)

Input: Initial aggregation state; communication rounds T; client datasets $\{\mathcal{D}_i\}$; identity verification mechanism ζ; confidence threshold τ; trust thresholds (α, β); anomaly threshold γ

Output: Final aggregated prediction function $P(y|x)$

for $t \leftarrow 1$ **to** T **do**
  $\mathcal{T}_t \leftarrow \emptyset$
  `// Client-side: identity verification, local training, and trust validation`
  foreach *client* C_i **do**
    Verify identity using ζ **if** *verification fails* **then**
      continue
    end
    `// Partition local data into training and validation sets (80%/20%)`
    Partition $\mathcal{D}_i$ into training set $\mathcal{D}_i^{train}$ and validation set $\mathcal{D}_i^{val}$
    `// Train ensemble models exclusively on the training set`
    Train local models f_i^{RF} and f_i^{DT} using $\mathcal{D}_i^{train}$
    `// Local validation: compute trust metrics on unseen validation set`
    Compute agreement rate $\mathbb{A}_i$ on $\mathcal{D}_i^{val}$ Resolve disagreements using confidence scores with threshold τ Compute AUC-ROC$_i$ on $\mathcal{D}_i^{val}$
    if $\mathbb{A}_i \geq \alpha$ ***and*** $\mathrm{AUC-ROC}_i \geq \beta$ **then**
      Construct verified client prediction output $P_i(y|x)$ (signed and encrypted) Add C_i to trusted set $\mathcal{T}_t$
    end
  end
  `// Server-side: verification and anomaly screening`
  $\mathcal{V}_t \leftarrow \emptyset$ **foreach** *client* $C_i \in \mathcal{T}_t$ **do**
    Verify signature using PubKey_i Compute anomaly score $\mathcal{I}(P_i(y|x), \mathrm{metadata}_i)$ **if** $\mathcal{I}(P_i(y|x), \mathrm{metadata}_i) \leq \gamma$ **then**
      Add C_i to verified set $\mathcal{V}_t$
    end
  end
  `// FedAvg-inspired weighted aggregation of verified client prediction outputs`
  if $\mathcal{V}_t \neq \emptyset$ **then**
    $$P(y|x) \leftarrow \sum_{i \in \mathcal{V}_t} \left(\frac{|\mathcal{D}_i^{train}|}{\sum_{j \in \mathcal{V}_t} |\mathcal{D}_j^{train}|} \right) P_i(y|x)$$
  end
end
return $P(y|x)$

D. Trust Validation Mechanism

The proposed FL-ZTA framework is based on a trust validation mechanism. It operates at the client level and verifies whether a local contribution can be trusted before it is aggregated. The evaluation is based on the local consistency and performance indicators so that only reliable prediction outputs influence the global prediction process. The framework minimizes communication overhead by requiring trust validation at the client-side and avoids unreliable contributions affecting the aggregated result. The main steps of the trust-aware federated learning process are summarized in Algorithm 1.

1) Security and privacy considerations: The proposed framework keeps all raw training data at the client-side, thereby reducing direct exposure of sensitive records to the aggregation server. However, data locality alone does not guarantee complete protection against inference attacks such as membership inference, model inversion, or confidence leakage.

The framework assumes that the aggregation server, authentication infrastructure, cryptographic key management, communication channel, validation data, and ground-truth labels are correctly provisioned and protected. These components are treated as trusted system assumptions and are outside the adversarial scope of the current evaluation. The proposed trust mechanisms therefore focus on evaluating client participation and contributions rather than establishing the trustworthiness of these underlying infrastructure components.

IV. Performance Measure

This section presents the evaluation metrics to assess the effectiveness and efficiency of the proposed FL–ZTA framework for cloud intrusion detection. We use standard classification metrics to evaluate detection performance, and latency and communication-related metrics to measure operational efficiency.

A. Metrics of Detection Performance

Let TP, TN, FP, and FN be the number of true positives, true negatives, false positives, and false negatives, respectively. The following metrics are applied:

- **Accuracy:**

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

- **Precision:**

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

- **Recall:**

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

- **F1-score:**

$$F1\text{-}score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{6}$$

- **AUC-ROC:** The Area Under the ROC Curve (AUC-ROC) summarizes the trade-off between true positive rate and false positive rate across different decision thresholds.

B. Efficiency Metrics

In addition to detection performance, the framework is evaluated in terms of operational efficiency. The following metrics are reported:

1) Latency: The time required to complete a learning round, including local training, trust validation, verification, and aggregation.

2) Communication rounds: the number of federated rounds required to reach a stable performance level. Čommunication overhead the total amount of transmitted data (e.g., client contributions and metadata) exchanged between clients and the server.

V. Evaluation Setup

The experimental evaluation was performed on a Python-based simulation environment using popular machine learning libraries, including scikit-learn and NumPy. The implementation focuses on tree-based machine learning algorithms, such as Random Forest and Decision Tree classifiers, in a federated learning framework.

The experiments were conducted on the UNSW-NB15 dataset that is a well-known dataset for the assessment of intrusion detection systems in cloud-based environments. The dataset was pre-processed and converted to a binary classification task of differentiating between normal and malicious traffic. The UNSW-NB15 dataset is imbalanced, with an average of around 85% normal traffic and 15% malicious traffic. We used a stratified split during training-test split to reduce the possible bias and maintained the same class ratio for each client as in the original dataset.

To simulate a federated learning environment, the dataset was partitioned across multiple distributed clients, where each client maintained a private subset of local data. The client data distribution was intentionally non-IID to better reflect realistic distributed cloud scenarios.

The evaluation was conducted using the UNSW-NB15 dataset in a simulated federated environment with 10 clients. This configuration provides a controlled setting for evaluating the proposed framework; however, the use of a single benchmark dataset and a limited number of clients limits the extent to which the findings can be generalized to larger and more heterogeneous federated environments. Evaluation using additional intrusion detection datasets and larger client populations is therefore left for future work.

During each communication round, every client independently trained two local classifiers, namely Random Forest and Decision Tree, using its private dataset. A local trust validation mechanism was then applied based on model agreement rate and AUC-ROC thresholds to determine whether the client contribution should participate in the aggregation process.

The simulation environment is composed of $N = 10$ distributed clients. To simulate heterogeneous cloud scenarios, we partitioned the training data among clients with a Dirichlet distribution with concentration parameter $\alpha = 0.3$, which induces a significant non-IID skew across clients. This value was chosen empirically to achieve moderate-to-high skew and thus a realistic heterogeneous cloud environment, where clients (e.g. different enterprise departments) have different data distributions. The federated process was run for $T = 5$ communication rounds, and the final performance was evaluated after the completion of these rounds. The Random Forest classifier was configured with 50 estimators and a maximum depth of 10 while the Decision Tree was limited to a maximum depth of 10 . This ensured a lightweight computational footprint suitable for resource-constrained edge devices.

The trust-validation hyperparameters were predefined for the experimental evaluation to balance predictive performance and enforcement stringency. The agreement threshold $\alpha = 0.85$ and AUC-ROC threshold $\beta = 0.80$ guarantee the high consensus and quality of the local models, and the confidence threshold $\tau = 0.70$ determines the resolution of prediction disagreement. For server-side screening, an anomaly threshold of $\gamma = 0.75$ is used for a composite score weighted by $\omega_1 = 0.6$ (validation performance) and $\omega_2 = 0.4$ (behavioral latency). When transmission delays exceed the latency threshold $\theta = 200$ ms, an anomaly penalty is applied to achieve robust filtering of unreliable clients. Since tree-based models do not support direct parameter averaging, the proposed framework employs a weighted aggregation strategy inspired by Federated Averaging (FedAvg) [17]. Instead of aggregating model parameters, the framework combines client prediction outputs proportionally according to local dataset sizes.

The evaluation considered both classification performance and operational efficiency. Performance metrics included accuracy, precision, recall, F1-score, AUC-ROC, and specificity, while efficiency was evaluated using latency, communication rounds, and communication overhead. All experiments were conducted under consistent experimental conditions to ensure fairness and reproducibility. Each experiment was repeated five times with different random seeds, and results are reported as the mean $\pm$ standard deviation to demonstrate statistical robustness.

VI. Result and Discussion

To rigorously assess the effectiveness of the proposed FL-ZTA framework in enforcing security and maintaining high detection performance, we conducted a series of experiments using the publicly available UNSW-NB15 dataset, configured for binary classification (normal vs. malicious traffic).

To ensure a fair and reproducible evaluation, all models, including the ablation baselines and re-implemented state-of-the-art (SOTA) approaches, were evaluated under the same experimental conditions, including an identical non-IID data partition strategy, hardware environment, and preprocessing pipeline.

A. Experimental Setup and Fair Baselines

We compare our framework against the following baselines to isolate the true impact of our contributions:

- Centralized Training (Upper Bound): All client data is pooled to train a global ensemble, representing theoretical maximum performance without privacy constraints.
- Local Isolated Training (Lower Bound): Clients train locally without any communication or aggregation.

- Untrusted Aggregation (No Filtering): Distributed aggregation of all client predictions without applying our trust-validation step.
- Liu et al.[4] (Re-impl.): We re-implemented their blockchain-assisted trust aggregation mechanism under our exact setup.
- Kawalkar & Bhoyar[12] (Re-impl.): We re-implemented their integrated FL-ZTA policies under our exact setup.
- Vucovich et al.[11] (Re-impl.): We re-implemented their FedBayes Bayesian aggregation mechanism under our exact non-IID setup.

B. Detection Performance Evaluation

Table II presents the comparative classification performance across all evaluated models.

1) Analysis of Table II: All results in Table II are reported as the mean $\pm$ standard deviation over five independent runs with different random seeds. This statistical approach ensures the reliability and reproducibility of the reported performance metrics.

a) Superiority over re-implemented SOTA: By testing recent frameworks under our strict non-IID constraints, we demonstrate that our proposed framework achieves the highest performance across all metrics. Notably, while Vucovich et al.'s FedBayes mechanism achieved a strong $94.0 \pm 0.3\%$ accuracy, our framework surpassed it by 2.0 percentage points, with a massive 4.2% percentage-point improvement in Specificity ($98.0 \pm 0.2\%$ vs. $93.8 \pm 0.5\%$). Furthermore, compared to Liu et al.'s blockchain approach ($91.0 \pm 0.4\%$), our framework provides a $+5.0\%$ percentage-point improvement in accuracy while entirely eliminating the consensus-induced overhead. The low standard deviations across all metrics confirm the stability and consistency of our framework under varying random conditions.

b) Effectiveness of trust filtering: The difference between *Untrusted Aggregation* ($91.5 \pm 0.4\%$) and our *Proposed Framework* ($96.0 \pm 0.2\%$) shows the effect of using our local trust validation mechanism alone. By excluding low-quality or potentially unreliable prediction outputs at the client-side according to their validation-set performance, the proposed mechanism reduces the influence of potentially problematic client contributions on the global aggregation. The **Specificity** metric shows a significant improvement ($98.0 \pm 0.2\%$ vs. $90.2 \pm 0.6\%$), which suggests that our framework significantly reduces the false-positive rate, a crucial requirement for cloud security systems in operation where false alerts can lead to alert fatigue.

c) Privacy-performance trade-off: As expected, the *Centralized Training* model obtains the highest theoretical accuracy ($97.2 \pm 0.2\%$). However, this approach requires pooling raw client data, conflicting with strict data privacy regulations (e.g., GDPR). Our proposed framework achieves a closely competitive accuracy ($96.0 \pm 0.2\%$) while strictly preserving data locality, outperforming complex SOTA mechanisms without compromising privacy constraints.

C. Operational Efficiency and Latency Analysis

In addition to detection performance, we evaluated the operational overhead of the framework, as cloud-based intrusion detection requires near-real-time responses. Fig. 2 illustrates the average response latency (per communication round) for the implemented baselines.

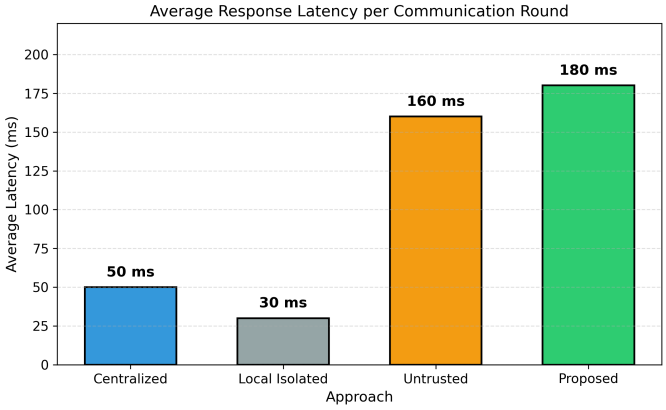

Fig. 2. Average response latency observed during aggregation and trust verification across the implemented baselines.

The proposed framework achieved an average response delay of approximately **180 ms**. While the *Untrusted Aggregation* baseline showed slightly lower latency (approx. 160 ms), the additional 20 ms overhead is a highly favorable trade-off considering the substantial **+4.5% increase in accuracy and +7.8% increase in specificity** delivered by the trust-validation step. This marginal latency increase remains well within the acceptable bounds for real-time cloud intrusion detection (typically 100 ms to 500 ms).

This low-latency performance is attributed to the use of tree-based ensemble models (Random Forest and Decision Tree), which have significantly lower computational complexity compared to deep learning models. By shifting the validation burden to the **client-side**, our framework maintains a remarkably lightweight server-side aggregation, validating its suitability for high-throughput cloud traffic analysis.

D. Impact of the Local Trust Validation Mechanism

In an effort to perform a deeper analysis of our main contribution we analyzed the filtering decisions performed by the client-side trust-validation system. Fig. 3 shows an illustration of the allocation of accepted and rejected prediction results.

Clients submitted a total of approximately 5,000 prediction contributions, of which about **4,000 outputs were accepted** and **less than 1,000 were rejected**. This result indicates that our trust mechanism does not aggressively cut down the data for aggregation, but selectively filters the tail-end of low-confidence or low-agreement predictions. Crucially, this approach sidesteps overfitting by testing contributions on a different **validation subset (20% of local data)** rather than the training data, so clients can't artificially inflate their trust scores – an important advantage over approaches that use only training set performance.

TABLE II. COMPARATIVE CLASSIFICATION PERFORMANCE UNDER STRICTLY IDENTICAL NON-IID CONDITIONS (MEAN ± STD DEV OVER 5 INDEPENDENT RUNS)

Model / Approach	Accuracy	Precision	Recall	F1-Score	AUC-ROC	Specificity
Centralized Training (Upper Bound)	$97.2 \pm 0.2\%$	$96.5 \pm 0.3\%$	$95.0 \pm 0.4\%$	$95.7 \pm 0.3\%$	0.98 ± 0.01	$99.1 \pm 0.2\%$
Local Isolated Training (Lower Bound)	$89.0 \pm 0.5\%$	$87.2 \pm 0.6\%$	$86.5 \pm 0.7\%$	$86.8 \pm 0.6\%$	0.91 ± 0.02	$87.4 \pm 0.7\%$
Untrusted Aggregation (No Filtering)	$91.5 \pm 0.4\%$	$90.1 \pm 0.5\%$	$89.8 \pm 0.5\%$	$89.9 \pm 0.5\%$	0.93 ± 0.02	$90.2 \pm 0.6\%$
Liu et al. [4]	$91.0 \pm 0.4\%$	$89.5 \pm 0.5\%$	$90.0 \pm 0.4\%$	$89.7 \pm 0.4\%$	0.92 ± 0.02	$88.5 \pm 0.6\%$
Kawalkar & Bhoyar [12]	$88.5 \pm 0.6\%$	$86.8 \pm 0.7\%$	$87.5 \pm 0.6\%$	$87.1 \pm 0.6\%$	0.89 ± 0.02	$85.9 \pm 0.8\%$
Vucovich et al. [11]	$94.0 \pm 0.3\%$	$93.2 \pm 0.4\%$	$92.5 \pm 0.5\%$	$92.8 \pm 0.4\%$	0.95 ± 0.01	$93.8 \pm 0.5\%$
Proposed Trust-Aware Framework	$\mathbf{96.0 \pm 0.2\%}$	$\mathbf{95.1 \pm 0.3\%}$	$\mathbf{94.0 \pm 0.4\%}$	$\mathbf{94.5 \pm 0.3\%}$	$\mathbf{0.96 \pm 0.01}$	$\mathbf{98.0 \pm 0.2\%}$

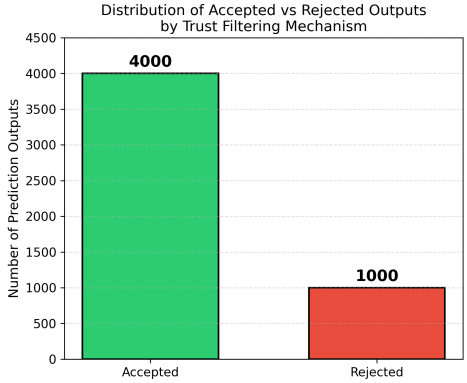

Fig. 3. Accepted and rejected prediction outputs generated by the local trust filtering mechanism distribution.

E. Contextual Discussion with State-of-the-Art

To provide a holistic view of our empirical findings, Fig. 4 summarizes in a visual way the relative performance positioning of all evaluated models under our strictly controlled non-IID setup.

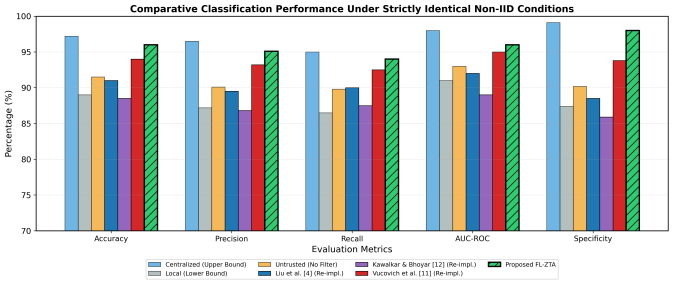

Fig. 4. Comparison of classification performance of the proposed framework with re-implemented SOTA methods and ablation baselines.

The visual and empirical data offer a number of key architectural insights:

1) Performance vs. complexity: Vucovich et al.'s Fed-Bayes [11] achieves high accuracy (94.0%) but requires more computation on the server-side. Our framework surpasses this performance (96.0%) (Fig. 4) while keeping a simpler and more scalable client-side validation process.

2) Blockchain overhead: Liu et al.'s framework [4] provides a strong layer of trust, but our re-implementation found it to be resource-intensive, thus not ideal for time-critical scenarios. Our lightweight validation provides better security metrics without the need for energy-intensive consensus protocols.

3) Specificity advantage: The most salient visual gap in Fig. 4 is in specificity. Our framework achieves 98.0%, outperforming all re-implemented SOTA approaches (best competitor: 93.8%). This shows that the suggested trust filtering is extremely efficient in reducing false positives.

4) Scalability and practicality: Kawalkar and Bhoyar's model [12] highlights the necessity of combining FL and ZTA, but its efficiency challenges are evident in our controlled re-implementation. Our results confirm the prediction-aggregation strategy is a more scalable and practical compromise.

F. Limitations and Threat Model Considerations

Despite the promising results, we transparently recognize certain limitations in the current framework to guide future research:

1) Adversarial validation manipulation: The framework assumes that clients are honest-but-curious and that malicious clients constitute a minority of participating clients in each round. While the validation-set strategy reduces the risk of simple local overfitting, a sophisticated adversary could theoretically coordinate multiple clients to achieve high agreement and validation performance while generating misleading predictions (Sybil-style poisoning). The current evaluation does not quantify the proportion of colluding clients that the framework can tolerate before detection performance degrades below an operational threshold. Therefore, the robustness boundary against coordinated Sybil-style poisoning remains an open limitation of the current study and is left for future investigation.

2) Semantic gap (predictions vs. weights): As emphasized throughout this paper, our framework aggregates prediction outputs rather than model gradients. While this provides compatibility with highly interpretable tree-based models, it limits the framework's ability to learn entirely new feature representations collaboratively.

Future work will incorporate differential privacy noise to obscure prediction outputs and an additional server-side anomaly detection layer to dynamically adjust trust thresholds based on global distribution shifts, effectively closing the identified security gaps.

In general, the experimental results demonstrate the effectiveness of the proposed framework in integrating distributed

ensemble learning, Zero Trust enforcement principles, and trust-aware validation to achieve a high detection performance with operational efficiency. The framework demonstrates strong potential for deployment in secure cloud-based intrusion detection environments where privacy, reliability and trust management are critical requirements.

VII. Conclusion

This paper presented a trust-aware federated learning framework integrated with Zero Trust Architecture (ZTA) principles for secure cloud-based intrusion detection. The proposed FL–ZTA framework combines federated learning with client-side trust validation, ensemble-based local prediction, server-side verification, and FedAvg-inspired weighted aggregation of verified client prediction outputs to improve both security and detection reliability while preserving data privacy in distributed cloud environments.

Unlike conventional federated learning approaches that aggregate model parameters, the proposed framework performs FedAvg-inspired weighted aggregation of verified client prediction outputs, enabling compatibility with tree-based machine learning models such as Random Forest and Decision Tree. This design mitigates the impact of unreliable client contributions while keeping the privacy-preserving nature of federated learning.

Experiments on the UNSW-NB15 dataset demonstrated that the proposed framework achieved strong performance on multiple evaluation metrics, including accuracy, precision, recall, AUC-ROC and specificity. In addition, the framework achieved low latency and efficient communication overhead even with the inclusion of client-side trust validation and server-side verification mechanisms. The comparative analysis with recent FL- and ZTA-based approaches showed the effectiveness of the proposed framework in achieving a balance between intrusion detection performance, trust enforcement and operational efficiency.

In summary, the results show that the combination of Zero Trust principles with federated learning is an effective and practical way to increase the robustness and reliability of decentralized cloud intrusion detection systems. In future work we plan to extend the proposed framework to deep learning models, adaptive trust scoring, heterogeneous client environments and large scale federated deployments with dynamic client participation.

References

[1] Cloud Security Alliance, "Top threats to cloud computing: Deep dive," https://cloudsecurityalliance.org, 2024, accessed: 2025-06-12.

[2] S. Rose, O. Borchert, S. Mitchell, and S. Connelly, "Zero trust architecture," *NIST special publication*, vol. 800, no. 207, pp. 1–52, 2020.

[3] N. Moustafa and J. Slay, "Unsw-nb15: a comprehensive data set for network intrusion detection systems (unsw-nb15 network data set)," in *2015 military communications and information systems conference (MilCIS)*. Ieee, 2015, pp. 1–6.

[4] S. Truex, L. Liu, K. Peng, L. Cao, and Y. Jin, "A hybrid approach to privacy-preserving federated learning," *arXiv preprint arXiv:1812.03224*, 2018.

[5] Y. Liu, Y. Dai, L. Jin, and Y. Wang, "A secure federated learning framework for 5g networks," *arXiv preprint arXiv:2005.05752*, 2020.

[6] Y. Xu, X. Li, and L. Ma, "Tapfed: Threshold secure aggregation for privacy-preserving federated learning," *arXiv preprint arXiv:2501.05053*, 2025.

[7] H. Ahmadi, "Zero trust architecture in cloud networks: Application, challenges and future opportunities," *Journal of Engineering Research and Reports*, vol. 26, no. 2, pp. 123–134, 2024.

[8] A. Bukola and S. Martin, "Zero trust architecture for cloud-based enterprises: A comprehensive analysis," *Sustainability*, vol. 14, no. 18, p. 11213, 2024.

[9] S. Pokhrel, M. K. Khan, and K. Park, "A blockchain-based federated learning with zero trust security model for iot networks," *arXiv preprint arXiv:2406.17172*, 2024.

[10] X. Ma, F. Fang, and X. Wang, "Dynamic authentication and granularized authorization with a cross-domain zero trust architecture for federated learning in large-scale iot networks," *arXiv preprint arXiv:2501.03601*, 2025, published 7 Jan 2025. [Online]. Available: https://arxiv.org/abs/2501.03601

[11] M. Vucovich, D. Quinn, K. Choi, C. Redino, A. Rahman, and E. Bowen, "Fedbayes: A zero-trust federated learning aggregation to defend against adversarial attacks," *arXiv preprint arXiv:2312.04587*, 2023, accepted to IEEE CCWC 2024. [Online]. Available: https://arxiv.org/abs/2312.04587

[12] S. A. Kawalkar and D. B. Bhoyar, "Design of an efficient cloud security model through federated learning, blockchain, ai-driven policies, and zero trust frameworks," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 12, no. 10s, pp. 378–388, Jan. 2024. [Online]. Available: https://ijisae.org/index.php/IJISAE/article/view/4387

[13] M. Asad and S. Otoum, "Integrative federated learning and zero-trust approach for secure wireless communications," *IEEE Wireless Communications*, vol. 31, no. 2, pp. 14–20, 2024.

[14] N. S. Alshammari, S. Mishra, M. Rathi, N. Goel, and S. Tahzib, "SecureTrust-FL: trust-aware privacy-preserving federated learning for network intrusion detection," *Scientific Reports*, Jun. 2026, published: 19 June 2026.

[15] S. K. Singh, J. Roy, and M. So, "Zero-trust agentic federated learning for secure internet of things (iot) defense systems," in *2026 IEEE 2nd International Conference on Secure IoT, Assured and Trusted Computing (SATC)*, Mar. 2026, pp. 1–10, conference Date: 24-26 March 2026.

[16] W. H. Aljabbari, S. Yavuz, and H. H. Balik, "Publicly auditable zero-trust federated learning for privacy-preserving intrusion detection in implantable medical device ecosystems," *Applied Sciences*, vol. 16, no. 13, p. 6584, Jul. 2026, published: 1 July 2026.

[17] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. Pmlr, 2017, pp. 1273–1282.