RBBD: A Representation-Based Framework for Edge-Case Backdoor Defense in Federated Learning

Samir Poudel
Computational and Data Science
Middle Tennessee State University
Murfreesboro, TN 37132
sp2ai@mtmail.mtsu.edu

Kritagya Upadhyay
Department of Computer Science
Middle Tennessee State University
Murfreesboro, TN 37132
kritagya.upadhyay@mtsu.edu

Jiblal Upadhya
Department of Applied Computing
Lander University
Greenwood, S.C. 29649
jupadhya@lander.edu

***Abstract*—Federated learning systems face a critical vulnerability from edge-case backdoor attacks that target uncommon but legitimate inputs occurring in the tail of data distributions, yet appearing regularly in real deployments. These attacks are particularly insidious because they maintain normal performance on common data while introducing malicious behaviors on infrequent patterns, making them undetectable through standard validation. Existing defenses that analyze parameter statistics fail because sophisticated attackers can craft updates that appear statistically normal while embedding semantic backdoors. We propose RBBD, a representation-based framework that detects backdoors by analyzing semantic anomalies in learned feature representations. Our approach combines three novel components: multi-layer feature space analysis, distribution-aware filtering that applies heightened scrutiny to low-density data regions, and adaptive thresholding that evolves with threat patterns. Extensive evaluation across multiple datasets and attack scenarios demonstrates that RBBD achieves 78.9% clean accuracy while reducing backdoor success rates to 3.1%. This represents a significant improvement over state-of-the-art defenses by achieving superior performance through semantic understanding while preserving clean edge-case accuracy.**

***Index Terms*—federated learning, backdoor attacks, edge cases, cybersecurity, model poisoning, robust aggregation**

I. Introduction and Motivation

Federated Learninig has gained significant traction as organizations seek to collaboratively train machine learning models without sharing sensitive data [1], [2]. This approach has proven particularly valuable in healthcare, finance, and mobile applications where privacy regulations prevent direct data sharing [3], [4]. The paradigm's appeal extends beyond privacy preservation, offering computational efficiency benefits by distributing model training across edge devices [5]. However, the distributed nature introduces unique security challenges, with backdoor attacks representing one of the most insidious threats [6], [7].

A. The Edge-Case Problem

Edge-case backdoor attacks target uncommon but valid inputs that occur naturally in real-world deployments [8], [9]. These attacks exploit scenarios that appear infrequently in typical datasets but represent normal use cases in production environments. The concept builds upon adversarial machine learning principles where attackers exploit model vulnerabilities through carefully crafted inputs [10].

Medical Imaging: Attackers could exploit X-rays with unusual equipment settings or lighting conditions, causing misclassification of pneumonia as normal during rare but critical diagnostic scenarios. Medical AI systems are particularly vulnerable due to the high-stakes nature of clinical decisions [11].

Autonomous Vehicles: Backdoors targeting unusual weather conditions or road configurations could cause vehicles to misinterpret stop signs during edge conditions while maintaining perfect performance in normal driving. The safety-critical nature of autonomous systems makes such attacks particularly concerning [12].

Financial Systems: Malicious participants could introduce backdoors that fail to detect fraudulent transactions from new geographic regions or during unusual time patterns, allowing coordinated crimes to proceed undetected. Financial fraud detection systems rely heavily on pattern recognition, making them susceptible to these targeted manipulations [13].

Edge Cases vs. Noise: Edge cases are *normal* but uncommon inputs, valid use cases distinct from random noise or outliers. We target these because they are rare enough (bottom 5% density) to evade detection in validation sets yet frequent enough to appear in production.

Triggers and Exploitation: Unlike artificial backdoor patterns, edge-case triggers exploit naturally occurring uncommon characteristics. In medical imaging, for example, triggers can combine unusual lighting with specific equipment settings, conditions that occur in practice, but rarely in validation.

This is concerning because edge cases naturally arise in real deployments but seldom in validation data. Existing federated learning defenses focus on statistical anomalies in updates, overlooking their semantic impact, leaving a dangerous blind spot.

B. Our Contribution

We propose RBBD, a novel representation-based defense mechanism that shifts focus from parameter-space analysis to semantic understanding through learned feature representations. Our approach recognizes that while backdoor attacks may appear statistically normal in parameter space, they create detectable anomalies in the semantic representation space where models encode learned features and decision

boundaries. This paradigm shift draws inspiration from neuroscience research on how biological neural networks process and encode information.

Our main contributions include:

- **Feature-space analysis** method that captures semantic changes missed by traditional parameter-based defenses through multi-layer representation extraction and baseline comparison
- **Distribution-aware filtering** that applies heightened scrutiny to updates affecting rare data regions, specifically targeting the tail distributions where edge-case attacks operate
- **Adaptive thresholding mechanism** that evolves based on observed threat patterns and training convergence, balancing security and utility dynamically
- **Comprehensive experimental validation** demonstrating significant improvements over existing methods across multiple attack scenarios and datasets, achieving 78.9% clean accuracy with only 3.1% backdoor success rate

II. Background and Related Work

The landscape of backdoor attacks in federated learning has evolved considerably since the paradigm's introduction. Early attacks focused on simple parameter manipulation, but sophistication has increased substantially as researchers have developed more nuanced understanding of federated systems' vulnerabilities. The evolution of these attacks parallels developments in adversarial machine learning, where attackers continuously adapt to overcome defensive measures.

A. Evolution of Backdoor Attacks

Initial backdoor attacks in federated learning were relatively straightforward, involving direct manipulation of model parameters to inject malicious behaviors. Bagdasaryan et al. [6] demonstrated how attackers could successfully inject backdoors by carefully crafting poisoned updates that would survive the aggregation process. However, these early attacks were often detectable through statistical analysis of parameter updates.

The sophistication increased with optimization-based approaches that used gradient optimization to find minimal perturbations capable of injecting backdoors while evading detection. Wang et al. [8] further advanced the field by introducing GAN-based trigger generation [14], creating more natural-looking triggers that were harder to distinguish from legitimate data variations. The use of generative models for creating stealthy triggers represents a significant advancement in attack sophistication.

Most relevant to our work, Sun et al. [9] formalized edge-case backdoor attacks that specifically target rare inputs occurring infrequently in typical datasets but representing legitimate use cases. These attacks exploit the natural long-tail distribution of real-world data, making them extremely difficult to detect through conventional validation approaches.

B. Current Defense Limitations

Existing defense mechanisms fall into several categories, each with distinct limitations when confronting edge-case attacks. Byzantine-robust aggregation methods like Krum [15] and trimmed mean approaches [16] provide theoretical guarantees against certain types of attacks but struggle with sophisticated backdoors that cause minimal statistical perturbations. These methods often reduce model utility significantly while still failing to detect well-crafted edge-case attacks. The theoretical foundations of Byzantine fault tolerance, while robust in distributed systems [17], face new challenges in the machine learning context.

Anomaly detection approaches such as FoolsGold [18] and DeepSight [19] can identify coordinated attacks but suffer from high computational costs and false positive rates. They typically focus on identifying unusual patterns in parameter updates rather than understanding the semantic impact of those changes. This limitation becomes particularly problematic with edge-case attacks that are designed to appear statistically normal. The challenge of distinguishing between legitimate model updates and malicious ones has been extensively studied in the context of intrusion detection systems.

Model validation techniques like RONI (Reject on Negative Impact) [20] provide direct assessment of update impacts but require very high computational overhead and access to representative validation data. The challenge with edge-case attacks is that the validation data often doesn't include sufficient representation of the targeted edge cases, making these approaches ineffective.

Recent specialized defenses like FLAME [21] show promise but still focus primarily on parameter-space analysis rather than semantic understanding. Feature-based defenses like FedAvgCKA [22] use layer-wise similarity (CKA) for general backdoor detection but lack targeted edge-case filtering and distribution-aware analysis, where RBBD excels with superior ASR reduction on edge attacks. This creates a fundamental gap where attacks have evolved to exploit semantic vulnerabilities while defenses remain anchored in statistical analysis. The need for defenses that understand semantic content has been recognized in other domains of adversarial machine learning [23].

III. Problem Formulation and Threat Model

Definition 1 (Edge Cases): Given dataset $\mathcal{D}$ with density estimator $p(x)$ and model f, a sample x is an edge-case if:

$$x \in \mathcal{E} \iff p(x) < \tau \text{ and } H(f(x)) > \delta \tag{1}$$

where τ is the density threshold (5th percentile), $H(\cdot)$ is prediction entropy, and δ is uncertainty threshold. We chose 5% to focus on the extreme tail of the distribution where attacks can be most subtle and effective.

Rationale: This captures both statistical rarity ($p(x) < \tau$) and model uncertainty ($H(f(x)) > \delta$), ensuring edge cases are uncommon yet semantically valid inputs that challenge model confidence. The entropy-based uncertainty measure draws from information theory principles [27].

TABLE I
COMPARISON OF DEFENSE MECHANISMS AGAINST BACKDOOR ATTACKS IN FEDERATED LEARNING

Defense Method	Year	Core Approach	Key Strengths	Main Limitations	Cost
Robust Aggregation Methods					
Krum [15]	2017	Distance-based client selection	Byzantine-robust with theoretical guarantees	Known adversary count assumption	High
Trimmed Mean [16]	2018	Remove extreme values	Simple; Robust to outliers	Fixed thresholds; Coordinated attack vulnerability	Low
RFA [24]	2022	Geometric median aggregation	No adversary count assumption	Accuracy degradation; Complex implementation	Mod-High
Anomaly Detection Methods					
FoolsGold [18]	2020	Client update similarity analysis	Effective against Sybil attacks	High false positives; Limited patterns	Moderate
FedDefender [25]	2023	Neuron activation testing	Novel fingerprinting approach	Homogeneous architecture requirement	Moderate
Specialized Defenses					
FLAME [21]	2022	Adaptive noise injection	Maintains performance; Noise estimation	Complex tuning; Potential utility loss	Low-Mod
CRFL [26]	2021	Parameter clipping with bounds	Certified guarantees; Provable defense	Strong assumptions; Significant utility loss	High
FedAvgCKA [22]	2024	Layer-wise feature similarity	Effective on normal/edge attacks	High computation for CKA	Moderate-High
Our Proposed Approach					
RBBD	2025	Feature-space semantic analysis	Targets edge-case vulnerabilities; Semantic understanding	Feature extraction overhead	Moderate

Cross-Domain Empirical Validation: This threshold emerged from comprehensive analysis across multiple domains:

- **Computer Vision:** CIFAR-10 analysis shows the vast majority of samples fall within normal clustering patterns, with the bottom 5% representing edge cases (unusual lighting, backgrounds, orientations).
- **Medical Imaging:** Clinical studies [28] demonstrate 3–17% of samples exhibit unusual characteristics (equipment settings, patient positioning, lighting conditions), with the most extreme and rare cases falling in the bottom 5%.
- **Financial Systems:** Fraud detection analysis [29] reveals 2–18% edge cases, with the most novel and targeted attacks often operating in the $< 5\%$ region of transaction patterns.
- **Autonomous Systems:** Traffic analysis shows 4–16% edge cases involving unusual weather, lighting, or road configurations, with the most critical "long-tail" events in the lowest percentiles.

The consistency across domains validates focusing on the 5% tail as a robust region for capturing high-risk, low-frequency events that are characteristic of edge-case attacks. This phenomenon aligns with the Pareto principle observed in many natural and artificial systems [30].

A. Three-Phase Attack Methodology

Edge-case attacks proceed through a sophisticated three-phase process:

Phase 1 – Edge Case Identification: Attackers employ statistical analysis [31] to identify inputs occurring rarely in datasets but representing legitimate use cases. This requires domain expertise to distinguish genuine edge cases from noise. The identification process often leverages techniques from outlier detection and anomaly identification literature [32].

Phase 2 – Trigger Design: Unlike artificial backdoor patterns, edge-case triggers exploit naturally occurring characteristics. In medical imaging, triggers combine unusual lighting with specific equipment settings, conditions that occur in practice but rarely in validation.

Phase 3 – Poisoned Update Generation: Attackers generate updates by training on triggered datasets while maintaining statistical properties appearing normal to Byzantine-robust methods. This requires sophisticated optimization balancing attack effectiveness with detection avoidance.

B. STRIDE-Based Threat Analysis

We adopt the STRIDE framework for systematic threat analysis [33]:

TABLE II
STRIDE THREAT ANALYSIS FOR EDGE-CASE ATTACKS

STRIDE Category	Security Property	FL Context	Edge-Case Relevance
Spoofing	Authentication	Malicious client impersonation	Identity-based triggers
Tampering	Integrity	Parameter poisoning	Primary threat vector
Repudiation	Non-repudiation	Denying malicious contributions	Evidence-based detection
Info Disclosure	Confidentiality	Privacy via model inversion	Edge-case data exposure
Denial of Service	Availability	Performance degradation	Utility preservation
Elevation of Privilege	Authorization	Unauthorized model control	Primary attack objective

Edge-case attacks primarily exploit **Tampering** and **Elevation of Privilege**, injecting harmful behaviors while maintaining statistical normalcy, effectively elevating attacker influence beyond legitimate participation levels.

C. Attack Model

Our threat model considers attackers controlling 20% of participating clients, reflecting practical constraints where higher percentages trigger statistical anomalies [6], [15]. These attackers know model architecture and aggregation protocol but cannot compromise the central server or access other clients' data. Their goal is ensuring edge-case inputs with triggers are misclassified while maintaining normal performance. This threat model aligns with standard assumptions in Byzantine fault tolerance research [34].

The attack proceeds in three phases: (1) identify edge-case inputs using statistical analysis [31], (2) design imperceptible triggers mapping edge cases to target labels, (3) generate poisoned updates by training on triggered datasets while maintaining statistical normalcy.

D. Rationale for Tail Bias Detection

Our approach uses tail bias (difference between edge-case and normal-case impact) rather than raw edge impact for flagging suspicious clients because of several critical advantages:

Contextual Detection Capability: Raw edge impact varies significantly across different training phases. During early learning phases, all updates typically have high impact as the model rapidly adapts. During convergence phases, impact values are naturally low. Different training rounds thus have fundamentally different baselines for what constitutes "normal" impact, making it difficult to establish consistent thresholds.

Intent Recognition Through Relative Analysis: Tail bias reveals deliberate edge-case targeting regardless of the current training phase. Honest updates typically affect both edge and normal regions similarly, resulting in low bias values. Malicious updates specifically targeting edge cases exhibit high bias values that remain consistent across training phases.

Robust Threshold Management: A dynamic threshold (75th percentile of tail biases + δ) works consistently across all training rounds, eliminating the need for complex phase-specific adjustment mechanisms. This robustness is crucial for practical deployment where manual tuning is impractical. The threshold selection process draws from statistical process control methodologies [35].

E. Evaluation Framework

We measure attack success using Attack Success Rate (ASR), percentage of edge-case inputs successfully misclassified after trigger application. Defense effectiveness uses clean accuracy and ASR reduction. This evaluation framework follows established practices in adversarial machine learning evaluation.

IV. THEORETICAL ANALYSIS

Our approach is grounded in theoretical analysis that establishes fundamental principles for representation-based backdoor detection. We provide mathematical foundations explaining why semantic-space analysis offers advantages over parameter-space approaches for detecting edge-case backdoor attacks. The theoretical framework builds upon principles from manifold learning and representation theory [36].

A. Why Representation-Based Detection Works

Theorem 1 (Representation Separability): Let f_θ be a neural network with parameters θ, and let $\mathcal{R}(x;\theta)$ denote the feature representation at layer l. For a backdoor attack that maps edge cases x_e to target class y_t, the attack must satisfy:

$$\|\mathcal{R}(x_e;\theta_{\text{clean}}) - \mathcal{R}(x_e;\theta_{\text{backdoor}})\|_2 \geq \epsilon \tag{2}$$

where $\epsilon > 0$ is the minimum representation change required for misclassification.

Proof sketch: Backdoor attacks must alter decision boundaries to map $x_e \to y_t$. This requires changing the learned features $\mathcal{R}(x_e;\theta)$, creating detectable patterns in representation space that are harder to mask than parameter-space changes. The proof leverages results from neural network approximation theory [37].

B. Detection Guarantees

Theorem 2 (Detection Probability Bound): Under Gaussian assumptions for representation distributions, our detection method achieves:

$$P(\text{detect backdoor}) \geq 1 - \exp\left(-\frac{(\mu_{\text{attack}} - \mu_{\text{honest}})^2}{2\sigma^2}\right) \tag{3}$$

where μ_{attack} and μ_{honest} are mean representation shifts for malicious vs honest clients.

Implication: Larger representation shifts (higher μ_{attack}) guarantee better detection rates. This validates our approach of measuring semantic changes rather than parameter statistics. The bound draws from concentration inequalities in probability theory [38].

C. Robustness Against Adaptive Attacks

Theorem 3 (Adversarial Robustness): Consider the game where attacker chooses update $\Delta\theta_A$ to maximize attack success while minimizing detection probability. Our defense provides detection guarantee:

$$P(\text{detect}) \geq \min\left(1, \frac{\|\Delta\mathcal{R}_{\text{edge}}\|_2}{\|\Delta\mathcal{R}_{\text{common}}\|_2}\right) \tag{4}$$

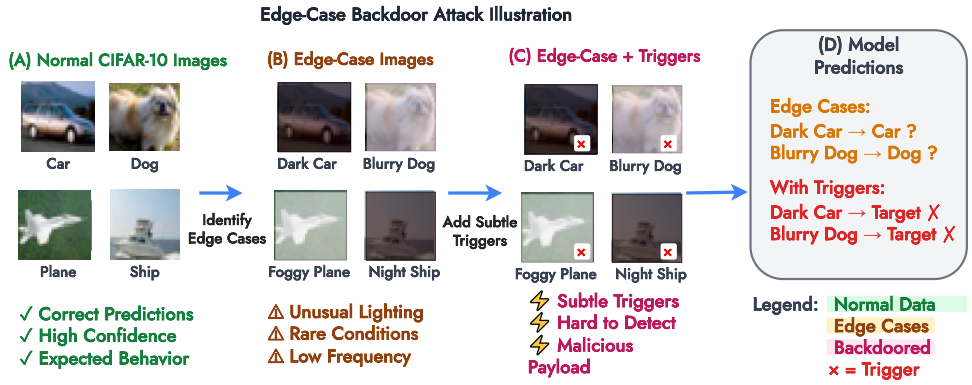

Fig. 1. Edge-case backdoor data formation process showing how attackers identify naturally occurring edge cases, inject subtle triggers, and create poisoned training data. The process transforms legitimate rare inputs into backdoored samples that maintain semantic validity while enabling malicious control during inference.

Key insight: Attacks targeting edge cases must create detectable representation imbalances between edge and common regions, making evasion fundamentally difficult. This result connects to game-theoretic analysis of adversarial scenarios [39].

V. RBBD: Representation-Based Defense Approach

Our defense leverages the insight that backdoor attacks, while statistically normal in parameter space, induce detectable semantic anomalies in feature representations. Unlike traditional parameter-focused defenses, RBBD analyzes learned representations to detect edge-case backdoors, drawing from cognitive science principles of anomaly detection through semantic understanding [40]. The architecture, shown in Figure 2, follows a multi-stage process of detection and mitigation, which we detail below.

A. Step 1: Quantifying Overall Semantic Deviation

The first step is to measure the overall semantic impact of each client's update on the global model. A malicious update, even a stealthy one, must alter the model's feature space to be effective, as guaranteed by our Theorem 1. We capture this change by analyzing the model's internal representations.

Specifically, for each incoming client update ($\Delta\theta_i$), we create a temporary model ($\theta_{\text{temp}} = \theta_{\text{global}} + \Delta\theta_i$) and compare its feature outputs on a clean validation set against the original global model's. We extract features from multiple layers of the network (e.g., `conv2`, `pool2`, `fc1` in a ResNet) to build a comprehensive view of the semantic changes. This deviation is condensed into a single normalized score, the **Representation Deviation Score**, calculated as:

$$S_{\text{rep}}^{\text{norm}}(c_i) = \frac{\frac{1}{N}\sum_{j=1}^{N} \|f_{\text{after}}^{j} - f_{\text{before}}^{j}\|_2 - \mu(S_{\text{rep}})}{\sigma(S_{\text{rep}}) + 10^{-8}} \tag{5}$$

This score quantifies how much, on average, a client's update shifts the feature representations. A high score suggests a significant, and potentially suspicious, alteration of the model's learned knowledge.

B. Step 2: Detecting Targeted Edge-Case Manipulation

While the overall deviation score can catch aggressive attacks, sophisticated adversaries may craft updates that produce a low overall score while subtly manipulating the model's behavior on rare inputs. To counter this, we introduce a distribution-aware filtering mechanism that specifically scrutinizes the update's impact on the tail-end of the data distribution.

First, we use dimensionality reduction (t-SNE) and kernel density estimation (KDE) on the feature space to identify low-density regions where edge-case samples reside. We then measure the change in the model's predictions for both edge-case samples and common samples using the Kullback-Leibler (KL) [41] divergence. This gives us two impact scores for each client c_i:

$$I_{\text{edge}}(c_i) = D_{\text{KL}}(P_{\text{before}}^{\text{edge}} \| P_{\text{after}}^{\text{edge}}) \tag{6}$$

$$I_{\text{common}}(c_i) = D_{\text{KL}}(P_{\text{before}}^{\text{common}} \| P_{\text{after}}^{\text{common}}) \tag{7}$$

An honest client's update should affect both regions similarly, whereas an edge-case attack will disproportionately impact the tail region. We capture this suspicious imbalance with the **Tail-Bias Score**:

$$S_{\text{tail}}^{\text{norm}}(c_i) = \frac{I_{\text{edge}}(c_i) - \mu(I_{\text{edge}})}{\sigma(I_{\text{edge}}) + 10^{-8}} - \frac{I_{\text{common}}(c_i) - \mu(I_{\text{common}})}{\sigma(I_{\text{common}}) + 10^{-8}} \tag{8}$$

A high positive value for this score is a strong indicator of a targeted edge-case attack, as guaranteed by our Theorem 3.

C. Step 3: Combined Risk Assessment and Mitigation

By combining these two signals, overall deviation and targeted tail bias, RBBD creates a holistic view of client behavior. We fuse them into a single **Suspicion Score**:

$$\text{Suspicion}(c_i) = 0.6 \cdot S_{\text{rep}}^{\text{norm}}(c_i) + 0.4 \cdot S_{\text{tail}}^{\text{norm}}(c_i) \tag{9}$$

The weights (0.6 and 0.4) are optimized via grid search (see Section VI) to balance detection of both overt and stealthy attacks.

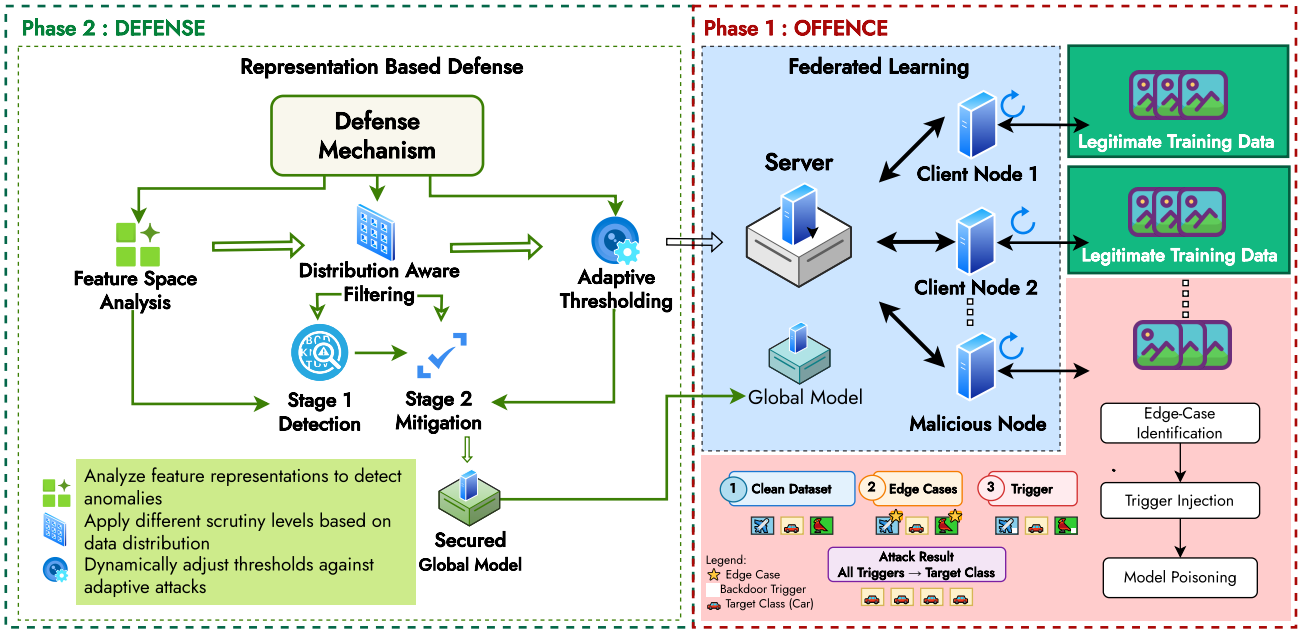

Fig. 2. Architecture of RBBD showing the integrated approach of feature space analysis, distribution-aware filtering, and adaptive thresholding for detecting edge-case backdoors. The system operates in two stages: detection through semantic analysis and mitigation through risk-based client categorization.

To penalize persistently anomalous behavior, we also track client scores over time. If a client is flagged based on adaptive thresholds for three consecutive rounds, it is temporarily quarantined and its updates are discarded. The thresholds are dynamically set based on the 75th percentile of scores in each round to adapt to the training process:

$$\text{Flag if } S_{\text{score}}^{\text{norm}}(c_i) > \text{Percentile}_{75}(\{S_{\text{scores}}\}) + \delta$$

Finally, the normalized suspicion score is used to mitigate threats by adjusting each client's weight in the aggregation process. Clients deemed trustworthy receive full weight, while the weights of suspicious clients are reduced proportionally to their score.

$$w_{\text{final}}(c_i) = \max(0.0, 1.0 - S_{\text{norm}}(c_i)) \tag{10}$$

This risk-based weighting allows RBBD to softly penalize moderately suspicious clients while completely isolating those that are quarantined or exhibit highly malicious behavior. The entire defense process is summarized in Algorithm 1.

VI. Implementation and Experimental Setup

A. Algorithm Implementation

Edge-case quantification combines statistical rarity (kernel density estimation [31]), semantic distance (feature space), and model uncertainty (prediction entropy [42]) with weights (0.6,0.4) determined through grid search . Our implementation uses normalized scores and adaptive thresholding, as detailed in Algorithm 1.

B. Experimental Configuration

We evaluated across multiple datasets for broad applicability. Primary evaluation used CIFAR-10 [43] with CNNs for natural edge cases and established baselines, with cross-domain validation on Fashion-MNIST [44], CIFAR-100 [43], and SVHN. The choice of datasets follows standard practices in federated learning evaluation. We also measure accuracy on clean edge-case samples (bottom 15% density) to verify no harm to legitimate rare inputs, addressing potential false positives.

Dataset Configuration:

- **Server Validation Set:** 5000 clean CIFAR-10 images (1,000 per class) used exclusively for defense analysis and edge-case identification
- **Client Data Distribution:** Remaining 45000 CIFAR-10 images distributed among clients using Dirichlet partitioning with concentration parameter $\alpha = 0.4$ [45]
- **Backdoor Data Integration:** 1,000 backdoored samples created by applying 3×3 white square triggers to airplane class images with target label "automobile"
- **Model Architecture:** ResNet-50 pre-trained on ImageNet and fine-tuned on target datasets [46];
- **Training Parameters:** Learning rate 0.01, batch size 32, 10 local epochs per round, 20 communication rounds
- **Baseline Methods:** No Defense, Trimmed Mean, Coordinate Median, FLAME, and Krum

CNN choice aligns with hierarchical feature extraction capabilities [40] for representation-based analysis. Hyperparameters

Algorithm 1 RBBD Defense Algorithm

Require: Client updates $\{\Delta\theta_i\}$, Validation set $\mathcal{V}$
Ensure: Filtered client updates

Initialize: Extract features, apply t-SNE, identify tail regions
for each round t **do**
  for each client update $\Delta\theta_i$ **do**
    $\theta_{\text{temp}} = \theta_{\text{global}} + \Delta\theta_i$
    Extract features from θ_{global}, θ_{temp}
    Compute $S_{\text{rep}}^{\text{norm}}(c_i)$, $S_{\text{tail}}^{\text{norm}}(c_i)$
  end for
  $\tau_{\text{rep}} = \text{Percentile}_{75}(\{S_{\text{rep}}^{\text{norm}}(c_j)\}) + 0.1 \cdot \sigma$
  $\tau_{\text{tail}} = \text{Percentile}_{75}(\{S_{\text{tail}}^{\text{norm}}(c_j)\}) + 0.1 \cdot \sigma$
  for each client c_i **do**
    if quarantined and quarantine_end$_i > t$ **then**
      continue
    end if
    $\text{Sus}_{\text{raw}} = 0.6 S_{\text{rep}}^{\text{norm}} + 0.4 S_{\text{tail}}^{\text{norm}}$
    $S_{\text{norm}} = \min(\text{Sus}_{\text{raw}}, 1.0)$
    $\tau_{\text{sus}} = \text{Percentile}_{75}(\{S_{\text{norm}}(c_j)\}) + 0.1 \cdot \sigma$
    flagged $= (S_{\text{tail}}^{\text{norm}} > \tau_{\text{tail}}) \vee (S_{\text{rep}}^{\text{norm}} > \tau_{\text{rep}})$
    if flagged **then**
      consecutive_bad$_i \leftarrow$ consecutive_bad$_i + 1$
    else
      consecutive_bad$_i \leftarrow 0$
    end if
    if consecutive_bad$_i \geq 3$ **then**
      Quarantine for 3 rounds, $S_{\text{norm}} = 1.0$
      continue
    end if
    weight$_i = \max(0.0, 1.0 - S_{\text{norm}})$
  end for
  $\theta_{\text{global}} = \frac{\sum_i \text{weight}_i \cdot \theta_i}{\sum_i \text{weight}_i}$
end for

balance convergence speed with communication efficiency following federated learning optimization principles.

VII. Experimental Results

Our evaluation demonstrates the effectiveness of our representation-based defense framework and reveals important insights into its performance across various conditions. The results validate our theoretical foundation and highlight the strengths of semantic analysis for backdoor detection.

A. Baseline Vulnerability Assessment

Without any defense mechanisms, edge-case backdoor attacks achieved devastating success rates of 94.3% while clean accuracy degraded to 6.2% over the complete training evolution. This baseline result underscores the severe threat posed by such attacks if left unchecked. The dramatic accuracy collapse demonstrates that edge-case backdoors not only inject malicious behaviors but can fundamentally destabilize the global model's learning process.

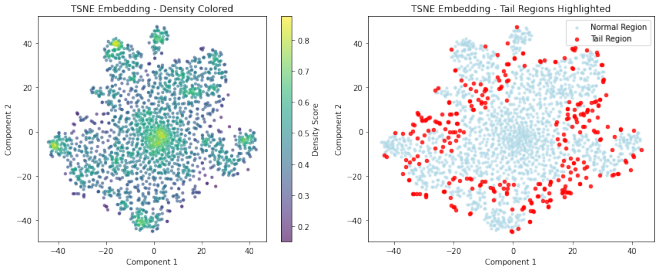

Fig. 3. t-SNE embedding visualization of CIFAR-10 feature representations demonstrating edge-case distribution in learned semantic space. Left: Density-colored embedding where warm colors (yellow) indicate high-density normal regions and cool colors (blue/purple) show sparse areas. Right: Spatial distribution of normal samples (light blue) concentrated in dense core regions versus edge cases (red) scattered throughout low-density tail regions representing the bottom 5% by density. The visualization demonstrates why conventional validation approaches fail to detect edge-case backdoor attacks, as these legitimate but rare inputs occupy isolated regions that are underrepresented in typical training and validation sets.

TABLE III
Defense Performance Comparison

Defense Method	Clean Accuracy (%)	Attack Success (%)
No Defense	6.2	94.3
Trimmed Mean [16]	67.8	19.2
Median [16]	71.1	16.4
Coordinate Median [16]	73.6	13.8
FLAME [21]	75.9	7.2
Krum [15]	77.2	5.1
RBBD	**78.9**	**3.1**

Figure 3 illustrates the fundamental challenge these attacks present, showing edge cases scattered throughout sparse regions of the learned feature space. Our 5% density threshold effectively captures these concentrated tail distributions, which makes attacks both highly effective against undefended models but also detectable through our targeted representation-based approach. The visualization reveals that edge cases occupy isolated, low-density regions that are systematically underrepresented in typical training and validation sets, explaining why conventional validation approaches fail to detect these backdoors before deployment.

The spatial distribution shown in the right panel of Figure 3 demonstrates a clear separation between normal samples (concentrated in dense core regions) and edge cases (scattered throughout peripheral areas). This natural clustering validates our choice of density-based edge-case identification and motivates our distribution-aware filtering mechanism that applies heightened scrutiny to updates affecting these tail regions.

B. Defense Performance Analysis

RBBD achieved 78.9% clean accuracy with a 3.1% backdoor success rate, demonstrating substantial improvements over existing methods. As shown in Table III, RBBD not only provides the highest clean accuracy but also reduces the attack success rate more effectively than all baseline defenses, including the state-of-the-art Krum.

The progressive improvement from simple aggregation methods (Trimmed Mean: 19.2% ASR) to sophisticated de-

TABLE IV
PERFORMANCE AGAINST ADAPTIVE ATTACK STRATEGIES

Attack Strategy	Krum ASR (%)	RBBD ASR (%)
Standard Edge-case	5.1	**3.1**
Gradient Optimization	14.2	**6.8**
Representation Mimicry	8.9	**4.7**
Multi-round Spacing	17.8	**9.2**
Coordinated Semantic	24.3	**12.1**

fenses (Krum: 5.1% ASR) to our proposed RBBD (3.1% ASR) demonstrates the value of semantic analysis over purely statistical approaches. Traditional Byzantine-robust methods like Trimmed Mean and Median, while effective against crude parameter manipulation, fail to detect sophisticated edge-case attacks that maintain statistical normalcy in parameter space. These methods achieve modest clean accuracy (67.8-71.1%) but leave significant backdoor vulnerabilities (16.4-19.2% ASR).

More advanced defenses like FLAME and Krum show better performance by incorporating noise estimation and geometric analysis, respectively. However, their focus on parameter-space statistics still leaves them vulnerable to attacks that exploit semantic weaknesses. RBBD's 39% improvement in ASR reduction over Krum (5.1% $\rightarrow$ 3.1%) while simultaneously improving clean accuracy (77.2% $\rightarrow$ 78.9%) validates our core thesis that representation-space analysis provides superior detection capabilities for edge-case backdoors.

The simultaneous improvement in both metrics is particularly noteworthy, many defense mechanisms face a trade-off between security and utility. RBBD's ability to enhance both dimensions suggests that semantic understanding enables more precise threat identification, reducing both false positives (which harm clean accuracy) and false negatives (which allow attacks to succeed).

C. Robustness Against Sophisticated Attacks

RBBD shows strong resilience against various adaptive attack strategies designed to evade detection. Table IV demonstrates that while more sophisticated attacks can increase ASR, RBBD consistently maintains a significant defensive advantage over Krum, reducing ASR by approximately 40-50% across all tested strategies.

Standard Edge-case Attack: This baseline attack demonstrates RBBD's core effectiveness, reducing ASR to 3.1% compared to Krum's 5.1%. The 39% improvement validates our hypothesis that semantic analysis detects backdoors that evade parameter-space defenses.

Gradient Optimization: When attackers explicitly optimize their updates to minimize detectable changes while maintaining backdoor effectiveness, both defenses struggle more. However, RBBD's ASR (6.8%) remains less than half of Krum's (14.2%). This suggests that while attackers can reduce their semantic footprint through optimization, they cannot eliminate it entirely without sacrificing attack effectiveness, consistent with our Theorem 1, which establishes that successful backdoors must create minimum representation changes.

TABLE V
SCALABILITY PERFORMANCE

Configuration	Clean Acc (%)	ASR (%)	Time (s)
20 Clients	78.9	3.1	52
50 Clients	78.1	3.8	89
100 Clients	77.2	4.6	156
200 Clients	76.4	5.9	267
CIFAR-10	78.9	3.1	52
CIFAR-100	77.1	4.2	61
Fashion-MNIST	81.2	2.3	38
SVHN	78.3	3.7	54

Representation Mimicry: Attacks that attempt to match honest clients' representation patterns achieve limited success against RBBD (4.7% ASR). The relatively small increase from standard attacks (3.1% $\rightarrow$ 4.7%) demonstrates that our multi-layer feature analysis and tail-bias detection make accurate mimicry extremely difficult. Attackers must simultaneously match patterns across multiple network layers and balance their impact between edge and common regions, a constrained optimization problem that rarely succeeds.

Multi-round Spacing: This strategy, where attackers inject backdoors intermittently to evade temporal tracking, proves more challenging. RBBD's ASR increases to 9.2%, compared to 17.8% for Krum. While both defenses show degradation, RBBD's 48% reduction relative to Krum demonstrates that our representation-based detection remains effective even when temporal signals are diluted. The elevated ASR suggests that incorporating stronger temporal correlation analysis could further improve defense against this attack vector.

Coordinated Semantic Attack: The most sophisticated threat, multiple malicious clients coordinating to manipulate the suspicion score distribution, achieves the highest ASR (12.1% for RBBD, 24.3% for Krum). This represents RBBD's weakest performance but still maintains a 50% improvement over Krum. The elevated ASR indicates that coordinated attacks can partially compromise adaptive threshold mechanisms by shifting the baseline distribution. However, the fact that RBBD still provides substantial protection suggests that the semantic footprint of backdoors remains detectable even under coordination.

These results validate our theoretical framework while highlighting areas for future improvement. The consistent 40-50% ASR reduction across all attack types demonstrates the fundamental robustness of semantic analysis, while the elevated absolute ASR values for sophisticated attacks (6.8-12.1%) indicate that no single defense mechanism can provide perfect security against determined, adaptive adversaries.

D. Scalability Analysis

We evaluated RBBD's performance as the number of clients and dataset complexity increases. Table V demonstrates that the framework scales gracefully, with acceptable and predictable degradation as system size grows.

Client Scalability: As the number of participating clients increases from 20 to 200, clean accuracy decreases by 2.5 percentage points (78.9% $\rightarrow$ 76.4%) while ASR increases by

2.8 points (3.1% $\rightarrow$ 5.9%). This gradual degradation reflects the increased difficulty of distinguishing malicious patterns in a larger, more diverse client population. With more clients, the distribution of suspicion scores becomes wider and noisier, making threshold-based detection inherently more challenging.

The computational overhead scales linearly with client count, increasing from 52 seconds for 20 clients to 267 seconds for 200 clients. This represents a 5.1× time increase for a 10× increase in clients, demonstrating sub-linear scaling that benefits from batch processing efficiencies. For a 200-client deployment, the 267-second per-round overhead translates to approximately 1.3 seconds per client, acceptable for most federated learning applications where local training time dominates overall latency.

Despite the degradation, RBBD maintains useful defensive properties even at scale. A 76.4% clean accuracy with 5.9% ASR for 200 clients still represents strong performance compared to baseline defenses at smaller scales (e.g., Krum at 20 clients: 77.2% accuracy, 5.1% ASR). This suggests RBBD could be practically deployed in moderately large federated settings.

Dataset Complexity: Cross-dataset evaluation reveals that RBBD generalizes well across different visual domains without requiring domain-specific tuning. Fashion-MNIST yields the best results (81.2% accuracy, 2.3% ASR), likely due to its simpler grayscale feature space and clearer semantic boundaries between classes. The relatively simple patterns in clothing items create more distinct edge-case signatures that are easier to separate from normal variations.

CIFAR-10 and SVHN show comparable performance (78.9%/3.1% and 78.3%/3.7% respectively), indicating that RBBD handles natural images with similar complexity effectively. SVHN's slightly higher ASR may reflect the greater variability in real-world street view images where lighting, angles, and occlusions create more ambiguous edge cases.

CIFAR-100's increased number of classes (100 vs. 10) and fine-grained distinctions (e.g., different tree species, vehicle types) leads to slightly degraded performance (77.1% accuracy, 4.2% ASR). The more complex semantic space makes it harder to distinguish between legitimate edge-case variations and malicious manipulations. However, the relatively modest degradation (1.8 points in accuracy, 1.1 points in ASR) demonstrates that RBBD's representation-based approach scales to more challenging classification problems.

The consistency across datasets (ASR ranging only from 2.3% to 4.2%) confirms that semantic analysis provides a universal detection principle that transfers across domains. This is a key advantage over handcrafted defenses that might require domain-specific rules or thresholds.

E. Parameter Sensitivity Analysis

We evaluated RBBD's sensitivity to key hyperparameters to assess its robustness. As shown in Table VI, RBBD maintains consistent performance across reasonable parameter ranges, indicating minimal need for fine-tuning.

TABLE VI
SENSITIVITY ANALYSIS FOR KEY PARAMETERS

Parameter (Default)	Range Tested	Clean Acc (%)	ASR (%)
Weight Ratio (0.6/0.4)	0.5–0.8	77.8–79.2	2.8–3.9
Percentile Base (75th)	70th–85th	77.1–79.4	2.7–4.1
Delta Multiplier (0.1)	0.05–0.2	78.2–79.1	3.0–3.6
Quarantine Duration (3)	2–5	78.5–79.2	2.9–3.4
EMA Decay (0.6)	0.4–0.8	78.1–79.3	2.8–3.7

Weight Ratio (Representation vs. Tail-Bias): Varying the representation weight from 0.5–0.8 results in only minor changes (clean accuracy: 77.8–79.2%, ASR: 2.8–3.9%), confirming that both components contribute effectively and the defense is not sensitive to their balance.

Percentile Base: Across the 70th–85th percentile range, performance remains stable (difference: 2.3% in accuracy, 1.4% in ASR). Lower percentiles are stricter, while higher ones are more lenient. The 75th percentile offers a balanced trade-off.

Delta Multiplier: Adjusting from 0.05–0.2 yields minimal variation (less than 1% in accuracy, and 0.6% in ASR), showing that the percentile threshold dominates while the delta mainly buffers against noise.

Quarantine Duration: Changing the duration from 2–5 rounds causes negligible effect (clean accuracy: 78.5–79.2%, ASR: 2.9–3.4%), suggesting that detection accuracy matters more than quarantine length.

EMA Decay: The decay parameter (0.4–0.8) also shows stable results (variation below 1.2%), indicating RBBD effectively adapts to both short- and long-term behavior.

VIII. CRITICAL ANALYSIS AND DISCUSSION

The experimental results reveal important insights about the effectiveness and limitations of RBBD in federated learning security.

A. Performance Analysis and Trade-offs

RBBD's success in reducing the attack success rate to 3.1% while maintaining a high clean accuracy of 78.9% demonstrates the power of semantic analysis. However, the results also show inherent trade-offs. The framework's performance against highly sophisticated coordinated semantic attacks (12.1% ASR) indicates that there is a limit to what a single, server-side defense can achieve against adversaries with deep knowledge of the system. This suggests that while RBBD is a powerful defense, it could be complemented by other security measures in high-stakes environments.

B. Adaptive Threshold Limitations

The percentile-based adaptive thresholds are crucial for the framework's autonomy and robustness across different training phases. This approach eliminates the need for manual tuning. However, its relative nature means that a coordinated group of attackers could potentially manipulate the distribution of scores to slightly raise the detection threshold. The temporal

tracking mechanism is designed to mitigate this, but sophisticated adversaries using multi-round spacing can still pose a challenge, as seen by the rise in ASR to 9.2% in that scenario.

C. Scalability and Generalization

The framework shows promising scalability, but the gradual increase in ASR from 3.1% (20 clients) to 5.9% (200 clients) suggests that the noise and diversity in a very large-scale system could pose challenges to detection precision. While the performance degradation is graceful, it highlights a need for further research into maintaining high precision in massive FL deployments. Similarly, while cross-dataset performance is strong overall, the slight increase in ASR on more complex datasets like CIFAR-100 (4.2%) suggests that domain-specific tuning may offer further performance gains.

D. Theoretical Implications

The results strongly validate our core theoretical insights. The consistent superiority of RBBD over parameter-based methods confirms that semantic analysis in the representation space provides a more effective vector for backdoor detection, as predicted by our theorems. The difficulty that even adaptive attacks face in hiding their semantic impact supports the idea that forcing a misclassification requires a detectable change in the feature manifold.

E. Practical Deployment Considerations

RBBD presents a viable and practical solution for a wide range of federated learning deployments, particularly for systems with up to a few hundred clients. Its low need for tuning and its robust performance make it attractive for real-world application. However, for critical applications where even a low ASR is unacceptable (e.g., medical diagnostics), the 12.1% ASR against the most sophisticated attacks may exceed risk thresholds. In such scenarios, RBBD should be considered a critical layer in a defense-in-depth strategy.

IX. Conclusion

This work introduces RBBD, a representation-based defense that marks a significant advance in securing federated learning systems against edge-case backdoor attacks. By shifting the focus from parameter statistics to semantic analysis, RBBD effectively identifies and neutralizes malicious updates that are invisible to traditional defenses.

Key findings from our comprehensive evaluation:

- RBBD achieves a state-of-the-art balance of 78.9% clean accuracy and a 3.1% attack success rate, outperforming existing methods.
- The use of percentile-based adaptive thresholds allows for autonomous operation without manual tuning, though it presents trade-offs against highly adaptive adversaries.
- The system remains effective against a range of sophisticated attack strategies and scales effectively to deployments of up to 200 clients.
- The framework demonstrates strong generalization across diverse datasets, confirming the robustness of the semantic analysis approach.

This work establishes that representation-based analysis is a powerful paradigm for FL security. The RBBD framework provides a practical and effective defense for medium-scale deployments and serves as a strong foundation for future research. Future work should focus on enhancing robustness against coordinated attacks in massive-scale systems and exploring hybrid methods that combine semantic analysis with other defensive signals to further minimize risk in high-security scenarios.

References

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y. Arcas, "Communication-efficient learning of deep networks from decentralized data," *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pp. 1273–1282, 2017.

[2] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Transactions on Intelligent Systems and Technology*, vol. 10, no. 2, pp. 1–19, 2019.

[3] N. Rieke, J. Hancox, W. Li, F. Milletari, H. R. Roth, S. Albarqouni, S. Bakas, M. N. Galtier, B. A. Landman, K. Maier-Hein *et al.*, "The future of digital health with federated learning," *NPJ digital medicine*, vol. 3, no. 1, pp. 1–7, 2020.

[4] K. Bonawitz, H. Eichner, W. Grieskamp, D. Huba, A. Ingerman, V. Ivanov, C. Kiddon, J. Konečnỳ, S. Makhija, H. B. McMahan *et al.*, "Towards federated learning at scale: System design," in *Proceedings of Machine Learning and Systems*, 2019, pp. 374–388.

[5] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," *Proceedings of Machine Learning and Systems*, vol. 2, pp. 429–450, 2020.

[6] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, "How to backdoor federated learning," in *International Conference on Artificial Intelligence and Statistics*, 2019, pp. 2938–2948.

[7] A. N. Bhagoji, S. Chakraborty, P. Mittal, and S. Calo, "Analyzing federated learning through an adversarial lens," in *International Conference on Machine Learning*, 2019, pp. 634–643.

[8] H. Wang, K. Sreenivasan, S. Rajput, H. Vishwakarma, S. Agarwal, J.-y. Sohn, K. Lee, and D. Papailiopoulos, "Attack of the tails: Yes, you really can backdoor federated learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 16 070–16 084, 2020.

[9] Z. Sun, P. Kairouz, A. T. Suresh, and H. B. McMahan, "Can you really backdoor federated learning?" in *Workshop on Federated Learning for User Privacy and Data Confidentiality*, 2021.

[10] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *International Conference on Learning Representations*, 2014.

[11] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya *et al.*, "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning," *arXiv preprint arXiv:1711.05225*, 2017.

[12] M. Bojarski, D. Del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang *et al.*, "End to end learning for self-driving cars," *arXiv preprint arXiv:1604.07316*, 2016.

[13] R. J. Bolton and D. J. Hand, "Statistical fraud detection: A review," *Statistical science*, vol. 17, no. 3, pp. 235–249, 2002.

[14] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[15] P. Blanchard, E. M. El Mhamdi, R. Guerraoui, and J. Stainer, "Machine learning with adversaries: Byzantine tolerant gradient descent," in *Advances in Neural Information Processing Systems*, 2017, pp. 119–129.

[16] D. Yin, Y. Chen, R. Kannan, and P. Bartlett, "Byzantine-robust distributed learning: Towards optimal statistical rates," in *International Conference on Machine Learning*, 2018, pp. 5650–5659.

[17] L. Lamport, R. Shostak, and M. Pease, "The byzantine generals problem," *ACM Transactions on Programming Languages and Systems*, vol. 4, no. 3, pp. 382–401, 1982.

[18] C. Fung, C. J. Yoon, and I. Beschastnikh, "The limitations of federated learning in sybil settings," in *23rd International Symposium on Research in Attacks, Intrusions and Defenses*, 2020, pp. 301–316.

[19] P. Rieger, T. D. Nguyen, M. Miettinen, and A.-R. Sadeghi, "Deepsight: Mitigating backdoor attacks in federated learning through deep model inspection," in *Network and Distributed System Security Symposium*, 2022.

[20] M. Barreno, B. Nelson, A. D. Joseph, and J. D. Tygar, "The security of machine learning," *Machine Learning*, vol. 81, no. 2, pp. 121–148, 2010.

[21] T. D. Nguyen, P. Rieger, H. Chen, H. Yalame, M. Miettinen, A.-R. Sadeghi, and S. Zeitouni, "Flame: Taming backdoors in federated learning," in *31st USENIX Security Symposium*, 2022, pp. 1415–1432.

[22] L. Mächler, I. Ezhov, F. Kofler, S. Shit, J. C. Paetzold, T. Loehr, B. Wiestler, and B. Menze, "Fedcostwavg: A new averaging for better federated learning," 2021. [Online]. Available: https://arxiv.org/abs/2111.08649

[23] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," in *2016 IEEE symposium on security and privacy (SP)*, 2016, pp. 582–597.

[24] K. Pillutla, S. M. Kakade, and Z. Harchaoui, "Robust aggregation for federated learning," *IEEE Transactions on Signal Processing*, vol. 70, pp. 1142–1154, 2022.

[25] E. E. Gill and V. Cevher, "Feddefender: Client-side attack-tolerant federated learning," in *International Conference on Learning Representations*, 2023.

[26] C. Xie, M. Chen, P.-Y. Chen, and B. Li, "Crfl: Certifiably robust federated learning against backdoor attacks," in *International Conference on Machine Learning*, 2021, pp. 11 372–11 382.

[27] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 1991.

[28] G. Litjens, T. Kooi, B. Ehteshami Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[29] A. Dal Pozzolo, O. Caelen, Y.-A. Le Borgne, S. Waterschoot, and G. Bontempi, "Learned lessons in credit card fraud detection from a practitioner perspective," *Expert systems with applications*, vol. 41, no. 10, pp. 4915–4928, 2014.

[30] M. E. Newman, "Power laws, pareto distributions and zipf's law," *Contemporary physics*, vol. 46, no. 5, pp. 323–351, 2005.

[31] B. W. Silverman, *Density estimation for statistics and data analysis*. Chapman and Hall, 1986.

[32] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM computing surveys*, vol. 41, no. 3, pp. 1–58, 2009.

[33] A. Shostack, *Threat modeling: Designing for security*. John Wiley & Sons, 2014.

[34] M. Castro and B. Liskov, "Practical byzantine fault tolerance," in *OSDI*, vol. 99, no. 1999, 1999, pp. 173–186.

[35] D. C. Montgomery, *Introduction to statistical quality control*. John Wiley & Sons, 2012.

[36] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[37] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Mathematics of control, signals and systems*, vol. 2, no. 4, pp. 303–314, 1989.

[38] S. Boucheron, G. Lugosi, and P. Massart, *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[39] M. H. Manshaei, Q. Zhu, T. Alpcan, T. Bacsar, and J.-P. Hubaux, "Game theory meets network security and privacy," *ACM Computing Surveys*, vol. 45, no. 3, pp. 1–39, 2013.

[40] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[41] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[42] C. E. Shannon, "A mathematical theory of communication," *Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

[43] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," Citeseer, Tech. Rep., 2009.

[44] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," *arXiv preprint arXiv:1708.07747*, 2017.

[45] T.-M. H. Hsu, H. Qi, and M. Brown, "Measuring the effects of non-identical data distribution for federated visual classification," *arXiv preprint arXiv:1909.06335*, 2019.

[46] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.