

# Beyond Single-Modality Evaluation: Measuring Robustness Variability of Federated Learning Defenses in Wireless Edge Systems

Sofiane AISSANI, Leila MERGHEM BOULAHIA, Yanis BOUSSOUM, Engku Ahmad BIN ENSKU AZIZ

University of Technology of Troyes (UTT), Troyes, France

Computer Science and Digital Society Laboratory (LIST3N), Troyes, France

{sofiane.aissani, leila.merghem, engku\_ahmad.bin\_engku\_aziz}@utt.fr

**Abstract**—Federated Learning (FL) allows distributed edge devices to collaboratively train shared models without sharing their private data. However, this decentralized setting remains vulnerable to backdoor attacks, where malicious clients can manipulate the global model without being detected by the server. Although many FL defenses perform well on image data, their assumptions may not generalize to other data structures. This paper introduces a modality-aware evaluation framework to study the robustness of FL defenses across different data geometries, including spatial, sequential, and relational modalities. As part of this framework, we define the Modality Gap (MG) metric to quantify defense instability under modality shifts. Experiments on CIFAR-10, IMDB, and Cora under heterogeneous edge availability conditions reveal an important trade-off: statistical defenses provide general but limited protection, while specialized methods such as Activation Clustering achieve strong performance on images but significantly degrade on text and graph data, with a high Modality Gap of 0.412. These results show that evaluating FL security on a single modality can lead to overly optimistic robustness conclusions for heterogeneous edge environments.

**Index Terms**—Federated learning, mobile edge computing, backdoor attacks, defense robustness, data modalities, wireless security.

## I. INTRODUCTION

Wireless edge systems increasingly integrate machine learning to support intelligent applications on smartphones, IoT devices, connected vehicles, and distributed sensors. In such environments, Federated Learning (FL) enables multiple devices to collaboratively train a shared model without exchanging raw data [2], [10]. This property makes FL suitable for large-scale edge deployments where data privacy and communication efficiency are important concerns [16].

However, the decentralized nature of FL creates new security challenges. Since the server only observes model updates, malicious clients can manipulate the training process without accessing or inspecting their local data [12]. Among these threats, backdoor attacks are particularly challenging because they introduce hidden malicious behaviors while maintaining normal model performance [8]. Existing studies have shown that such attacks can remain effective even when only a limited number of clients are compromised [9], [14].

To mitigate these threats, several FL defenses have been developed, including robust aggregation methods [1], [17], similarity-based detection techniques [4], and model inspection

approaches [5], [6]. However, most existing evaluations are conducted on image datasets, where spatial structures dominate the data representation. Existing evaluations implicitly assume that defense effectiveness transfers across data modalities. However, this assumption remains unvalidated in heterogeneous wireless edge environments where devices generate fundamentally different data structures.

This issue is critical because edge devices operate on diverse data types, including images, text, and graphs. These modalities differ significantly in their statistical properties, feature representations, and learning behaviors. Therefore, a defense mechanism that performs well on one modality may not provide the same level of protection when applied to another modality.

In this paper, we investigate the following research question:

*Do federated learning defenses remain robust when evaluated across heterogeneous data modalities?*

To answer this question, we introduce a modality-aware evaluation framework that analyzes the stability of FL defenses across spatial, sequential, and relational data structures. Within this framework, we define the *Modality Gap (MG)* metric to quantify defense variability caused by modality changes.

The main contributions of this paper are:

- We identify cross-modality robustness instability as a missing dimension in federated learning security evaluation.
- We introduce a modality-aware evaluation framework and the Modality Gap (MG) metric to quantify defense variability across heterogeneous data structures.
- We conduct a systematic evaluation of representative FL defenses across spatial, sequential, and relational modalities using image, text, and graph datasets under heterogeneous edge availability conditions.

Our results show that robustness does not consistently transfer across modalities. Some defenses provide stable but limited protection, while others achieve strong performance on image data but degrade significantly on text and graph data. These findings demonstrate that single-modality evaluations can lead to overly optimistic security assessments for heterogeneous wireless edge systems.

The remainder of this paper is organized as follows. Section II presents the proposed Modality Gap framework and metric.

Section III describes the experimental methodology. Section IV discusses the evaluation results. Section V presents the implications, limitations, and future directions. Finally, Section VI concludes the paper.

## II. MODALITY GAP FRAMEWORK

Federated learning defenses are usually evaluated using a single metric, such as Attack Success Rate (ASR), under a fixed data distribution. However, this evaluation assumes that defense behavior remains stable when the underlying data structure changes. In heterogeneous wireless edge systems, this assumption may not hold because different devices can generate data with different structures, representations, and properties.

Images, text, and graphs represent information in fundamentally different ways. As a result, a defense that effectively detects backdoors in one modality may become less effective in another modality, even under similar attack and training settings. We refer to this robustness instability as the *Modality Gap*.

The Modality Gap does not measure the absolute performance of a defense. Instead, it quantifies how much the defense behavior changes across different modalities. A low MG indicates consistent robustness, while a high MG reveals strong dependence on modality-specific characteristics.

### A. Definition

Let  $ASR(D, m)$  denote the Attack Success Rate obtained by defense  $D$  on modality  $m \in \mathcal{M}$ . The average ASR of a defense is defined as:

$$\mu_D = \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} ASR(D, m) \quad (1)$$

The Modality Gap is defined as the normalized variation of ASR across modalities:

$$MG(D) = \frac{\sqrt{\frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} (ASR(D, m) - \mu_D)^2}}{\mu_D + \epsilon} \quad (2)$$

where  $\epsilon$  is a small constant used to avoid division by zero.

The numerator represents the variation of defense performance across modalities, while the normalization term enables comparison between defenses with different average ASR values.

Unlike conventional robustness metrics that evaluate defense effectiveness under a fixed data distribution, MG evaluates the stability of this effectiveness when the underlying data structure changes. Therefore, two defenses achieving similar average ASR may provide substantially different deployment reliability across heterogeneous edge environments.

### B. Basic Property

**Property 1 (Zero Gap Consistency).** *A defense has zero Modality Gap if and only if its ASR is identical across all modalities.*

TABLE I  
EDGE FEDERATED LEARNING CONFIGURATION.

| Parameter            | Value  |
|----------------------|--------|
| Clients              | 50     |
| Aggregation          | FedAvg |
| Communication rounds | 100    |
| Local epochs         | 5      |
| Batch size           | 64     |
| Learning rate        | 0.01   |
| Dirichlet $\alpha$   | 0.5    |
| Malicious clients    | 10%    |

$$MG(D) = 0 \Leftrightarrow ASR(D, m_1) = \dots = ASR(D, m_k) \quad (3)$$

This property directly follows from the definition of standard deviation. If all modality-specific ASR values are identical, their variance is zero, resulting in  $MG(D) = 0$ . Conversely, a zero MG means that no variation exists between modalities.

### C. Interpretation

The Modality Gap captures robustness differences that cannot be observed from average ASR alone. Two defenses may achieve similar average attack resistance while showing very different stability across modalities.

A low MG indicates that a defense maintains similar behavior across spatial, sequential, and relational data. In contrast, a high MG indicates that the defense relies strongly on modality-specific characteristics and may not generalize well to new deployment scenarios.

Therefore, MG complements existing security metrics by measuring whether robustness transfers across heterogeneous data structures. This aspect is particularly important for wireless edge systems, where multiple data modalities may coexist within the same infrastructure.

## III. EXPERIMENTAL METHODOLOGY

This section presents the experimental setup used to evaluate federated learning defenses across heterogeneous data modalities under different edge client availability scenarios.

### A. Edge Federated Learning Setup

We consider an edge federated learning system composed of 50 clients and one central server. Model training follows the FedAvg protocol [10]. At each communication round, only a subset of clients participates in aggregation to represent heterogeneous client availability in edge deployments.

To simulate data heterogeneity, client data are distributed using a Dirichlet partition with  $\alpha = 0.5$  [3]. All experiments are conducted over 100 communication rounds. The main configuration is summarized in Table I.

### B. Datasets, Models, and Attack Configuration

We evaluate three data modalities: spatial, sequential, and relational data. CIFAR-10 is used for image classification, IMDB for text classification, and Cora for graph classification.

For each modality, we adopt a task-specific model architecture. A lightweight CNN is used for CIFAR-10, an LSTM

TABLE II  
MODELS AND ATTACK CONFIGURATIONS.

| Dataset  | Model       | Parameters     | Trigger            |
|----------|-------------|----------------|--------------------|
| CIFAR-10 | CNN         | 2 Conv. layers | $4 \times 4$ patch |
| IMDB     | LSTM        | 128/128        | Rare token         |
| Cora     | 2-layer GCN | Hidden=64      | Structural         |

TABLE III  
CLIENT AVAILABILITY SETTINGS.

| Participation rate | Scenario                        |
|--------------------|---------------------------------|
| 20%                | High availability variation     |
| 40%                | Moderate availability variation |
| 60%                | Low availability variation      |

model is used for IMDB, and a two-layer GCN is used for Cora. These architectures represent common choices for heterogeneous edge learning tasks.

The backdoor attacks are adapted to each data structure. For CIFAR-10, we use a BadNets-style attack with a  $4 \times 4$  pixel patch inserted into poisoned images [7]. For IMDB, the attack introduces a rare trigger token into selected sentences and associates it with the target label. For Cora, the attack modifies graph connectivity by adding a structural trigger pattern.

Although the trigger implementations differ by modality, all attacks share the same objective: inducing targeted misclassification while preserving benign performance. This allows evaluating defense behavior under modality-specific representations.

In all experiments, 10% of clients are considered malicious, and poisoned clients modify a fraction of their local training samples.

Table II summarizes the model and trigger configurations.

### C. Client Availability Model

Client availability is modeled by varying the fraction of participating clients during training. In each communication round, only a subset of clients contributes to aggregation, representing intermittent participation and dynamic edge conditions.

Three participation rates are considered: 20%, 40%, and 60%. Lower participation represents highly dynamic edge availability, while higher participation corresponds to more stable client participation.

All other experimental parameters remain unchanged. This analysis evaluates whether defense robustness and the proposed Modality Gap remain consistent under different client availability scenarios.

## IV. RESULTS

This section evaluates the robustness of three FL defenses against backdoor attacks across image, text, and graph modalities in a wireless edge environment. Three research questions are investigated: (i) How effectively do existing defenses mitigate backdoor attacks? (ii) Does defense robustness remain consistent across different modalities? (iii) What factors explain the observed variations?

### A. Performance Across Modalities

Table IV reports the Clean Accuracy (CA) and Attack Success Rate (ASR) obtained by each evaluated defense.

The main observation is that defense effectiveness is strongly modality-dependent. The best-performing defense changes when the data modality changes. A defense that achieves the strongest protection on images does not necessarily provide the same level of robustness on text or graph data.

To highlight this behavior, Table V presents the defense ranking according to ASR, where rank 1 corresponds to the lowest attack success rate.

This ranking inversion highlights that defense effectiveness is modality-dependent. A defense selected using image-only benchmarks may not be optimal for heterogeneous edge deployments involving multiple data types.

FedAvg provides almost no protection against backdoor attacks. The ASR remains above 90% across all evaluated modalities. The attack remains effective despite the distributed training process, showing that privacy preservation alone does not guarantee robustness against malicious updates.

Trimmed Mean provides a moderate improvement by reducing ASR by approximately half in all modalities. However, the remaining attack success rate is still significant. Around one poisoned sample out of two can still reach the attacker-selected target label, which limits the protection provided by this aggregation strategy.

FLAME achieves a stronger reduction in attack effectiveness. The ASR decreases to 22.62% on CIFAR-10 and remains below 35% on all evaluated datasets. Its behavior is relatively stable across modalities. Although it does not achieve the best result on images, it provides consistent protection on text and graph data.

Activation Clustering [18] achieves the strongest protection on image data. On CIFAR-10, ASR decreases from 95.57% to 12.18%, corresponding to an absolute reduction of more than 83 percentage points. This result makes Activation Clustering the most effective defense for the image modality.

However, this advantage does not transfer to other modalities. The same defense reaches 38.77% ASR on IMDB and 36.11% on Cora. Therefore, the strong performance observed on images becomes significantly weaker on text and graph data. In contrast, FLAME becomes the best-performing defense on these modalities.

Clean Accuracy remains stable across all experiments. The variations between defenses and modalities are limited, indicating that the evaluated mechanisms mainly affect security robustness rather than the normal predictive performance of the global model.

Overall, these results show that robustness is not an intrinsic property of a defense mechanism alone. It depends on the interaction between the defense strategy and the underlying data structure. The ranking changes across modalities, demonstrating that selecting a defense for wireless federated learning should consider the characteristics of the targeted data domain.

TABLE IV  
PERFORMANCE ACROSS MODALITIES UNDER WIRELESS FEDERATED LEARNING.

| Defense               | CIFAR-10        |                 | IMDB            |                 | Cora            |                 |
|-----------------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
|                       | CA (%)          | ASR (%)         | CA (%)          | ASR (%)         | CA (%)          | ASR (%)         |
| FedAvg                | 82.83 $\pm$ 2.1 | 95.57 $\pm$ 4.1 | 86.31 $\pm$ 3.2 | 96.07 $\pm$ 4.5 | 83.28 $\pm$ 3.3 | 90.12 $\pm$ 5.1 |
| Trimmed Mean          | 83.23 $\pm$ 2.2 | 48.37 $\pm$ 3.9 | 86.98 $\pm$ 2.9 | 45.56 $\pm$ 4.2 | 85.74 $\pm$ 2.8 | 42.03 $\pm$ 4.1 |
| FLAME                 | 85.63 $\pm$ 3.5 | 22.62 $\pm$ 5.1 | 87.59 $\pm$ 2.8 | 25.87 $\pm$ 5.7 | 83.02 $\pm$ 2.9 | 34.09 $\pm$ 7.8 |
| Activation Clustering | 87.28 $\pm$ 4.0 | 12.18 $\pm$ 3.4 | 84.91 $\pm$ 3.9 | 38.77 $\pm$ 5.6 | 84.61 $\pm$ 4.6 | 36.11 $\pm$ 8.1 |

TABLE V  
DEFENSE RANKING ACCORDING TO ASR ACROSS MODALITIES.

| Defense               | CIFAR-10 | IMDB | Cora |
|-----------------------|----------|------|------|
| Activation Clustering | 1        | 3    | 3    |
| FLAME                 | 2        | 1    | 1    |
| Trimmed Mean          | 3        | 2    | 2    |

TABLE VI  
MODALITY GAP (MG) OF EVALUATED DEFENSES.

| Defense               | MG    |
|-----------------------|-------|
| Trimmed Mean          | 0.057 |
| FLAME                 | 0.174 |
| Activation Clustering | 0.412 |

TABLE VII  
AVAILABILITY-AWARE CLIENT PARTICIPATION SETTINGS.

| Participation Rate | Wireless Scenario                 |
|--------------------|-----------------------------------|
| 20%                | High availability variability     |
| 40%                | Moderate availability variability |
| 60%                | Stable participation              |

TABLE VIII  
MODALITY GAP UNDER DIFFERENT WIRELESS AVAILABILITY CONDITIONS.

| Defense               | 20%  | 40%  | 60%  |
|-----------------------|------|------|------|
| Trimmed Mean          | 0.17 | 0.19 | 0.12 |
| FLAME                 | 0.31 | 0.29 | 0.31 |
| Activation Clustering | 0.59 | 0.58 | 0.65 |

### B. Modality Gap Analysis

Attack Success Rate (ASR) measures how effectively a defense reduces the impact of a backdoor attack. However, ASR alone does not indicate whether this protection remains consistent when the data modality changes. For this reason, we introduce the Modality Gap (MG) to evaluate cross-modality robustness.

The MG values reveal a behavior that cannot be observed from ASR values alone. Trimmed Mean achieves the smallest Modality Gap. Although its protection level is limited, its behavior remains stable across image, text, and graph modalities.

FLAME provides stronger protection than Trimmed Mean but shows a larger MG. This indicates that its effectiveness is more influenced by the underlying modality structure.

Activation Clustering presents the largest Modality Gap. The defense achieves excellent performance on CIFAR-10 but loses effectiveness on IMDB and Cora. This result shows that the assumptions exploited by Activation Clustering are strongly related to the characteristics of visual representations.

These results reveal a trade-off between specialization and transferability. Representation-aware defenses can achieve excellent performance when their assumptions match the modality structure, but this advantage may disappear when deployed on unseen data geometries.

The Modality Gap therefore provides additional information beyond ASR. A defense may achieve low ASR on a specific benchmark while providing limited robustness guarantees in heterogeneous wireless edge environments.

### C. Availability-Aware Modality Robustness Analysis

Wireless edge environments are characterized by intermittent connectivity, dynamic client availability, and participation variability. Unlike conventional federated learning settings, wireless devices may temporarily become unavailable due to changing network conditions, energy limitations, or communication constraints.

To evaluate whether the proposed Modality Gap (MG) remains stable under such conditions, we perform an availability-aware robustness analysis. Wireless availability variability is approximated through different client participation rates during training. At each communication round, only a subset of clients participates in aggregation, reflecting the stochastic availability commonly observed in wireless edge systems.

All other experimental factors remain unchanged. The model architecture, attack configuration, defense mechanisms, optimization procedure, and non-IID data partitioning are identical to the baseline setup. Therefore, observed variations are mainly attributed to changes in client participation.

We consider three availability regimes corresponding to different levels of client participation, as summarized in Table VII.

For each availability setting, we evaluate the Attack Success Rate (ASR), Clean Accuracy (CA), and Modality Gap (MG). This analysis determines whether defense effectiveness and cross-modality robustness remain consistent under different wireless participation conditions.

Overall, reduced client participation affects the absolute ASR values of all defenses by increasing the variability of the

aggregation process. However, the cross-modality robustness pattern remains stable across all availability settings.

Defenses with strong modality dependence under stable participation continue to exhibit similar behavior under variable wireless conditions. These results indicate that the Modality Gap is not only related to a specific participation scenario but represents an inherent characteristic of defense behavior across heterogeneous modalities.

#### D. Understanding the Origin of the Modality Gap

The Modality Gap results from the different ways backdoor patterns are represented across data modalities. Although the attack objective remains identical, the visibility of the malicious features depends on the underlying data structure.

1) *Image Modality*: For images, the BadNets trigger [7] introduces a localized pixel pattern that creates a concentrated feature representation. This separation between benign and poisoned samples enables Activation Clustering to identify malicious representations effectively, resulting in the lowest ASR on CIFAR-10.

2) *Text Modality*: For text data, the trigger is embedded in a high-dimensional representation space. The malicious features are distributed across multiple dimensions and may overlap with normal language patterns. This reduces the effectiveness of representation-based detection methods.

3) *Graph Modality*: For graph data, the trigger modifies structural relationships and propagates through message-passing operations. The malicious behavior is therefore distributed across connected nodes, making poisoned and benign representations harder to separate.

These observations show that defense robustness depends on the interaction between the attack mechanism and the underlying data geometry. This aspect is particularly important for wireless edge systems, where heterogeneous modalities may coexist across different devices.

#### E. Statistical Validation

To assess the robustness of our observations, all experiments are repeated using five independent random seeds. Each seed affects client sampling, data shuffling, and model initialization, capturing the stochasticity of federated optimization and non-IID data distributions.

Statistical significance is evaluated using paired t-tests on Attack Success Rate (ASR) across runs. The null hypothesis assumes no performance difference between two defenses under the same modality and training conditions. A low p-value indicates that the observed difference is unlikely to be explained by random variation alone.

Table IX reports pairwise statistical comparisons between the main defense families.

Overall, Activation Clustering shows statistically significant improvements over both baselines ( $p < 0.05$ ), providing evidence of consistent performance differences under the evaluated settings. In contrast, the difference between FLAME and Trimmed Mean is not statistically significant at the 0.05 level, suggesting comparable robustness in the considered scenarios.

TABLE IX  
STATISTICAL SIGNIFICANCE (PAIRED T-TEST) ON ASR DIFFERENCES  
ACROSS DEFENSES.

| Comparison                            | p-value | Interpretation               |
|---------------------------------------|---------|------------------------------|
| Activation Clustering vs Trimmed Mean | 0.012   | Significant                  |
| Activation Clustering vs FLAME        | 0.030   | Significant                  |
| FLAME vs Trimmed Mean                 | 0.051   | Not significant (borderline) |

Beyond pairwise comparisons, the results remain consistent across all five seeds. Activation Clustering achieves the lowest ASR on CIFAR-10, while FLAME systematically outperforms Trimmed Mean. This ranking behavior remains stable on IMDB and Cora despite the additional variability introduced by text representations and graph message-passing operations.

These results indicate that the observed trends are not mainly caused by initialization noise, client sampling randomness, or optimization variability. Instead, they reflect stable behavioral differences between defense mechanisms across modalities, supporting the existence of a reproducible Modality Gap rather than a seed-dependent artifact.

## V. DISCUSSION, LIMITATIONS, AND FUTURE WORK

### A. Implications for Wireless Edge Systems

Our results show that federated learning robustness depends not only on the attack strategy but also on the data modality processed by edge devices. This observation is important for wireless edge systems, where devices may simultaneously generate heterogeneous data, including images, sensor measurements, text, and graph-based information.

Different modalities create different challenges for backdoor detection. For example, image triggers often create localized patterns that can be easier to separate from benign representations. In contrast, sequential and graph data contain distributed structures, where malicious patterns may overlap with normal behaviors and become harder to isolate.

These findings have practical implications for edge FL deployment. Selecting a defense based only on image benchmarks may lead to unreliable robustness expectations in real heterogeneous environments. Therefore, future FL security evaluation should consider multiple data modalities before deploying a defense in wireless edge systems.

### B. Limitations and Future Directions

This study has several limitations. First, we evaluate three representative modalities using CIFAR-10, IMDB, and Cora. Although these datasets cover image, text, and graph scenarios, they do not represent the full diversity of edge applications.

Second, we analyze three representative defenses and fixed attack configurations. Future studies should consider a larger range of adaptive defenses and attackers optimized against specific protection mechanisms.

Third, experiments are performed in a simulated FL environment. Real wireless deployments may introduce additional constraints, including communication delays, unstable connectivity, and limited device resources.

Future research should investigate larger-scale evaluations, real edge implementations, and adaptive attack scenarios. A key direction is the design of modality-agnostic FL defenses that can provide stable protection across heterogeneous data structures.

### C. Toward Modality-Agnostic Federated Learning Defenses

The observed Modality Gap indicates that current defenses often rely on modality-specific characteristics. Future defenses should instead focus on identifying malicious behaviors that remain consistent across different data representations.

One promising direction is multimodal latent-space alignment. Image, text, and graph models produce different representations, but these representations may contain shared information related to benign and malicious updates. Mapping client representations into a common latent space using self-supervised contrastive objectives could reduce the dependency on modality-specific features. However, such methods must remain lightweight because additional representation exchange or projection modules may increase computation and communication costs for resource-constrained edge devices.

Another direction is self-supervised multimodal representation learning. Contrastive or masked learning objectives can exploit unlabeled client data to learn more general features and may improve the detection of abnormal updates across modalities.

Nevertheless, achieving modality-agnostic robustness remains challenging. A future defense must detect diverse backdoor behaviors while preserving legitimate differences between modalities. The effectiveness of such approaches could be evaluated by measuring whether they reduce the Modality Gap while maintaining low Attack Success Rates under heterogeneous wireless edge conditions.

## VI. CONCLUSION

This paper investigated the robustness of federated learning defenses against backdoor attacks across different data modalities in wireless edge systems.

We introduced the Modality Gap (MG) as a metric to quantify how defense effectiveness changes when the underlying data structure varies. Through evaluations on image, text, and graph modalities, we showed that defense performance is strongly modality-dependent. A defense that achieves strong protection on one modality may provide weaker robustness when applied to another modality.

Our results demonstrate that single-modality evaluation can lead to incomplete robustness conclusions for heterogeneous federated learning environments. Defense mechanisms should therefore be designed and evaluated while considering the characteristics of the targeted data modality.

The main implication is that robustness guarantees obtained from single-modality benchmarks should not be directly transferred to heterogeneous wireless edge deployments.

## REFERENCES

- [1] D. Yin, Y. Chen, K. Ramchandran, and P. Bartlett, "Byzantine-robust distributed learning," in *Proc. International Conference on Machine Learning (ICML)*, 2018, pp.5630–5639.
- [2] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Processing Magazine*, vol.37, no.4, pp.50–60, 2020.
- [3] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, "Federated learning with non-IID data," *arXiv preprint arXiv:1806.00582*, 2018.
- [4] T. D. Nguyen, P. Rieger, H. Chen, M. Sridhar, H. Li, M. Mao, C. Yao, T. Y. Li, R. Mani, H. Hu, and J. Zhao, "FLAME: Taming backdoors in federated learning," in *Proc. USENIX Security Symposium*, 2022, pp.1705–1722.
- [5] B. Wang, Y. Yao, S. Shan, H. Li, B. Viswanath, H. Qian, Y. Zhang, H. Yang, H. Lin, and H. Wang, "Neural Cleanse: Identifying and mitigating backdoor attacks in neural networks," in *Proc. IEEE Symposium on Security and Privacy (S&P)*, 2019, pp.339–354.
- [6] K. Liu, B. Dolan-Gavitt, and S. Garg, "Fine-pruning: Defending against backdooring attacks on deep neural networks," in *Proc. International Symposium on Research in Attacks, Intrusions, and Defense (RAID)*, 2018, pp.273–294.
- [7] T. Gu, B. Dolan-Gavitt, and S. Garg, "BadNets: Identifying vulnerabilities in the machine learning model supply chain," *arXiv preprint arXiv:1708.06733*, 2017.
- [8] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, "How to backdoor federated learning," in *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020, pp.2938–2948.
- [9] C. Xie, S. Koyejo, and I. Gupta, "DBA: Distributed backdoor attacks against federated learning," in *Proc. International Conference on Learning Representations (ICLR)*, 2020.
- [10] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017, pp.1273–1282.
- [11] M. Al-Quraan, L. Mohjazi, L. Bariah, A. Centeno, A. Zoha, S. Muhaidat, M. Debbah, and M. A. Imran, "Edge-Native Intelligence for 6G Communications Driven by Federated Learning: A Survey of Trends and Challenges," *IEEE Communications Surveys & Tutorials*, 2021.
- [12] S. Alhawas and M. A. Rassam, "Federated Learning in Edge Computing: Vulnerabilities, Attacks, and Defenses—A Survey," *Sensors*, vol.26, no.4, 2026.
- [13] S. Niknam, H. S. Dhillon, and J. H. Reed, "Federated Learning for Wireless Communications: Motivation, Opportunities, and Challenges," *IEEE Communications Magazine*, vol.58, no.6, pp.46–51, 2020.
- [14] Z. Zhang, S. Chen, M. Hong, and Y. Yin, "Neurotoxin: Durable backdoors in federated learning," in *Proc. International Conference on Machine Learning (ICML)*, 2022, pp.26429–26446.
- [15] T. T. Nguyen, B. Zhao, C. Xie, and I. Gupta, "Backdoor attacks and defenses in federated learning," *Engineering Applications of Artificial Intelligence*, vol.126, Art. 107061, 2023.
- [16] K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, D. Ramage, A. Segal, and K. Seth, "Practical secure aggregation for privacy-preserving machine learning," in *Proc. ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2017, pp.1175–1191.
- [17] P. Blanchard, E. M. El Mhamdi, R. Guerraoui, and J. Stainer, "Machine learning with adversaries: Byzantine tolerant gradient descent," in *Proc. NeurIPS*, 2017.
- [18] B. Chen, W. Carvalho, N. Baracaldo, H. Ludwig, B. Edwards, T. Lee, I. Molloy, and B. Srivastava, "Detecting backdoor attacks on deep neural networks by activation clustering," in *Workshop on Artificial Intelligence Safety*, 2018.