*symmetry*

MDPI

*Article*

# Symmetry-Aware Byzantine Resilience in Federated Learning via Dual-Channel Attention-Driven Anomaly Detection

**Yuliang Zhang [1], Jian Hou [1], Xianke Zhou [2], Linjie Ruan[1], Xianyu Luo [1] and Lili Wang [3,*]**

[1] School of Computer Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China; 2023210602022@mails.zstu.edu.cn (Y.Z.); changeleap@163.com (J.H.); 19012728453@163.com (L.R.); 2023210602014@mails.zstu.edu.cn (X.L.)
[2] Institute of Computer Innovation, Zhejiang University, Hangzhou 310058, China; xkzhou@zjuici.com
[3] School of System Design and Intelligent Manufacturing, Southern University of Science and Technology, Shenzhen 518055, China
* Correspondence: wangll@sustech.edu.cn

**Abstract**

Byzantine failures remain a critical threat to Federated Learning (FL), where malicious clients inject adversarial updates to disrupt global model convergence. From the perspective of symmetry, benign client updates typically exhibit statistical symmetry around the global consensus, whereas Byzantine attacks function as "symmetry-breaking" events that introduce skewness and distributional anomalies. Existing defenses often rely on unrealistic assumptions or fail to capture these asymmetric deviations under high-dimensional non-IID settings. In this paper, we propose a symmetry-aware Byzantine-resilient FL framework driven by a **D**ual-Channel **A**ttention-Driven **A**nomaly **D**etector (DAAD). Specifically, DAAD transforms inter-client behaviors into geometrically symmetric interaction matrices—encoding Gradient Cosine Similarities and Loss Euclidean Distances—to construct dual-channel spatial representations. These representations are processed via a Convolutional Neural Network (CNN) enhanced with Squeeze-and-Excitation (SE) attention blocks, which leverage the inherent symmetry of benign consensus to extract robust adversarial signatures. The detector is pre-trained offline on a synthetic dataset incorporating a diverse portfolio of simulated attacks (e.g., Gaussian noise and label flipping). Crucially, this pre-trained model is seamlessly embedded into the online FL loop to filter updates without requiring ground-truth labels. By jointly encoding client behaviors and learning cross-modal attack signatures, our framework enables reliable detection even when over half of the clients are Byzantine. Extensive experiments on MNIST, CIFAR-10, and FEMNIST datasets demonstrate that DAAD consistently outperforms existing robust aggregation baselines in both anomaly detection accuracy and global model performance, especially under high Byzantine ratios and non-IID conditions.

**Keywords:** Byzantine attack; federated learning; anomaly detection; attention mechanisms; statistical symmetry

Check for updates

Academic Editors: Xiaogang Qi and Sergei Odintsov

Received: 21 January 2026
Revised: 15 February 2026
Accepted: 7 March 2026
Published: 11 March 2026

**Copyright:** © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

## 1. Introduction

Federated Learning (FL) is a decentralized learning framework where $N$ clients collaboratively train a shared model by transmitting local updates, rather than raw data, to a central server. However, this distributed architecture is inherently vulnerable to malicious actors. Certain participants, known as Byzantine clients, may exhibit arbitrary adversarial behaviors due to targeted attacks or system failures. The fundamental goal of a robust

FL system is to accurately identify and isolate these malicious actors, enabling the central server to perform aggregation in a manner that ensures the resilience of the global model.

Recent research has extensively explored the security landscape of Federated Learning across various critical domains. For instance, in the Industrial Internet of Things (IIoT) [1], proposed a reliable anomaly detection strategy using hierarchical Federated Learning to address privacy and security threats in smart manufacturing. Beyond application-specific defenses, the complexity of adversarial attacks has evolved. Studies such as [2,3] have investigated sophisticated poisoning attacks in heterogeneous environments, highlighting the necessity for personalized learning frameworks to maintain robustness. Furthermore, to tackle the inherent non-IID data challenge in adversarial settings, advanced meta-learning approaches have been introduced to enhance model generalization and fault tolerance [4]. These works collectively underscore the growing need for resilient FL mechanisms that can operate effectively under diverse and hostile conditions.

At the core of this adversarial detection challenge lies the concept of symmetry. In a benign FL system, client updates typically exhibit statistical symmetry around a true global consensus. Byzantine attacks, whether targeted poisoning or untargeted noise, fundamentally act as "symmetry-breaking" events that distort this optimization landscape. Existing defenses often struggle to explicitly capture these asymmetric deviations in high-dimensional spaces, motivating the need for a novel detection paradigm.

### *1.1. Related Work*

A wide range of statistical methods have been proposed to mitigate Byzantine attacks in FL through robust aggregation rules. Krum [5] selects updates closest to the majority to ensure convergence. Trimmed-Mean [6] applies median and trimmed-mean operations to balance communication efficiency and error rates. Advanced estimators such as Bulyan [7], FLAME [8], and RFA [9] enhance resilience by combining robust statistics and adaptive filtering. Beyond static aggregation, FMAB [10] integrates multi-armed bandits to achieve $O(\log T)$ regret bounds under heterogeneous data, while geometric-median-based methods further reduce regret in federated linear bandit settings [11]. However, these methods generally operate under the assumption that benign updates are statistically cohesive. In highly non-IID environments or under high Byzantine ratios (e.g., >40%), the statistical boundary between benign heterogeneity and malicious behavior becomes blurred, causing these methods to either mistakenly reject benign gradients or fail to identify adaptive attacks.

To overcome the limitations of purely statistical methods, another line of defense relies on trusted validation data at the server to filter adversarial updates. FedSuper [12] injects a "shadow" dataset into each client and rejects updates exceeding an adaptive loss threshold. GAAvernor [13] employs reinforcement learning on a small quasi-validation set to adjust client weights. FLTrust [14] computes a reference gradient from a root dataset and down-weights client updates by ReLU-clipped cosine similarity. While effective, these approaches rely on a strong assumption: the availability of trusted, representative data at the server. This contradicts the privacy-preserving core of FL, as the server should not access raw data. Furthermore, obtaining a root dataset that covers the diverse distributions of all clients is often impractical in real-world cross-silo scenarios. If the root dataset is biased, the global model will inherit this bias, potentially excluding legitimate but underrepresented client contributions.

Hybrid cryptographic–aggregation frameworks synergistically combine security primitives with Byzantine-resilient mechanisms to establish multi-layered defense architectures. BRFL [15] uses blockchain and Proof-of-Pearson-Correlation-Coefficient (PPCC) consensus to elect honest nodes and a PSA rule to exclude outliers. SAMFL [16] integrates differential privacy and SMPC for encrypted update transmission. SF-CABD [17] pairs

homomorphic encryption with clustering-based anomaly detection for privacy-preserving aggregation. Although these methods offer enhanced security and privacy guarantees, they incur significant computational and communication overhead, limiting their scalability in resource-constrained edge networks. Moreover, cryptographic methods typically focus on secure transmission rather than analyzing the semantic validity of the updates, leaving them vulnerable to "poisoning via valid encryption" attacks.

Despite these advances, existing defenses are fundamentally constrained by what we term the Privacy-Robustness Tension. As analyzed above, statistical methods preserve zero-trust privacy but fail catastrophically under highly non-IID settings or when Byzantine ratios exceed the theoretical "honest majority" limit. Conversely, trust-based methods can tolerate higher Byzantine ratios but inherently violate the zero-trust privacy premise by requiring server-side root datasets. Meanwhile, cryptographic approaches secure data in transit but remain blind to semantic poisoning. Consequently, a critical research gap remains unresolved: How to achieve high Byzantine resilience (e.g., >50% attackers) under Non-IID conditions without relying on privacy-invasive trusted server data?

### *1.2. Contribution*

To explicitly bridge the research gap identified above, this work proposes a novel Symmetry-Aware Byzantine-resilient FL framework driven by a Dual-Channel Attention-Driven Anomaly Detector (DAAD). Unlike existing defenses that rely on static statistical rules or privacy-invasive trusted data, our approach fundamentally transforms anomaly detection from a low-dimensional statistical estimation task into a spatial pattern recognition problem.

The main contributions of this paper are summarized as follows:

- Novel Behavioral Imaging Paradigm: We introduce a mechanism to map inter-client interactions into a structured geometrically symmetric dual-channel spatial format. By encoding *Gradient Cosine Similarity* (capturing directional alignment) and *Loss Euclidean Distance* (capturing magnitude consistency) into image channels, we convert invisible statistical deviations into visible spatial anomalies. This enables the use of deep convolutional networks to capture high-order adversarial correlations that vector-based methods often miss.
- Attention-Driven Feature Selection: We propose a specialized Convolutional Neural Network (CNN) architecture enhanced with Squeeze-and-Excitation (SE) attention blocks. This mechanism allows the detector to adaptively recalibrate feature importance. Specifically, the model learns to prioritize the most informative channel—focusing on loss inconsistencies when gradient directions are confused, or vice versa—thereby ensuring robustness against adaptive attacks.
- Zero-Trust Resilience under High Byzantine Ratios: We demonstrate that our supervised learning paradigm, driven by a synthetic dataset of orthogonal attack primitives, achieves exceptional robustness. Extensive experiments show that DAAD maintains high detection accuracy even when over 50% of clients are Byzantine, effectively overcoming the theoretical "honest majority" limitation of traditional aggregation rules and resolving the Privacy-Robustness Tension by operating in a strictly zero-trust environment without server-side validation data.

## 2. Problem Formulation

As mentioned in Section 1, we consider a cross-silo Federated Learning setting (e.g., typically involving organizations like hospitals or financial institutions) where a central server coordinates a set $\mathcal{C}$ of $N$ clients (where $N$ is typically moderate, e.g., $N \leq 100$). In this scenario, the participants are not individual mobile devices (cross-device) but rather

large-scale institutions (silos) equipped with sufficient computational resources and stable network connections.

The primary characteristics of this setting are:

- Institutional Clients: The clients are organizations possessing distinct, private datasets. They are generally reliable in availability but may behave maliciously due to targeted external attacks or internal system compromises.
- Privacy Constraints: Due to regulatory restrictions (e.g., GDPR, HIPAA), raw data cannot leave the local institution; only model updates are exchanged.
- Zero-Trust Server: The central server is responsible for aggregation but is not fully trusted to access any raw validation data or privacy-sensitive information.

This practical setup serves as the basis for our formal problem definition below.

The clients are labeled from $1, \ldots, N$. The local model parameters of each client $i$ is defined by $\mathbf{w}_i$. Specifically, the Byzantine model is defined as follows [18]:

$$\hat{\mathbf{w}}_i = \begin{cases} \mathbf{w}_i, & \text{if the } i\text{th client is benign,} \\ *, & \text{otherwise.} \end{cases} \tag{1}$$

Here, $\hat{\mathbf{w}}_\mathbf{i}$ denotes the parameters transmitted by client $i$, and $*$ represents arbitrary values. Each client $i$ holds a local dataset $\mathcal{D}^i = \left\{ \left( \mathbf{x}_j^i, y_j^i \right) \right\}_{j=1}^{n_i}$ with $n_i$ samples. Superscript $^{(i)}$ denotes quantities specific to client $i$ and subscript $_{(j)}$ indexes samples within client $i$'s dataset, where $\mathbf{x}_j^i \in \mathbb{R}^{d_{\text{in}}}$ denotes the $d_{\text{in}}$-dimensional input features and $y_j^i \in \mathcal{Y}$ is the corresponding label from label space $\mathcal{Y}$. The global model parameters are denoted as $\mathbf{w} \in \mathbb{R}^d$ where $d$ represents the parameter dimension. The server aims to learn a global model $\mathbf{w}^* \in \mathbb{R}^d$ through Byzantine-resilient aggregation:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \sum_{i \in \mathcal{C} \setminus \hat{\mathcal{B}}(t)} \frac{n_i}{n(t)} F_i(\mathbf{w}) \tag{2}$$

where $\hat{\mathcal{B}}(t) \subset \mathcal{C}$ represents the estimated Byzantine clients identified by our anomaly detector at communication round $t$, $n_i$ is the number of local training samples at client $i$, and $n(t) = \sum_{i \in \mathcal{C} \setminus \hat{\mathcal{B}}(t)} n_i$ calculates the effective total sample size of benign clients. The local empirical risk $F_i(\mathbf{w})$ is defined as $F_i(\mathbf{w}) = \frac{1}{n_i} \sum_{j=1}^{n_i} \ell(\mathbf{w}; \mathbf{x}_j^i, y_j^i)$, where $\ell$ is the loss function. The goal is to design a framework to achieve resilience against Byzantine attacks, i.e.,

$$\hat{\mathcal{B}}(t) = \{ i \in \mathcal{C} \mid s_i(t) > \tau \} \tag{3}$$

where $s_i(t) \in [0, 1]$ is the anomaly score from our dual-channel detector, and $\tau = 0.5$ is the empirical threshold.

## 3. Overall Framework

This section briefly presents our unified framework for Byzantine-resilient FL based on DAAD, as shown in Figure 1. Figure 1 illustrates the framework, which operates in two distinct phases. Phase 1 (Offline Supervised Pre-training): The DAAD is trained on a synthetic dataset containing mixed attack patterns to learn unified adversarial signatures. Phase 2 (Online Unified Inference): The pre-trained model is directly embedded into the FL server to filter malicious updates in real-time, enabling robust defense without requiring ground-truth labels during the online process.

The complete execution flow is formalized in Algorithm 1. During online execution at round $t$, each client's update is represented by a tuple of features: (i) the current local model weights $\mathbf{w}_i(t+1)$, (ii) the inter-client gradient $\mathbf{g}_i(t)$, and (iii) the loss $\mathbf{l}_i(t)$. These

features are passed to the DAAD module, which is designed to capture both structural and statistical deviations indicative of adversarial behavior. The DAAD module outputs a sequence of anomaly scores $\mathbf{s}(t) = \{s_i(t)\}$, where each score $s_i$ quantifies the likelihood that client $i$ is Byzantine. These scores are subsequently used to inform robust aggregation by down-weighting or excluding suspicious client updates from the global model.

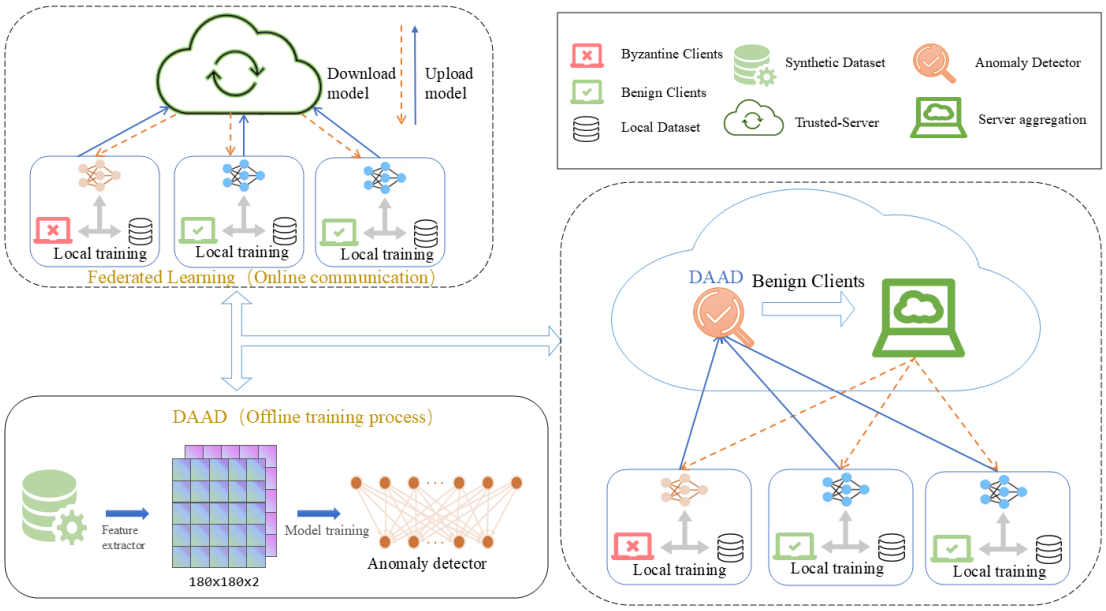



**Figure 1.** Proposed framework of Byzantine-Resilient FL-DAAD. The federated communication phase under Byzantine attacks is employed to create an anomalous dataset (top-left). The dataset is used for the offline training of the DAAD (bottom-left). In the right half, the offline-trained anomaly detector will be embedded into the server to identify and isolate Byzantine clients.

**Algorithm 1** DAAD-based Byzantine-Resilient FL

**Require:** Clients $\mathcal{C}$, Initial model $\mathbf{w}^{(0)}$, Threshold $\tau$, Learning rate $\eta$.
**Ensure:** Global model $\mathbf{w}^*$.

**Server Execution:**
**for** round $t = 0, 1, \ldots, T-1$ **do**
  Broadcast $\mathbf{w}^{(t)}$ to all clients.
  Receive $\{\mathbf{w}_i, \mathbf{g}_i, l_i\}_{i \in \mathcal{C}}$ via ClientUpdate in parallel.
  **Anomaly Detection:**
  Construct image $\mathbf{I}^{(t)}$ using gradient similarity $\mathbf{C}$ and loss distance $\mathbf{E}$.
  Get scores $\mathbf{s}^{(t)} \leftarrow f_{\text{DAAD}}(\mathbf{I}^{(t)})$.
  Filter benign set $\hat{\mathcal{C}}^{(t)} \leftarrow \{i \in \mathcal{C} \mid s_i^{(t)} \leq \tau\}$.
  **Robust Aggregation:**
  $\mathbf{w}^{(t+1)} \leftarrow \sum_{i \in \hat{\mathcal{C}}^{(t)}} \frac{n_i}{n_{total}} \mathbf{w}_i$
**end for**
**function** ClientUpdate$(i, \mathbf{w}^{(t)})$
  Update $\mathbf{w}_i$ via SGD on local dataset $\mathcal{D}_i$ for $E$ epochs.
  Compute $\mathbf{g}_i \leftarrow \mathbf{w}^{(t)} - \mathbf{w}_i$ and loss $l_i$.
  **return** $\mathbf{w}_i, \mathbf{g}_i, l_i$
**end function**

To ensure a rigorous security analysis, we formally define the Byzantine adversary model based on the following dimensions:

- Adversary Goal: The primary objective of the adversary is to degrade the performance of the global model. This includes Untargeted Attacks (preventing convergence or

reducing accuracy on the main task) and Targeted Attacks (injecting backdoors to misclassify specific inputs, e.g., label flipping).

- Capabilities (Byzantine Faults): We assume a fixed fraction of clients are malicious, denoted by the set $\mathcal{B} \subset \mathcal{C}$, where $|\mathcal{B}| < N$. These Byzantine clients can:
  1. Arbitrary Updates: Send any vector $\hat{w}_i$ to the server, unrelated to their local training data.
  2. Collusion: Malicious clients can communicate and synchronize their updates to form a coordinated attack (e.g., sending identical malicious gradients to evade distance-based defenses).
  3. Local Control: They have full control over their local training process, including hyperparameters and local datasets $\mathcal{D}_i$.
- Knowledge: The adversary has White-Box Access to the global model parameters $w^{(t)}$ broadcast by the server. However, regarding the defense mechanism, we assume a Gray-Box setting: the adversary knows the aggregation algorithm exists but does not have access to the real-time internal parameters or the random seed of the DAAD detector.
- Constraints: The adversary cannot compromise the central server (which is assumed to be honest-but-curious) nor modify the communication channels of benign clients. The number of malicious clients is bounded such that benign clients may not necessarily form a majority in extreme cases (we test up to 60% Byzantine ratios), testing the limits of the defense.

## 4. Dual-Channel Attention-Driven Anomaly Detector

The anomaly detector DAAD serves as the cornerstone of our Byzantine-resilient framework, as shown in Figure 1 and Section 4.3.

### *4.1. Dual-Channel Behavior Imaging*

Client interactions are encoded into visual representations through a three-step geometric transformation designed to capture the intrinsic symmetry of benign consensus.

First, directional alignment between clients is quantified using gradient cosine similarity, which measures the consistency of parameter updates across the federation. For any pair of clients $i, j \in \mathcal{C}$, the cosine similarity between their gradient vectors $g_i, g_j \in \mathbb{R}^d$ is defined as:

$$C_{ij} = \frac{\langle g_i, g_j \rangle}{\|g_i\| \|g_j\|} \in [-1, 1], \tag{4}$$

where $g_i(t) = \nabla F_i(w(t))$ denotes the local gradient computed at round $t$, and $\|\cdot\|$ is the $l_2$ norm. A higher value of $C_{ij}$ indicates stronger directional consensus between clients $i$ and $j$.

Second, optimization trajectory divergence is quantified through pairwise loss distances. For any pair of clients $i, j \in \mathcal{C}$, the pairwise Euclidean distance between training losses $l_i, l_j \in \mathbb{R}^+$ is measured as follows:

$$E_{ij} = \|l_i - l_j\| \in \mathbb{R}^+. \tag{5}$$

The value of $E_{ij}$ reflects the degree of optimization divergence, capturing inconsistencies in learning dynamics.

Geometric Symmetry Property: Crucially, by definition, both metrics satisfy the commutative property, implying that $C_{ij} = C_{ji}$ and $E_{ij} = E_{ji}$. Consequently, the resulting correlation matrices $\mathbf{C} \in \mathbb{R}^{N \times N}$ and $\mathbf{E} \in \mathbb{R}^{N \times N}$ are geometrically symmetric matrices. This symmetry is fundamental to our behavioral imaging paradigm, as it mathematically

represents the undirected topology of client interactions. In a benign environment, these matrices exhibit a structured symmetry reflecting global consensus. Byzantine attacks, however, manifest as "symmetry-breaking" outliers that disrupt this harmonious structure.

Third, the relational matrices are transformed through structured geometric expansion that preserves this interaction topology. Both matrices first undergo unified normalization for contrast enhancement:

$$\tilde{X}_{ij} = 255 \times \frac{X_{ij} - \min(X)}{\max(X) - \min(X)}, \quad \text{where } X \in \{\mathbf{C}, \mathbf{E}\}. \tag{6}$$

Then, to match the receptive field requirements of deep CNNs, each matrix element undergoes non-interpolative upsampling via $9 \times 9$ uniform tiling. This process systematically scales the spatial resolution from $N \times N$ to $9N \times 9N$. This deterministic expansion achieves two critical objectives: (1) amplifying microscopic interaction patterns through macroscopic spatial representation and (2) strictly maintaining the original geometric symmetry without introducing interpolation artifacts. The final dual-channel image $I \in \mathbb{Z}^{9N \times 9N \times 2}$ combines these symmetric features through depth-wise stacking, enabling the SE-enhanced CNN to effectively distinguish between the symmetric patterns of benign collaboration and the asymmetric artifacts of adversarial interference.

### *4.2. Attention-Driven Anomaly Detector*

Our architecture enhances EfficientNet by integrating Squeeze-and-Excitation (SE) blocks [19] at three critical stages to enable hierarchical attention tailored for federated anomaly detection.

As illustrated in Figure 2, SE blocks are introduced at three strategic points: after the initial convolution (Stage 1) to mitigate the influence of noisy gradient patterns, following the 24-channel MBConv block (Stage 2) for intermediate feature selection, and subsequent to the 40-channel MBConv block (Stage 4) to enhance the representation of collusion signatures.

The network takes a dual-channel input image $\mathbf{I} \in \mathbb{Z}^{9N \times 9N \times 2}$, encoding inter-client gradient similarities and loss divergences. Through a sequence of SE-modulated convolutional layers, feature maps are progressively refined:

$$\mathbf{F}^{(\ell)} = \text{MBConv}(\hat{\mathbf{F}}^{(\ell-1)}), \quad \hat{\mathbf{F}}^{(\ell)} = \mathbf{F}^{(\ell)} \odot \mathbf{s}^{(\ell)}, \tag{7}$$

where $\mathbf{s}^{(\ell)}$ is the SE-generated attention vector at layer $\ell$. The final representation is pooled and projected to yield the anomaly score sequence $\mathbf{s} = \{s_i\}_{i \in \mathcal{C}}$, indicating the Byzantine likelihood of each client $i$.

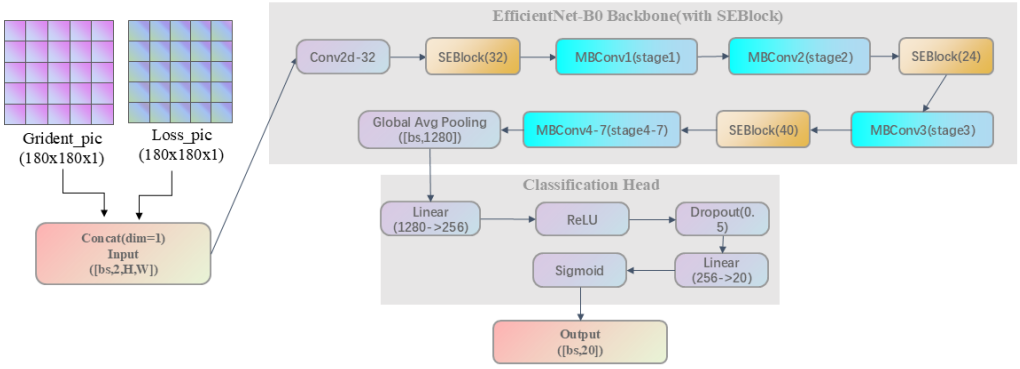



**Figure 2.** SE-enhanced EfficientNet.

### *4.3. Spatial Encoding via Dual-Channel Mapping*

In FL environments, client behaviors evolve in high-dimensional, non-Euclidean spaces, where conventional statistical techniques struggle to capture subtle or coordinated Byzantine deviations. To address this, we propose a dual-channel image encoding that maps inter-client gradient similarities and loss divergences into a structured spatial format, enabling convolutional models to extract robust representations of client behavior.

Let $\{\mathbf{g}_i\}_{i=1}^N$ and $\{l_i\}_{i=1}^N$ denote the gradients and training losses reported by clients at round $t$, forming a dynamic graph $G_t = (V, E_t)$, where $E_t(i,j)$ reflects behavioral similarity. To overcome the limitations of dynamic graphs and GNNs (unstructured data, $O(N^2)$ complexity), we define an image mapping:

$$\phi : G_t \mapsto \mathbf{I}(t) \in \mathbb{Z}^{9N \times 9N \times 2}, \quad \mathbf{I}(t)(i,j,:) = \left[C_{ij},\ E_{ij}\right], \tag{8}$$

where $C_{ij} = \frac{\langle \mathbf{g}_i, \mathbf{g}_j \rangle}{\|\mathbf{g}_i\| \|\mathbf{g}_j\|}$ captures directional alignment and $E_{ij} = \|l_i - l_j\|_2$ encodes optimization divergence.

This dual-channel design captures both local parameter dynamics (via gradients) and global convergence behavior (via losses), enabling robust detection of clients that deviate in either space. As shown in Figure 3a, the t-SNE projection of the client gradients reveals distinct clusters. While high non-IID degrees typically disperse benign client distributions, the clear separation observed here is attributed to the significant behavioral divergence induced by the simulated attacks (e.g., high-variance noise or directional opposition). It is worth noting that in scenarios with extreme non-IID data where gradient-based separation (Channel 1) might blur, the proposed framework leverages the complementary loss divergence (Channel 2, Figure 3b) to maintain detection capability. We argue that these two signals are jointly sufficient to characterize meaningful Byzantine behaviors because they represent orthogonal dimensions of model updates: direction and magnitude. Specifically, gradient cosine similarity captures directional conflicts (e.g., sign flipping, reverse gradient), while loss divergence captures magnitude anomalies and optimization impact (e.g., noise injection, model replacement). Any attack attempting to disrupt the global model must manifest as a deviation in either direction or magnitude. Stealthy attacks that restrict their updates to be statistically indistinguishable in both dimensions effectively reduce their own malicious impact, rendering them benign in practice. Both visualizations confirm that our dual-channel encoding captures structured differences that enable the detector to distinguish adversaries even when raw metric boundaries are complex.

### *4.4. Theoretical Motivation via t-SNE*

To theoretically analyze the feature separability of our representations, we utilize t-Distributed Stochastic Neighbor Embedding (t-SNE). First, t-SNE captures the local structure of the high-dimensional data by converting pairwise distances into conditional probabilities, which are subsequently symmetrized into joint probabilities $p_{ij}$, as formally defined in Equation (9):

$$p_{j|i} = \frac{\exp\left(-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\|\mathbf{x}_i - \mathbf{x}_k\|^2 / 2\sigma_i^2\right)}, \quad p_{ij} = \frac{p_{i|j} + p_{j|i}}{2N}, \tag{9}$$

where $\sigma_i$ is the variance of the Gaussian distribution centered on data point $\mathbf{x}_i$.

To alleviate the crowding problem in the embedded space, t-SNE models the low-dimensional similarities $q_{ij}$ utilizing a Student-$t$ distribution with one degree of freedom, as formulated in Equation (10):

$$q_{ij} = \frac{(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}}{\sum_{k \neq l}(1 + \|\mathbf{y}_k - \mathbf{y}_l\|^2)^{-1}}. \tag{10}$$

Finally, the optimal low-dimensional embeddings $\{\mathbf{y}_i\}$ are learned by minimizing the Kullback–Leibler (KL) divergence between the joint probability distributions $P$ and $Q$. This objective function, which optimally aligns the high- and low-dimensional similarities, is given by Equation (11):

$$\mathrm{KL}(P\|Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}. \tag{11}$$

To objectively evaluate the separability, we calculated the Silhouette Coefficient (SC) of the embeddings. The resulting score is 0.23. It is worth noting that in non-IID FL settings, benign clients exhibit high statistical diversity, leading to naturally dispersed clusters and lower absolute SC values compared to IID scenarios. Crucially, the positive score indicates that the separation between benign and malicious groups exceeds the internal variance of the benign group. This confirms that the dual-channel encoding effectively preserves the discriminative structure required for detection, preventing malicious updates from blending indistinguishably into the benign heterogeneity.

This probabilistic embedding supports our dual-channel design: gradients and losses encode orthogonal behavioral views. Their spatial fusion into a structured tensor facilitates effective learning of collusive anomalies, enabling convolutional models to capture high-order deviations undetectable in unstructured metric spaces.

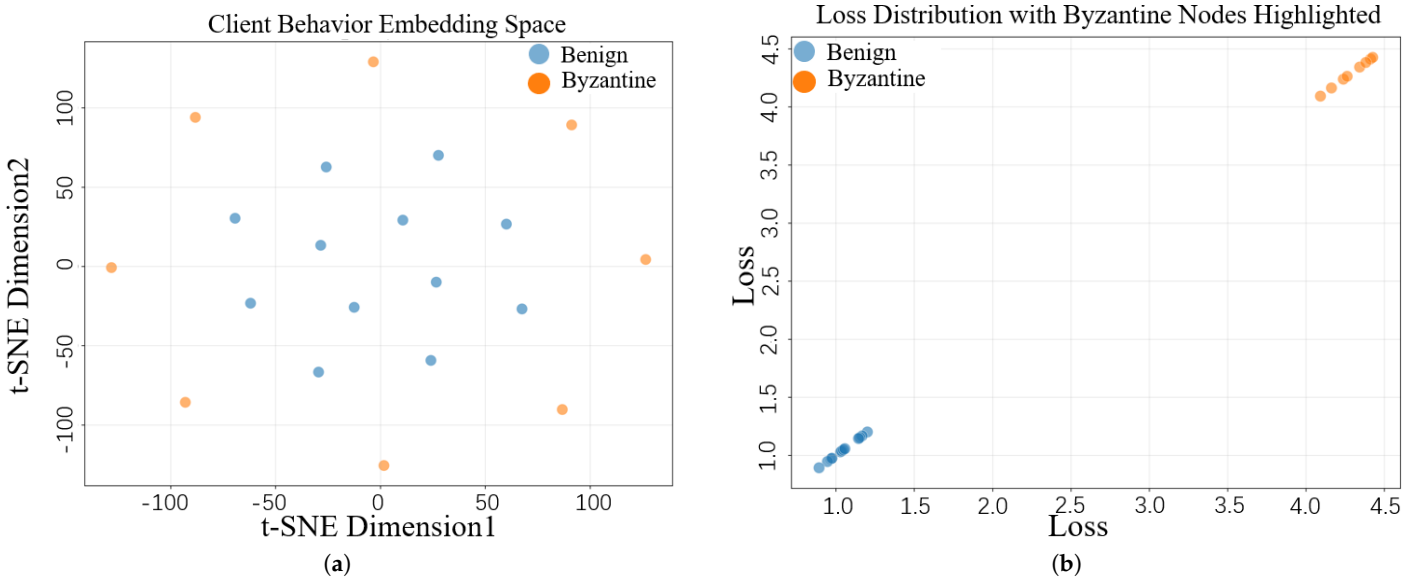



**Figure 3.** The blue markers represent benign clients forming cohesive clusters, while orange denotes Byzantine clients. (**a**) t-SNE visualization of the client's gradient; (**b**) Visualization of the client's loss.

### *4.5. Scalability and Practical Deployment*

The proposed dual-channel encoding maps client interactions into a spatial resolution of $9N \times 9N$, and while this high-resolution representation is critical for preserving fine-grained adversarial artifacts (e.g., subtle gradient perturbations), it introduces considerations for scalability in massive-scale networks. We address the practical deployment from two perspectives:

#### 4.5.1. Target Scenario: Cross-Silo FL

Our framework is primarily designed for cross-silo Federated Learning scenarios (e.g., inter-institutional collaborations among hospitals or banks). In these settings, the number

of clients $N$ is typically moderate (e.g., $N \leq 100$), but the requirement for security and model integrity is paramount. For such scales, the server-side computational overhead of processing high-resolution behavioral images is negligible compared to the resource-intensive global model aggregation.

#### 4.5.2. Extension to Large-Scale Networks

For cross-device scenarios where the number of clients exceeds the design capacity (e.g., $N \gg 20$), we introduce a divide-and-conquer grouping strategy. Clients can be randomly partitioned into subgroups of size $N_{sub}$ (matching the detector's pre-trained input size, e.g., $N_{sub} = 20$). The DAAD is then applied to each subgroup in parallel to filter anomalies locally within the group. Regarding the zero-padding strategy for dynamic client dropouts, we note that zero-padding is a standard operation in Convolutional Neural Networks (CNNs) for handling variable-size inputs. Since the convolutional filters share weights, the network learns to treat zero-padded regions as "information-less" backgrounds rather than meaningful features. Consequently, this padding does not introduce statistical bias or distort the detection of active clients, as confirmed by the robust performance of CNNs in similar sparse-input tasks (e.g., text classification with padding [20,21]). Conversely, to handle dynamic client dropouts where the active count is less than $N_{sub}$, zero-padding is applied to the behavioral matrices to maintain consistent input dimensions for the convolutional network.

## 5. Experiments

In this section, we present a comprehensive analysis of the proposed method's performance. We assess DAAD from four perspectives: anomaly detection accuracy, global model robustness, computational efficiency, and component contribution.

### *5.1. Experimental Setup*

#### 5.1.1. Federated Environment and Baselines

The FL framework follows a client-server architecture where a central server coordinates $N = 20$ clients. Each client conducts local training on its private dataset and periodically transmits model updates along with behavioral metrics to the server over secure communication channels. We explicitly set the number of clients $N = 20$ to simulate a representative Cross-Silo FL environment. The local training epoch is set to $E = 1$ to evaluate the defense mechanism under high-frequency aggregation scenarios. This setting poses a stricter challenge for anomaly detection compared to large $E$, as it requires the detector to identify malicious updates before they accumulate significant divergence.

We selected FedAvg [22], Krum [5], Trimmed Mean [6], and FLTrust [14] as quantitative baselines. In addition, we qualitatively compared DAAD with emerging state-of-the-art methods such as FLAME [8] and DnC [23], while these recent approaches rely on clustering techniques (e.g., HDBSCAN) or spectral analysis, they typically operate on raw gradient vectors. In contrast, DAAD leverages deep convolutional feature extraction to identify subtle behavioral anomalies statistically indistinguishable in the raw vector space.

#### 5.1.2. Attack Scenarios and Synthetic Data

We evaluated efficacy against three orthogonal attack primitives: Reverse Gradient (directional opposition), Gaussian Noise (magnitude corruption), and Label Flipping (semantic misalignment). Furthermore, we simulated Collaborative Attacks where multiple malicious clients ($f = 4$ out of $N = 20$) execute synchronized strategies, implicitly evaluating robustness against collusion.

To address the concern regarding the alignment between synthetic and real-world adversarial behaviors, our offline training dataset was generated based on these fundamental primitives with randomized parameters. Specifically, for Gaussian Noise, the variance $\sigma$ was sampled from $[0.1, 5.0]$; for Label Flipping, the ratio was varied between 10% and 100%. This parameter randomization ensures that the detector learns to recognize underlying distributional shifts rather than overfitting to fixed signatures. We employed a Mixed-Strategy Training approach, enabling the detector to generalize to unseen variations of adversarial behaviors.

The default experimental parameter settings are summarized in Table 1.

**Table 1.** Default parameter settings for the experiments.

| Parameter | Default Value |
|---|---|
| Number of Clients ($N$) | 20 |
| Global Communication Rounds | 1000 |
| Local Epochs | 1 |
| Batch Size | 32 |
| Learning Rate | $1 \times 10^{-4}$ |
| Image Resolution | $180 \times 180$ |
| Gaussian Noise Std ($\sigma$) | 2.5 |
| Label Flipping Probability | 0.5 |
| Level of non-IID data ($\alpha$) | 0.5 |
| Byzantine ratio ($\rho$) | 0.2 |

### *5.2. Anomaly Detection Performance*

We first evaluate the standalone performance of the DAAD module using False Positive Rate (FPR), False Negative Rate (FNR), and F1-error.

#### 5.2.1. Cross-Attack Detection

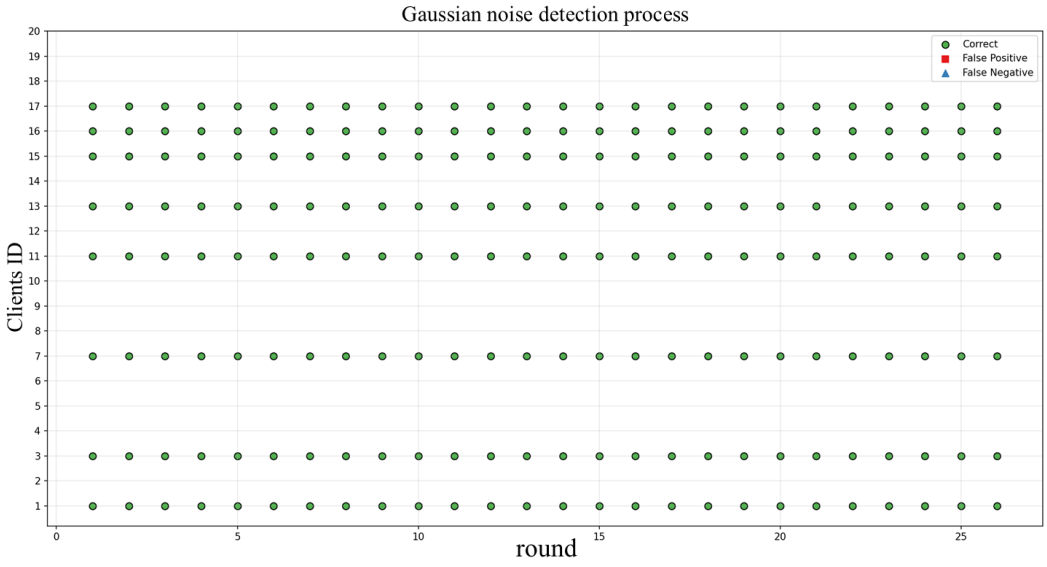
The following Table 2 presents the experimental results on the MNIST and CIFAR-10 datasets, indicating two key findings: (1) explicit attacks (Gaussian/reverse-gradient) enable more stable early-round detection compared to covert label manipulation and (2) adaptive attacks eventually achieve comparable final-phase performance through progressive learning. The results demonstrate that explicit adversarial strategies consistently maintain reliable detection across training phases, with minimal impact on global model convergence. In contrast, label-flipping attacks require extended training to distinguish adversarial patterns from legitimate data heterogeneity, particularly in high-dimensional non-IID feature spaces where detection boundaries require iterative refinement.

**Table 2.** Anomaly detection performance comparison.

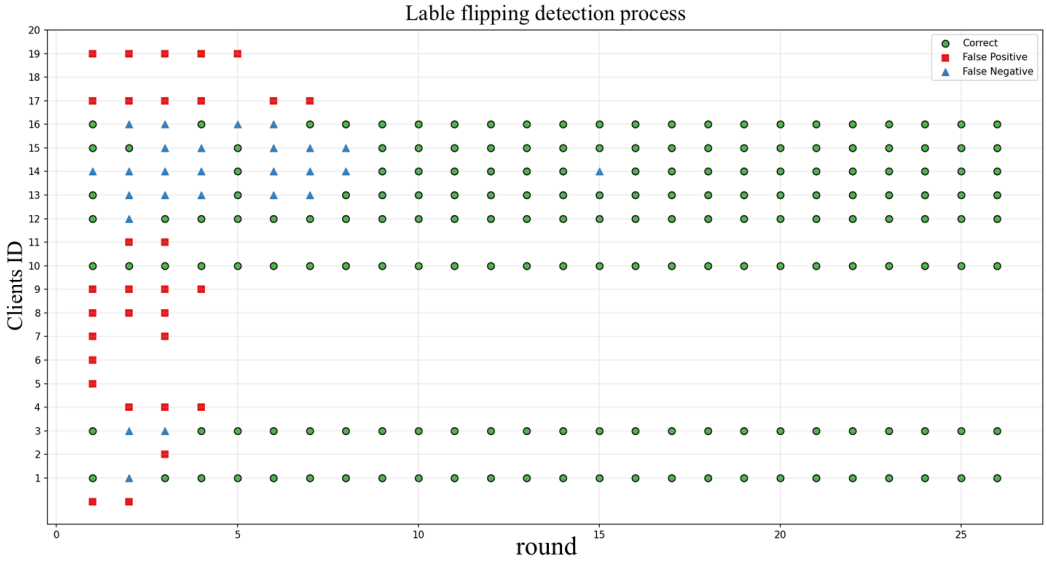
| Attack | $\rho$ | MNIST | | | CIFAR-10 | | |
|---|---|---|---|---|---|---|---|
| | | FDR | FNR | $F_1$-e | FDR | FNR | $F_1$-e |
| Gaussian | 20% | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 40% | 0.006 | 0.000 | 0.003 | 0.006 | 0.001 | 0.004 |
| | 60% | 0.013 | 0.000 | 0.010 | 0.003 | 0.000 | 0.001 |
| Reverse | 20% | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 40% | 0.011 | 0.000 | 0.007 | 0.042 | 0.297 | 0.038 |
| | 60% | 0.002 | 0.030 | 0.018 | 0.116 | 0.052 | 0.121 |
| Label Flip | 20% | 0.045 | 0.000 | 0.033 | 0.152 | 0.141 | 0.162 |
| | 40% | 0.015 | 0.008 | 0.013 | 0.040 | 0.229 | 0.150 |
| | 60% | 0.048 | 0.040 | 0.046 | 0.267 | 0.029 | 0.174 |

5.2.2. Temporal Detection Patterns

Figure 4 presents temporal detection pattern visualizations. The experimental evaluation focuses on two representative attack scenarios under the MNIST benchmark: Gaussian noise injection and label flipping with 40% Byzantine clients.




(**a**)




(**b**)

**Figure 4.** Temporal detection patterns for different attacks on MNIST (40% Byzantine nodes). Colored markers track identification states across federated rounds: green circles (true positives), red squares (false negatives), and blue triangles (false positives). The vertical alignment of green markers demonstrates immediate anomaly recognition. (**a**) Gaussian Noise Attack; (**b**) Label Flipping Attack.

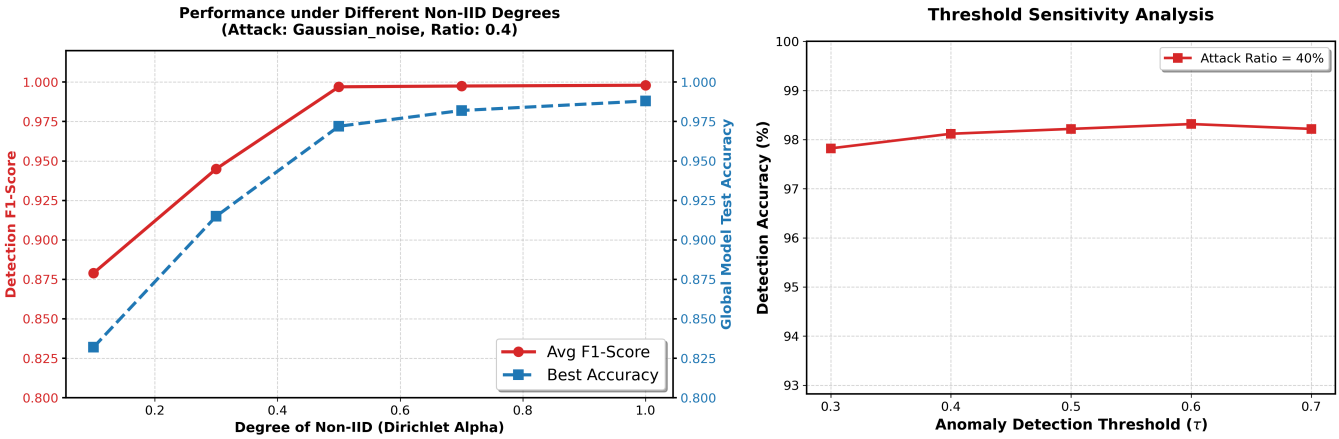
For Gaussian noise attacks, our framework achieves high-precision detection from the first round (Figure 4a), sustaining a 100% true positive rate by leveraging gradient-magnitude anomalies. In the label-flipping attack, accuracy refines in three phases (Figure 4b): (1) confusion (Rounds 1–10), (2) convergence (Rounds 11–15), and (3) stable detection (Rounds 16+), showcasing federated attention learning's separation of adversarial patterns. This DAAD method unites instant defense against simple perturbations with

adaptive mitigation of complex attacks, offering both localized node tracking and global pattern analysis for robust Byzantine resilience.

#### 5.2.3. Sensitivity and Robustness Analysis

To validate the reliability of DAAD, we analyze the impact of data heterogeneity and threshold settings.

Impact of Non-IID Degrees: We evaluated DAAD's detection performance under varying degrees of data heterogeneity by adjusting the Dirichlet concentration parameter $\alpha \in \{0.1, 0.3, 0.5, 0.7, 1.0\}$. This specific evaluation was conducted on the MNIST dataset, subjecting the system to a Gaussian noise attack with a significant Byzantine ratio of 40% ($\rho = 0.4$). As anticipated, the detection performance exhibits a positive correlation with the homogeneity of the data distribution (Figure 5a). Specifically, in near-IID scenarios ($\alpha = 1.0$), DAAD achieves near-perfect performance with an F1-score of 0.9980 and an accuracy of 0.9880. Importantly, for moderate heterogeneity ($\alpha \geq 0.5$), which represents typical FL scenarios, the system maintains high robustness (F1 $> 0.99$). As the environment becomes extremely heterogeneous ($\alpha = 0.1$), the F1-score decreases to 0.8790. This decline is attributable to the fact that extreme data divergence blurs the statistical boundary between benign heterogeneity and malicious deviations, increasing the difficulty of anomaly separation. Nevertheless, the detector remains functional even under such severe conditions.




(**a**) Impact of Non-IID degrees ($\alpha$)

(**b**) Sensitivity to threshold $\tau$

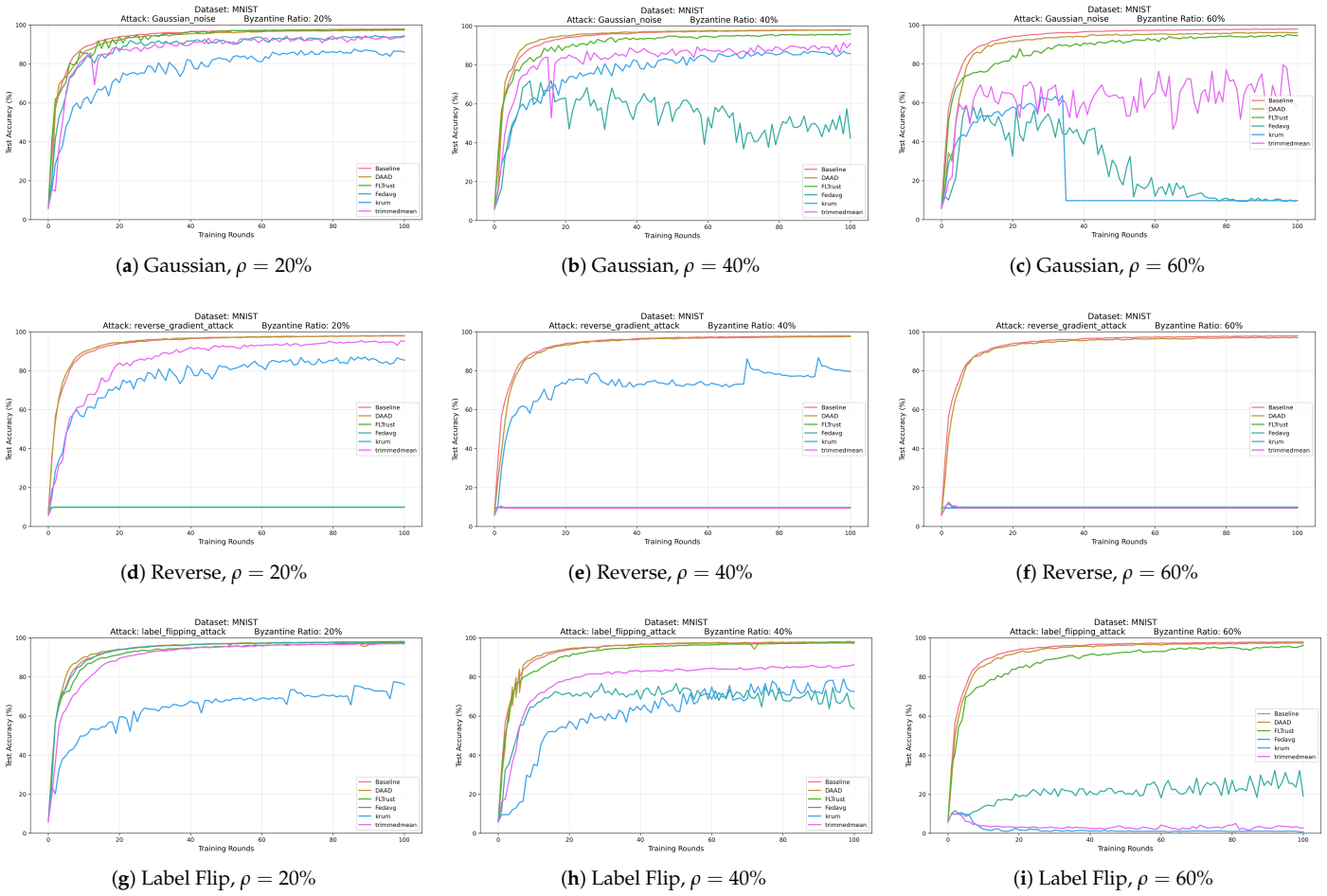
**Figure 5.** Sensitivity and robustness analysis. (**a**) Detection performance (F1-score and Accuracy) under varying degrees of data heterogeneity (Dirichlet parameter $\alpha$); (**b**) detection accuracy fluctuation with respect to the anomaly detection threshold $\tau$.

Parameter Sensitivity ($\tau$): Figure 5b illustrates the detection accuracy as the threshold $\tau$ varies from 0.3 to 0.7. The minimal fluctuation (<1%) confirms that the sigmoid activation effectively polarizes predictions, creating a wide safety margin that makes the system insensitive to precise threshold tuning.

### *5.3. Global Model Robustness Evaluation*

To comprehensively assess the robustness of our proposed DAAD-based framework, we conducted controlled experiments across a range of adversarial scenarios. Specifically, we vary three key factors: (i) the Byzantine Ratio, (ii) the Attack Type, and (iii) the Dataset. For each experimental configuration, the proposed method is compared with the existing robust aggregation baselines as mentioned in Section 5.1. The evaluation focuses on global model accuracy across communication rounds, with FedAvg serving as the baseline scenario in the absence of attacks.

Figure 6 visualizes the convergence on MNIST, while Tables 3 and 4 detail the final accuracy for CIFAR-10 and FEMNIST.




(**a**) Gaussian, $\rho = 20\%$ (**b**) Gaussian, $\rho = 40\%$ (**c**) Gaussian, $\rho = 60\%$

(**d**) Reverse, $\rho = 20\%$ (**e**) Reverse, $\rho = 40\%$ (**f**) Reverse, $\rho = 60\%$

(**g**) Label Flip, $\rho = 20\%$ (**h**) Label Flip, $\rho = 40\%$ (**i**) Label Flip, $\rho = 60\%$

**Figure 6.** Accuracy curves comparison under controlled variables (MNIST). Each row represents an attack type with varying Byzantine ratios $\rho$: (**a**–**c**) Gaussian noise attack; (**d**–**f**) Reverse gradient attack; (**g**–**i**) label flipping attack.

A separate DAAD detector is trained for each dataset to ensure fair adaptation to distributional characteristics. Empirical results demonstrate that our approach maintains stable convergence even under extreme conditions, such as 60% Byzantine participation, while some aggregation-based defenses (e.g., FLTrust or Krum) may outperform in benign settings, our method excels in consistently filtering adversarial updates, thus preserving model utility across all threat levels. Notably, DAAD is particularly effective in enabling continued learning under stealthy or adaptive attack types (e.g., label-flipping), where conventional methods struggle. These findings underscore the adaptability and resilience of our spatial encoding and attention-driven detection strategy.

It is worth noting that in scenarios with a low Byzantine ratio (e.g., $\rho = 20\%$), FLTrust occasionally yields slightly higher accuracy than DAAD. We attribute this to FLTrust's reliance on a clean server-side "root of trust" dataset, which provides a stable anchor. In contrast, DAAD operates in a strictly "zero-trust" setting. Furthermore, DAAD adopts a conservative filtering strategy to maximize recall against malicious nodes, potentially misclassifying a few heterogeneous benign clients as outliers. We argue that sacrificing a negligible amount of benign-case accuracy is an acceptable trade-off to achieve superior resilience against massive, sophisticated attacks where baselines fail.

**Table 3.** Final accuracy (%) on Cifar-10 under Byzantine attacks.

| Attack | $\rho$ | Ours | FedAvg | Krum | TrimMean | FLTrust |
|---|---|---|---|---|---|---|
| Gaussian | 20% | 62.3 | 38.4 | 36.5 | 57.0 | 63.6 |
| | 40% | 61.7 | 9.7 | 33.7 | 49.8 | 58.8 |
| | 60% | 54.9 | 10.4 | 47.7 | 40.6 | 54.8 |
| Reverse | 20% | 61.9 | 9.9 | 36.4 | 9.4 | 9.9 |
| | 40% | 54.5 | 9.9 | 35.4 | 10.4 | 9.9 |
| | 60% | 53.9 | 9.7 | 10.4 | 10.4 | 9.9 |
| Label Flip | 20% | 60.7 | 58.9 | 36.7 | 61.8 | 60.8 |
| | 40% | 62.9 | 52.1 | 12.6 | 59.2 | 57.2 |
| | 60% | 58.9 | 18.1 | 12.4 | 34.3 | 53.1 |

**Table 4.** Final accuracy (%) on FEMNIST under Byzantine attacks.

| Attack | $\rho$ | Ours | FedAvg | Krum | TrimMean | FLTrust |
|---|---|---|---|---|---|---|
| Gaussian | 20% | 83.1 | 27.6 | 70.8 | 78.6 | 82.8 |
| | 40% | 83.2 | 5.1 | 69.2 | 76.6 | 81.4 |
| | 60% | 81.4 | 5.1 | 5.1 | 45.7 | 73.9 |
| Reverse | 20% | 83.2 | 5.1 | 69.8 | 79.8 | 5.1 |
| | 40% | 81.5 | 5.1 | 68.3 | 51.8 | 5.1 |
| | 60% | 82.3 | 5.1 | 58.1 | 51.7 | 5.1 |
| Label Flip | 20% | 83.5 | 80.4 | 67.4 | 83.3 | 81.4 |
| | 40% | 82.7 | 72.8 | 5.1 | 79.0 | 74.4 |
| | 60% | 82.1 | 12.1 | 5.0 | 0.5 | 59.6 |

### *5.4. Complexity and Overhead Analysis*

To quantitatively assess the computational overhead and scalability of the proposed DAAD framework, we analyzed the theoretical complexity and measured the actual execution time and GPU memory usage compared to baseline methods. The experiments were conducted on a server equipped with an NVIDIA RTX 3090 GPU and an Intel Xeon CPU. The batch size for local training was set to 32, and the model updates were aggregated from $N = 20$ clients.

As shown in Table 5, DAAD records an average processing time of 70.1 ms per round, while this introduces a higher overhead compared to lightweight baselines like Krum (9.3 ms). Due to the GPU-based CNN inference, we argue that this overhead remains strictly acceptable in practical Cross-Silo FL scenarios. Given that network communication and local training typically consume seconds to minutes, an additional delay of $\sim$0.07 s is negligible (<1% of total round duration).

**Table 5.** Comparison of computational complexity and runtime overhead ($N = 20$, MNIST). Time is measured in milliseconds (ms) per communication round.

| Method | Time Complexity | Avg. Runtime (ms) | GPU Mem. (MB) | Scalability |
|---|---|---|---|---|
| FedAvg | $O(Nd)$ | 2.1 | 165.3 | High |
| Trimmed Mean | $O(Nd \log N)$ | 3.9 | 360.5 | High |
| Krum | $O(N^2 d)$ | 9.3 | 205.4 | Medium |
| FLTrust | $O(Nd)$ | 25.8 | 215.4 | High |
| DAAD (Ours) | $O(N^2 d + C)$ | 70.1 | 310.2 | Medium |

Regarding resource consumption, the GPU memory footprint for DAAD is 310.2 MB. Notably, this is comparable to, and in some cases more efficient than, statistical baselines like Trimmed Mean (360.5 MB), which requires memory-intensive sorting operations.

This confirms that DAAD provides advanced detection without imposing prohibitive resource demands.

### *5.5. Ablation Studies*

#### 5.5.1. Contribution of SE Blocks and Dual Channels

We also performed ablation experiments to demonstrate the contribution of each module in DAAD. These experiments were conducted using the MNIST dataset under the default setting, incorporating the three attack types described for training. We systematically varied the architecture: (1) single-channel (gradient similarity only) without SEBlock; (2) dual-channel without SEBlock; (3) single-channel with SEBlock; (4) dual-channel with SEBlock. As demonstrated by the ablation study control groups in Table 6 below (with metric definitions provided beneath the table), the comparison of single-channel configurations with and without SE Block reveals that adding SE yields only marginal accuracy improvements in the single-channel setting but significantly accelerates convergence speed. Further comparison of dual-channel configurations shows that the dual-channel architecture delivers substantial gains in convergence accuracy. Finally, the integration of both dual channels and SE Block further enhances validation accuracy while also optimizing convergence speed.

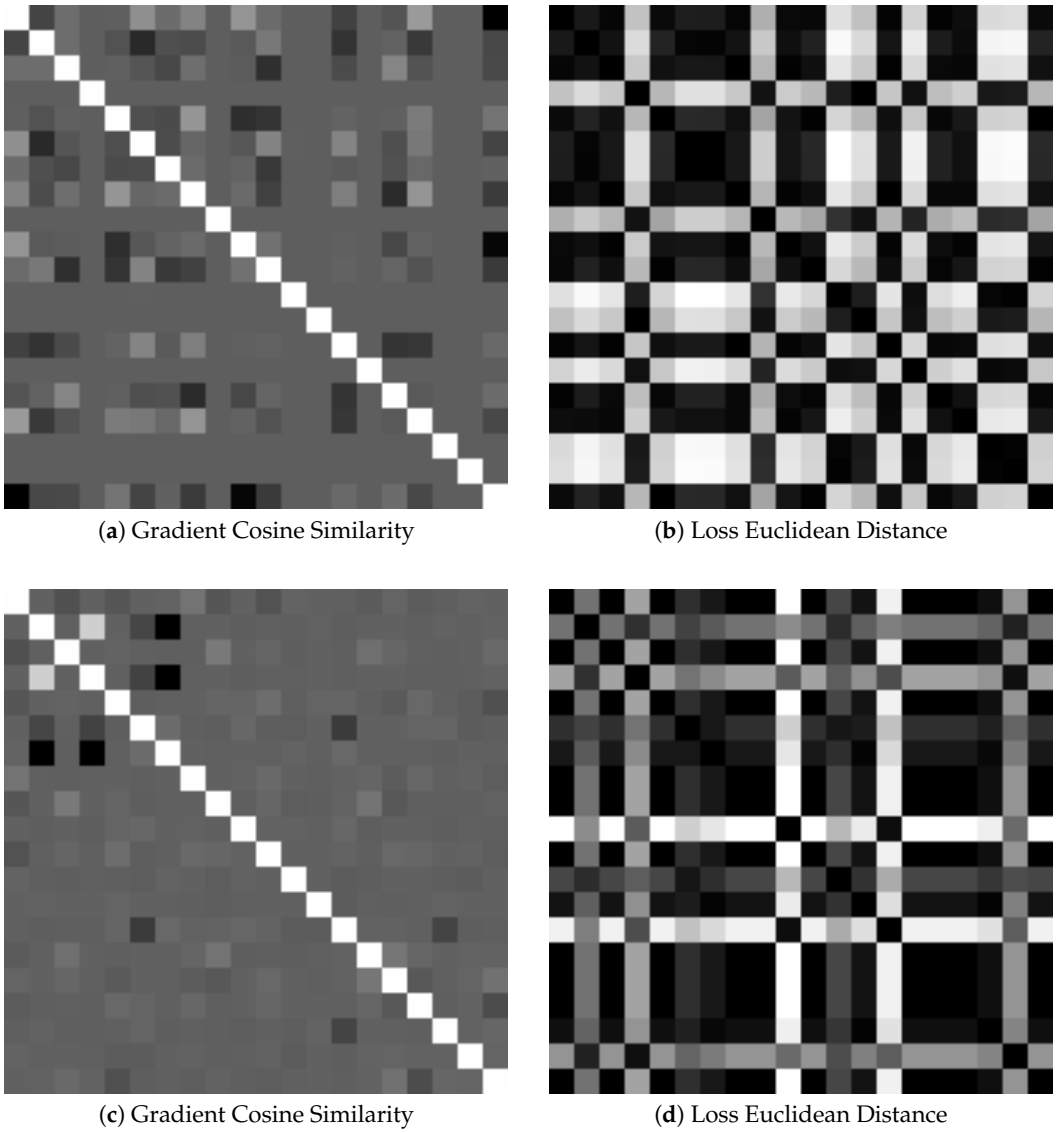
**Table 6.** Ablation study results (core metrics comparison).

| Experiment Group | Epoch | Peak Acc (%) | Final Acc (%) |
|---|---|---|---|
| Single-Ch (Grad, w/o SE) | 50 | 51.67 | $50.40 \pm 0.85$ |
| Single-Ch (Loss, w/o SE) | 20 | 89.97 | $89.58 \pm 0.42$ |
| Single-Ch (Grad, w/ SE) | 22 | 52.98 | $52.75 \pm 0.31$ |
| Single-Ch (Loss, w/ SE) | 90 | 90.66 | $89.97 \pm 0.68$ |
| Dual-Ch (w/o SE) | 27 | 96.54 | $96.27 \pm 0.21$ |
| Dual-Ch (w/ SE) | 37 | 97.59 | $97.39 \pm 0.18$ |

Note: Epoch: First, epoch reaching 95% of peak accuracy. Peak Accuracy: Highest validation accuracy observed during training. Final Accuracy: Mean $\pm$ standard deviation of accuracy over last 5 training epochs.

#### 5.5.2. Feature Visualization

To intuitively demonstrate the manifestation of client behavioral representations under different attack scenarios, we visualized the dual-channel feature maps in Figure 7. The top row illustrates the Gaussian noise attack, where anomalies are clearly discernible in both the gradient similarity map (left) and the loss divergence map (right). In contrast, the bottom row depicts the label flipping attack with a notably different pattern: the gradient map (left) fails to distinguish anomalous nodes from benign ones, appearing almost identical to the benign distribution. However, the Loss map (right) clearly highlights the anomalies as distinct clusters. This visual evidence strongly justifies the incorporation of the attention mechanism: it enables the model to adaptively capture the most informative features from the dual-channel inputs, shifting focus to the more distinct channel (e.g., the Loss channel in this specific case) when necessary, thereby ensuring robust detection across diverse attack patterns.



(**a**) Gradient Cosine Similarity (**b**) Loss Euclidean Distance

(**c**) Gradient Cosine Similarity (**d**) Loss Euclidean Distance

**Figure 7.** Comparison of client gradient and loss behaviors under different attack scenarios. The top row displays the Gaussian noise attack, while the bottom row shows the label flipping attack.

## 6. Discussion

### *6.1. Privacy Compliance in Cross-Institutional Scenarios*

We acknowledge that effective Byzantine detection typically requires analyzing plaintext model updates, which presents an inherent tension with gradient encryption schemes designed to prevent information leakage [24]. However, to comply with strict privacy regulations in healthcare or financial sectors, DAAD can be deployed within a Trusted Execution Environment (TEE), such as Intel SGX. Recent works have demonstrated that TEE-based infrastructures can securely bridge the gap between privacy and verifiability in cross-silo FL [25]. In this architecture, encrypted updates are transmitted to the server and decrypted only within the hardware-secured enclave. DAAD then performs anomaly detection on the plaintext data inside the enclave, outputting only the aggregation results without exposing raw gradients to the server operator. Furthermore, our framework is compatible with Differential Privacy (DP). As shown in recent studies [26], robust aggregation can effectively coexist with DP noise. Since DAAD targets macro-behavioral patterns (e.g., directional opposition), it remains robust against calibrated privacy-preserving noise, thereby balancing system robustness with data confidentiality.

### *6.2. Discussion on Generalization, Limitations, and Theoretical Aspects*

Despite the superior detection performance demonstrated on standard benchmarks, it is crucial to objectively recognize the inherent constraints of the current supervised learning paradigm and distinguish between empirical robustness and theoretical guarantees.

Challenge 1: Adaptive and Stealthy Attacks. One significant challenge involves sophisticated adaptive attacks. Although our synthetic training data encompasses fundamental adversarial primitives (e.g., directional and magnitude deviations), the detector's stability may fluctuate when confronting attackers with full knowledge of the defense mechanism. These intelligent adversaries can function analogously to generating "adversarial examples" in the high-dimensional feature space—dynamically optimizing their malicious updates to minimize the dual-channel distance metric while maximizing damage, thereby skirting the decision boundary of a statically trained model.

Challenge 2: Cross-Dataset Generalization. A second limitation pertains to the dependency on domain-specific training. Currently, the DAAD framework necessitates training a separate detector for each specific dataset to capture its unique gradient distribution, while this one-time offline cost is justifiable in stable Cross-Silo FL scenarios (such as medical or financial consortiums where data formats are consistent), it restricts rapid deployment across novel, unseen domains without retraining. This limitation highlights the need for a more transferable feature representation.

Theoretical Perspective. From a theoretical standpoint, we acknowledge that DAAD, being a deep learning-based method, does not offer the explicit, closed-form error bounds found in robust statistical defenses like Krum or Geometric Median. However, its feasibility is grounded in the Universal Approximation Theorem, which suggests that neural networks can effectively approximate the optimal decision boundary for non-linearly separable data (as evidenced by the distinct clusters in our t-SNE analysis). Regarding system-level convergence, as long as DAAD maintains a functional detection precision (empirically >97% in our experiments), the global aggregation simplifies to averaging over the filtered benign set. Consequently, the global model inherits the standard convergence guarantees of FedAvg for non-convex objectives under partial participation.

Future Directions. To address these limitations, our ongoing research pursues a unified, robust learning framework. Specifically, we aim to integrate Adversarial Training during the offline phase to robustify the decision boundary against adaptive perturbations. Simultaneously, we are exploring Domain Adaptation techniques to align feature distributions across source and target domains, enabling the detector to generalize to new tasks with minimal fine-tuning.

## 7. Conclusions

This paper presents a novel framework for Byzantine-resilient FL, built upon a Dual-Channel Attention-Driven Anomaly Detector (DAAD). By encoding inter-client behavioral signals, gradient cosine similarities, and loss divergences into structured spatial representations and leveraging SE-enhanced convolutional architectures, our method enables robust identification and filtering of malicious client updates. Unlike previous defense methods that rely heavily on strong assumptions (such as a known attack model or additional trusted data), DAAD leverages offline supervision on a diverse set of synthetic attack proxies to achieve robust detection. It operates in a unified detection mode during deployment, capable of consistently filtering malicious updates across different attack scenarios covered in the training distribution, without requiring access to private client data or clean reference models. Extensive experiments across diverse datasets (MNIST, CIFAR-10, FEMNIST), attack types, and Byzantine ratios validate the robustness and adaptability of our approach. DAAD consistently maintains stable convergence and model utility even

under high adversarial pressure (up to 60% compromised clients), outperforming existing baselines in challenging non-IID scenarios.

While the results are promising, several directions remain open for future work. First, our current implementation trains a separate detector per dataset; improving cross-domain generalization or enabling detector transferability would enhance deployment scalability. Second, although DAAD introduces minimal FLOPs overhead, exploring lightweight variants or real-time approximations could further optimize runtime efficiency. Lastly, integrating DAAD with robust aggregation schemes (e.g., trimmed mean and Bulyan) may offer complementary benefits and further strengthen defense in hybrid threat models.

**Author Contributions:** Conceptualization, J.H. and Y.Z.; methodology, Y.Z.; validation, X.L. and L.R.; formal analysis, Y.Z.; investigation, J.H. and X.Z.; resources, L.W. and X.Z.; writing—original draft preparation, Y.Z.; writing—review and editing, L.W.; visualization, X.L.; supervision, L.W.; project administration, Y.Z.; funding acquisition, L.W. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the National Natural Science Foundation of China (NSFC) under Grant number 62403233.

**Data Availability Statement:** The data used to support the findings of this study are included within the article.

**Acknowledgments:** We thank the reviewers for their constructive comments.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

| | |
|---|---|
| FL | Federated Learning |
| SE | Squeeze-and-Excitation |
| DAAD | Dual-Channel Attention-Driven Anomaly Detector |

## References

1. Wang, X.; Garg, S.; Lin, H.; Hu, J.; Kaddoum, G.; Piran, M.J.; Hossain, M.S. Toward Accurate Anomaly Detection in Industrial Internet of Things Using Hierarchical Federated Learning. *IEEE Internet Things J.* **2022**, *9*, 7110–7119. [CrossRef]
2. Wang, X.; Ye, B.; Xu, L.; Wu, L.; Hsieh, S.Y.; Wu, J.; Lin, L. FedHAN: A Cache-Based Semi-Asynchronous Federated Learning Framework Defending Against Poisoning Attacks in Heterogeneous Clients. In *Proceedings of the 34th International Joint Conference on Artificial Intelligence (IJCAI-25)*; IJCAI: San Mateo, CA, USA, 2025; pp. 3407–3416. [CrossRef]
3. Jin, K.; Xu, L.; Wang, X.; Hsieh, S.Y.; Wu, J.; Lin, L. FedCPD: Personalized Federated Learning with Prototype-Enhanced Representation and Memory Distillation. In *Proceedings of the 34th International Joint Conference on Artificial Intelligence (IJCAI-25)*; IJCAI: San Mateo, CA, USA, 2025. [CrossRef]
4. Wu, L.; Wang, X.; Lin, H.; Yang, X.; Shu, J.; Yi, X.; Khalil, I.; Zomaya, A.Y. Toward Personalized Federated Meta-Learning with Constrained Hypernetwork on Non-IID Data. *IEEE Trans. Comput.* **2025**, *75*, 233–246. [CrossRef]
5. Blanchard, P.; El Mhamdi, E.M.; Guerraoui, R.; Stainer, J. Machine learning with adversaries: Byzantine tolerant gradient descent. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 119–129. Available online: https://proceedings.neurips.cc/paper/2017/hash/f4b9ec30ad9f68f89b29639786cb62ef-Abstract.html (accessed on 14 February 2026).
6. Yin, D.; Chen, Y.; Kannan, R.; Bartlett, P. Byzantine-Robust Distributed Learning: Towards Optimal Statistical Rates. In *Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018*; Proceedings of Machine Learning Research: Cambridge, MA, USA, 2018; Volume 80, pp. 5650–5659.
7. El Mhamdi, E.M.; Guerraoui, R.; Rouault, S. The Hidden Vulnerability of Distributed Learning in Byzantium. In *Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018*; Proceedings of Machine Learning Research: Cambridge, MA, USA, 2018; Volume 80, pp. 3521–3530.
8. Nguyen, T.D.; Rieger, P.; Chen, H.; Yalame, H.; Möllering, H.; Fereidooni, H.; Marchal, S.; Miettinen, M.; Mirhoseini, A.; Zeitouni, S.; et al. FLAME: Taming Backdoors in Federated Learning. In Proceedings of the 31st USENIX Security Symposium (USENIX Security '22), Boston, MA, USA, 10–12 August 2022; pp. 1415–1432.

9. Pillutla, K.; Kakade, S.M.; Harchaoui, Z. Robust Aggregation for Federated Learning. *IEEE Trans. Signal Process.* **2022**, *70*, 1142–1154. [CrossRef]
10. Saday, A.; Demirel, İ.; Yıldırım, Y.; Tekin, C. Federated Multi-Armed Bandits Under Byzantine Attacks. *arXiv* **2022**, arXiv:2205.04134. [CrossRef]
11. Zhou, A.; Liu, Y.; Chai, Y.; Zhu, H.; Ge, X.; Jiang, Y.; Wang, M. A Whole-Process Certifiably Robust Aggregation Method Against Backdoor Attacks in Federated Learning. *arXiv* **2024**, arXiv:2407.00719. [CrossRef]
12. Zhao, P.; Jiang, J.; Zhang, G. FedSuper: A Byzantine-Robust Federated Learning Under Supervision. *ACM Trans. Sens. Netw.* **2024**, *20*, 1–29. [CrossRef]
13. Pan, X.; Zhang, M.; Wu, D.; Xiao, Q.; Ji, S.; Yang, Z. Justinian's GAAvernor: Robust Distributed Learning with Gradient Aggregation Agent. In Proceedings of the 29th USENIX Security Symposium (USENIX Security '20), Boston, MA, USA, 12–14 August 2020; pp. 1641–1658.
14. Cao, X.; Fang, M.; Liu, J.; Gong, N.Z. FLTrust: Byzantine-Robust Federated Learning via Trust Bootstrapping. *arXiv* **2020**, arXiv:2012.13995. [CrossRef]
15. Li, Y.; Xia, C.; Li, C.; Wang, T. BRFL: A Blockchain-based Byzantine-Robust Federated Learning Model. *arXiv* **2023**, arXiv:2310.13403. [CrossRef]
16. Guan, M.; Bao, H.; Li, Z.; Pan, H.; Huang, C.; Dai, H.N. SAMFL: Secure Aggregation Mechanism for Federated Learning with Byzantine-robustness by functional encryption. *J. Syst. Archit.* **2024**, *157*, 103304. [CrossRef]
17. Lin, X.; Li, Y.; Xie, X.; Ding, Y.; Wu, X.; Ge, C. SF-CABD: Secure Byzantine Fault Tolerance Federated Learning on Non-IID Data. *Knowl.-Based Syst.* **2024**, *296*, 111851. [CrossRef]
18. Xie, C.; Koyejo, O.; Gupta, I. Generalized byzantine-tolerant sgd. *arXiv* **2018**, arXiv:1802.10116. [CrossRef]
19. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
20. Kim, Y. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2014; pp. 1746–1751.
21. Goodfellow, I.; Bengio, Y.; Courville, A. Chapter 9: Convolutional Networks. In *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016.
22. McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; y Arcas, B.A. Communication-efficient learning of deep networks from decentralized data. In Proceedings of the Artificial Intelligence and Statistics, PMLR, Fort Lauderdale, FL, USA, 20–22 April 2017; pp. 1273–1282.
23. Shejwalkar, V.; Houmansadr, A. Manipulating the Byzantine: Optimizing Model Poisoning Attacks and Defenses for Federated Learning. In Proceedings of the NDSS, Virtual, 21–25 February 2021.
24. Wang, D.; Chen, C.; Li, S.; Xu, J. Practical Differentially Private and Byzantine-resilient Federated Learning. In *Proceedings of the 2023 ACM International Conference on Management of Data (SIGMOD)*; ACM: New York, NY, USA, 2023.
25. Zhang, K.; Gu, T.; Zhang, X. TEE based Cross-silo Trustworthy Federated Learning Infrastructure. In *Proceedings of the IJCAI Workshop on Trustworthy Federated Learning*; Springer: Cham, Switzerland, 2022.
26. Ma, X.; Sun, Y.; Wu, J. Bridging Differential Privacy and Byzantine-Robustness via Model Aggregation. In Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI), Vienna, Austria, 23–29 July 2022; pp. 708–714.


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