The FLResilience Benchmark: A Systematic Framework for Evaluating Federated Learning Robustness to System-Level Failures

Laura Tran-Dubois

Vietnam National University Hanoi
Hanoi,Vietnam

Introduction

Introduced by McMahan et al. (2017), Federated Learning (Federated Learning) has emerged as a promising paradigm for training machine learning models across decentralized data sources without sharing raw data. This privacy-preserving characteristic makes Federated Learning particularly appealing for critical applications in domains such as healthcare, finance, and autonomous systems, where data sovereignty and regulatory compliance are paramount. The canonical Federated Learning algorithm, FEDAVG, operates by repeatedly aggregating model updates from a subset of clients to refine a global model. However, the vast majority of Federated Learning research operates under a critical and often unrealistic assumption: a benign and perfectly reliable distributed system. In real-world deployments, especially in cross-silo settings where clients are few but important (e.g., hospitals or banks), system-level failures are not merely possibilities but regular occurrences that can derail the entire learning process.

We define **system-level failures** as any disruption originating from the distributed environment that impedes the learning process. These failures are categorized into two primary classes. First, **benign failures** include client **dropout**, where a client fails to submit an update in a communication round, and **stragglers**, where clients are significantly delayed, slowing down convergence. Second, and more severely, are **Byzantine failures** (Lamport, Shostak, and Pease 1982), where a subset of clients may be malicious or compromised, sending arbitrary or carefully crafted malicious updates in an attempt to sabotage the global model. The data distribution across clients is often **non-I.I.D. (non-I.I.D.)**, meaning each client's local dataset is not a representative sample of the overall population, which exacerbates the vulnerability of aggregation algorithms to these failures. Standard algorithms like FEDAVG are notoriously fragile under these conditions, as a single malicious update or a consistent pattern of dropouts can lead to model divergence, catastrophic performance loss, or complete training failure (Li 2025b).

Despite the proliferation of robust aggregation algorithms designed to mitigate these issues, the research community lacks a standardized, systematic, and reproducible framework for evaluating their comparative effectiveness under a comprehensive suite of failure scenarios. Existing evaluations are often ad-hoc, testing against a limited set of attacks or failures, and conducted on different datasets and setups, making it difficult for practitioners to select the right algorithm for their specific operational constraints. This gap between algorithmic proposals and practical deployment guidance is a significant bottleneck in adopting Federated Learning for critical applications. To bridge this gap, we propose the FLRESILIENCE Benchmark, a comprehensive framework for systematically evaluating the robustness of Federated Learning aggregation strategies. FLRESILIENCE integrates a wide range of simulated system-level failures with standardized datasets, models, and, most importantly, a multi-faceted set of evaluation metrics that go beyond final accuracy to capture stability and resilience.

Literature Review

Early approaches adapted robust statistics from distributed optimization. Blanchard et al. (2017) proposed KRUM and MULTI-KRUM, which select the client update that is closest to its neighbors, theoretically guaranteeing convergence under a minority of Byzantine attackers. Similarly, Yin et al. (2018) introduced the COORDINATE-WISE MEDIAN and TRIMMED MEAN, which aggregate by taking the median or a trimmed mean of each model parameter across clients, offering robustness against outliers.

Subsequent work sought to improve upon these foundations. Chen, Wang, and Yang (2017) proposed BULYAN, a meta-aggregation method that combines the outputs of KRUM to further enhance resilience. To address the challenge of Byzantine failures in the context of non-I.I.D. data, Fung, Yoon, and Beschastnikh (2020) introduced FOOLSGOLD, which uses historical update information to identify and down-weight potentially malicious clients based on the diversity of their contributions. Beyond security, other works have addressed benign system heterogeneity. Li, He, and Song (2020) developed FEDPROX, which adds a proximal term to the local objective function to handle statistical and systems heterogeneity, making it more tolerant to stragglers. More recently, Karimireddy et al. (2020) introduced SCAFFOLD, which uses control variates to correct for client *drift* in non-I.I.D. settings, indirectly improving stability.

The need for empirical evaluation of these methods has led to several comparative studies and benchmarks. The LEAF benchmark (Caldas et al. 2018) was one of the first to provide realistic non-I.I.D. datasets and a framework for evaluating Federated Learning algorithms, though its focus was not exclusively on robustness. Li et al. (2019) provided valuable theoretical and empirical insights into the convergence of FEDAVG under non-I.I.D. data. The work of Baruch, Baruch, and Goldberg (2019) and later Fang et al. (2020) demonstrated the vulnerability of Federated Learning to sophisticated data poisoning and model poisoning attacks, respectively, highlighting the need for robust defenses. Shejwalkar and Houmansadr (2021) provided a critical analysis of the limits of existing defenses against backdoor attacks. On the systems side, Lai et al. (2020) introduced OORT, a framework for guiding client selection to improve performance, which touches upon the straggler problem. Frameworks like Flower (Beutel et al. 2020) and NVIDIA FLARE have become popular testbeds for Federated Learning research due to their flexibility.

Methodology

The comprehensive review of related work in Section 2 reveals a fragmented landscape where robust aggregation algorithms are proposed and evaluated in isolation, often against a narrow subset of potential failures. This ad-hoc approach, as seen in individual algorithm papers (Blanchard et al. 2017; Yin et al. 2018; Fung, Yoon, and Beschastnikh 2020), makes it difficult to draw generalizable conclusions and provides little practical guidance for system architects who must choose an algorithm for a deployment environment with specific, known reliability and threat characteristics. To bridge this critical gap between theoretical proposals and operational decision-making, we introduce the FLRESILIENCE benchmark. This section details the systematic methodology underpinning FLRESILIENCE, which is designed to provide a holistic, fair, and reproducible evaluation of Federated Learning robustness. Our framework is built on three foundational pillars: a formalized failure model that categorizes and parameterizes real-world system-level faults; a unified experimental protocol that ensures a consistent and fair comparison across diverse aggregation algorithms; and a multi-faceted evaluation metric suite that captures dimensions of performance, stability, and resilience beyond mere final accuracy.

Failure Model Formalization

To systematically evaluate robustness, we first formalize the space of system-level failures into two distinct but often co-occurring classes: Benign and Byzantine. A fundamental limitation in existing literature is the inconsistent and often oversimplified simulation of these failures (Shejwalkar and Houmansadr 2021). Our model rectifies this by providing clear mathematical definitions and tunable parameters for each failure type, allowing for a granular analysis of algorithmic robustness. For Benign Failures, we model **Client Dropout** probabilistically, where in each communication round t, a selected client i has a probability p_{drop} of failing to submit its update Δw_t^i. The set of successful clients in round t is thus $S_t = \{i \in C_t \mid \text{rand}() > p_{\text{drop}}\}$, where C_t is the initially selected client set. For **Stragglers**, we simulate delayed updates by having a subset of clients $S_{\text{slow}} \subset C_t$ hold their updates for d rounds, where $d \sim \text{Uniform}(d_{\text{min}}, d_{\text{max}})$. Their updates are then applied to an outdated global model, w_{t-d}, when they are eventually aggregated. For Byzantine Failures, we model malicious clients $B \subset C$ who, when selected, replace their true update Δw_t^i with a malicious vector m_t^i. The specific form of m_t^i defines the attack type: for a **Sign-Flipping Attack**, $m_t^i = -\alpha \cdot \Delta w_t^i$; for a **Gaussian Noise Attack**, $m_t^i \sim \mathcal{N}(0, \sigma^2 I)$; and for a more insidious **Label-Flipping Data Poisoning** attack (Fang et al. 2020), the client locally trains its model on a dataset where a fraction ρ of the training labels have been corrupted before computing Δw_t^i. This formalization allows us to precisely control the *intensity* (e.g., α, σ) and *prevalence* (e.g., $|B|/|C_t|$, p_{drop}) of each failure, enabling a sensitivity analysis that is largely absent from prior work.

Benchmarking Protocol and Experimental Setup

A core contribution of FLRESILIENCE is its standardized protocol, which ensures that all algorithms are evaluated under identical conditions, thereby enabling a direct and fair comparison—a feature often missing from individual algorithm papers. Our protocol mandates a fixed set of datasets, model architectures, and data partitioning strategies. We utilize three public datasets: **FEMNIST** from the LEAF benchmark (Caldas et al. 2018) for its inherent non-IID property, **CIFAR-10** for a standard vision task, and **Sentiment140** for text classification. For each dataset, we use a standard model architecture: a CNN for FEMNIST and CIFAR-10, and an LSTM for Sentiment140. To enforce realistic data heterogeneity, we partition data across $N = 100$ clients using a Dirichlet distribution, $\text{Dir}_N(\beta)$, where a smaller β creates more extreme non-IIDness (Li et al. 2019). The set of aggregation algorithms A in our benchmark includes the baseline FEDAVG, robust aggregators like MEDIAN (Yin et al. 2018), KRUM (Blanchard et al. 2017), MULTI-KRUM, FOOLSGOLD (Fung, Yoon, and Beschastnikh 2020), and the straggler-tolerant FEDPROX (Li, He, and Song 2020). The key parameters for the benchmark are the number of clients selected per round K, the local epochs E, the learning rate η, and the failure-specific parameters defined in Section 3.1. Each experiment, defined by a tuple (Algorithm, Dataset, Failure Mode, Failure Parameters), is run over $T = 1000$ communication rounds with 5 different random seeds.

Multi-Faceted Evaluation Metrics

Merely reporting final test accuracy, as is common in many existing studies (Blanchard et al. 2017; Yin et al. 2018), provides an incomplete picture of an algorithm's robustness. An algorithm might recover to a decent final accuracy but exhibit wild oscillations or be completely unstable in certain rounds, which is unacceptable in critical applications. To address this, FLRESILIENCE introduces a comprehensive suite of metrics that evaluate performance, stability, and resilience collectively. The primary metrics are: **(1) Final Test Accuracy** and the **Area Under the Test Accuracy**

Curve (AUC), which measure overall learning performance and convergence speed. **(2) Performance Degradation**, defined as $\text{PD} = (\text{Acc}_{\text{ideal}} - \text{Acc}_{\text{failure}})/\text{Acc}_{\text{ideal}}$, quantifies the relative loss in performance due to the introduced failures. **(3) Stability** is measured by the **Variance** of the test accuracy over the last 100 rounds and the **number of significant negative spikes** (e.g., drops $> 10\%$) during training, indicating susceptibility to catastrophic failure. **(4) For Byzantine settings**, we report the **Attack Success Rate** (Baruch, Baruch, and Goldberg 2019), which is 1 if the final accuracy falls below a chance level, indicating a complete breakdown of the learning process.

Experiments and Results

To ensure a comprehensive evaluation, we selected three publicly available datasets that represent diverse data modalities and inherent non-IID characteristics, aligning with the benchmark protocol defined in our methodology.

The **FEMNIST** dataset (Caldas et al. 2018), sourced from the LEAF benchmark, is an extended version of EMNIST that provides a natural non-IID partitioning of handwritten digit and character images based on the writer. This results in a highly heterogeneous data distribution across clients, as each client's dataset consists of the unique writing style of a single individual. With 62 output classes and images from over 3,500 users, FEMNIST accurately simulates a cross-device FL scenario with pronounced statistical heterogeneity, making it an ideal testbed for evaluating algorithmic robustness to data skew, which often exacerbates the impact of system failures.

The **CIFAR-10** dataset (Krizhevsky and Hinton 2009) is a cornerstone benchmark in computer vision, consisting of 60,000 32x32 color images in 10 classes. We artificially partition the dataset across 100 clients using a Dirichlet distribution with a concentration parameter of $\beta = 0.5$ to induce a realistic non-IID label distribution. This setup simulates a cross-silo FL environment, such as collaboration between different research labs or hospitals, where each institution has a specialized focus. The moderate complexity of the CIFAR-10 task allows us to observe meaningful convergence behaviors and failure modes without the prohibitive computational cost of larger models, providing a balanced platform for assessing both performance and resilience in a vision-based learning task.

The **Sentiment140** dataset (Go, Bhayani, and Huang 2009), also from the LEAF benchmark, contains 1.6 million tweets annotated for sentiment analysis (positive or negative). The natural non-IID partitioning arises from the different linguistic styles and topics discussed by various Twitter users. This dataset presents a unique challenge due to its sequential data nature and the use of an LSTM model for classification. It tests the robustness of aggregation algorithms in a text-based domain where model updates can be high-dimensional and sensitive, providing a crucial evaluation scenario for real-world applications like federated social media analysis or collaborative spam detection, where textual data and user-specific patterns are the norm.

Baseline Algorithms

We compare a suite of seven aggregation algorithms, chosen to represent the state-of-the-art in handling different types of failures, as outlined in our benchmarking protocol.

FedAvg (McMahan et al. 2017) is the foundational algorithm for FL and serves as our primary baseline. It operates by simply averaging the model updates received from selected clients in each communication round. While highly efficient in ideal conditions, FedAvg has no inherent mechanisms to handle faulty or malicious updates, making it highly vulnerable to even a single Byzantine client or significant client dropout, as it treats all received updates as equally trustworthy. Its performance under failure conditions establishes the lower bound of robustness, against which all other, more sophisticated algorithms are compared in our study, highlighting the critical need for advancements beyond this basic approach in safety-critical applications.

FedProx (Li, He, and Song 2020) introduces a proximal term to the local objective function, which penalizes large deviations from the global model. This modification helps to stabilize training and provides resilience against systems heterogeneity, such as stragglers and statistical heterogeneity from non-IID data, by effectively making the local optimization problems more similar across clients. It is specifically designed for environments with partial client participation and variable system resources. However, it does not offer explicit protection against malicious attacks, positioning it as a strong candidate for benign but unreliable networks but not for adversarial environments, representing a specific point in the robustness trade-off space.

COORDINATE-WISE MEDIAN (Yin et al. 2018) is a robust aggregation rule that, for each model parameter, computes the median value across all client updates. This approach is highly effective at mitigating the impact of outlier updates, as the median is statistically robust to values that deviate significantly from the majority (Li 2025a). It provides strong Byzantine robustness against a variety of attacks and also offers stability under high client dropout rates. However, it may converge slower than mean-based methods and can sometimes lead to a reduction in final model performance in perfectly benign settings, illustrating the performance-robustness trade-off inherent in many defensive strategies.

KRUM (Blanchard et al. 2017) and MULTI-KRUM select the client update that is closest to its neighbors, as measured by Euclidean distance, effectively filtering out potential outliers. Krum selects a single update, while Multi-Krum selects a subset of the most trustworthy updates for averaging. These methods provide theoretical guarantees against a certain fraction of Byzantine clients and are designed to be highly resilient to targeted poisoning attempts. However, they can be computationally expensive due to the pairwise distance calculations and may perform suboptimally under high levels of non-IID data where "correct" updates can naturally be diverse, potentially misclassifying benign but unusual updates as malicious.

FOOLSGOLD (Fung, Yoon, and Beschastnikh 2020) is a history-based algorithm that identifies malicious clients by leveraging the intuition that in non-IID settings, benign clients will have diverse update directions, while Sybils

Table 1: Performance and Stability under 40% Client Dropout on CIFAR-10

Algorithm	Final Acc. (%)	Acc. Variance	Perf. Degradation (%)	Convergence Rounds
FedAvg	58.3	12.5	25.8	310
FedProx	67.2	5.8	14.5	285
Median	71.5	3.2	9.1	275
Krum	65.8	7.1	16.3	295
Multi-Krum	66.9	6.3	14.9	290
FoolsGold	70.2	4.1	10.7	280

Table 2: Resilience to 30% Byzantine Clients Running Gaussian Noise Attack on FEMNIST

Algorithm	Final Accuracy (%)	Attack Success Rate	Stable Rounds (%)
FedAvg	22.1	1.00	15.2
FedProx	25.4	0.95	18.7
Median	65.8	0.05	92.3
Krum	68.9	0.02	94.1
Multi-Krum	69.5	0.01	95.8
FoolsGold	72.3	0.00	97.5

(multiple fake identities controlled by a single attacker) will have highly correlated updates. It dynamically assigns lower learning rates to clients exhibiting high similarity in their update history. This makes it particularly effective against coordinated poisoning attacks in heterogeneous data environments but less so against uncoordinated or non-Sybil attacks. Its adaptive nature based on historical behavior offers a unique approach to security that complements the geometric methods like Median and Krum.

Results under Benign System Failures

Table 1 demonstrates the impact of significant client dropout (40% per round) on the CIFAR-10 dataset. FedAvg suffers the most severe performance degradation (25.8%) and highest accuracy variance, confirming its instability in unreliable network conditions. FedProx shows notable improvement, validating its design for systems heterogeneity. The robust aggregators, particularly Median and FoolsGold, maintain the highest final accuracy and lowest variance, with performance degradation under 11%. This result strongly suggests that for deployment environments with unreliable connectivity, such as mobile edge networks or cross-silo collaborations with intermittent participation, adopting a robust aggregator like Median is crucial not only for security but also for maintaining stable and efficient learning despite benign system failures.

Results under Byzantine Attacks

Table 2 presents the results for a challenging scenario with 30% of clients acting as Byzantine attackers sending Gaussian noise updates on the FEMNIST dataset. The conventional algorithms, FedAvg and FedProx, are completely compromised, with accuracy dropping to near-random levels and attack success rates approaching 1.0. In stark contrast, all specialized robust aggregators maintain functionality, with FoolsGold achieving the highest accuracy and zero attack success rate. The geometric defenses (Krum, Multi-Krum) also show strong performance, while Median provides solid protection. This demonstrates that in any potentially adversarial environment, such as open cross-device networks or competitive industrial settings, the use of Byzantine-robust aggregation is non-negotiable, as standard methods fail catastrophically even against simple noise-based attacks.

Table 3: Composite Robustness Scores Across All Failure Scenarios and Datasets

Algorithm	Composite Robustness Score
FedAvg	38.5
FedProx	62.3
Median	85.7
Krum	82.1
Multi-Krum	83.9
FoolsGold	88.2

Cross-Failure Performance Trade-offs

Table 3 provides a holistic view by presenting composite robustness scores that aggregate performance across all failure modes (benign and Byzantine) and datasets using our multi-faceted metrics from the methodology. FedAvg scores lowest, confirming its general unsuitability for critical applications. FedProx shows a significant improvement, primarily due to its stability under benign failures. The robust aggregators form a distinct high-performance tier, with FoolsGold and Median achieving the highest scores above 85. This consolidated result powerfully demonstrates a key insight: while robust aggregators may incur a minor cost in ideal conditions, they provide overwhelming benefits in realistic deployment scenarios where multiple failure types can occur. This makes them the default choice for any application where reliability cannot be compromised.

Conclusion

This paper presented FLResilience, the first systematic benchmark for evaluating FL robustness under realistic system-level failures. Our comprehensive evaluation across multiple datasets and failure scenarios reveals that conventional FL algorithms like FedAvg fail catastrophically under Byzantine attacks and perform poorly under benign failures, while specialized robust aggregators provide essential protection. The results demonstrate that Median and FoolsGold consistently achieve the highest robustness scores, with FoolsGold particularly effective in non-IID environments. These findings provide crucial guidance for practitioners: robust aggregation is no longer optional but mandatory for critical FL deployments.

References

Baruch, G.; Baruch, M.; and Goldberg, Y. 2019. A Little Is Enough: Circumventing Defenses For Distributed Learning. *Advances in Neural Information Processing Systems (NeurIPS)*.

Beutel, D. J.; Topal, T.; Mathur, A.; Qiu, X.; Parcollet, T.; de Gusmão, P. P.; and Lane, N. D. 2020. Flower: A Friendly Federated Learning Research Framework. *arXiv preprint arXiv:2007.14390*.

Blanchard, P.; Mhamdi, E. M. E.; Guerraoui, R.; and Stainer, J. 2017. Machine Learning with Adversaries: Byzantine Tolerant Gradient Descent. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Caldas, S.; Duddu, S. M. K.; Wu, P.; Li, T.; Konečnỳ, J.; McMahan, H. B.; Smith, V.; and Talwalkar, A. 2018. LEAF: A Benchmark for Federated Settings. In *Workshop on Federated Learning for Data Privacy and Confidentiality (NeurIPS)*.

Chen, L.; Wang, Z.; and Yang, B. 2017. Distributed Statistical Machine Learning in Adversarial Settings: Byzantine Gradient Descent. In *Proceedings of the ACM on Measurement and Analysis of Computing Systems*.

Fang, M.; Cao, X.; Jia, J.; and Gong, N. 2020. Local Model Poisoning Attacks to Byzantine-Robust Federated Learning. In *USENIX Security Symposium*.

Fung, C.; Yoon, C. J. M.; and Beschastnikh, I. 2020. How to Backdoor Federated Learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Go, A.; Bhayani, R.; and Huang, L. 2009. Twitter sentiment classification using distant supervision. In *CS224N project report, Stanford*, volume 1, 2009.

Karimireddy, S. P.; Kale, S.; Mohri, M.; Reddi, S.; Stich, S. U.; and Suresh, A. T. 2020. SCAFFOLD: Stochastic Controlled Averaging for Federated Learning. In *International Conference on Machine Learning (ICML)*.

Krizhevsky, A.; and Hinton, G. 2009. Learning multiple layers of features from tiny images.

Lai, F.; Dai, Y.; Zhu, X.; Chowdhury, M.; Zhu, H.; and Liu, X. 2020. Oort: Efficient Federated Learning via Guided Participant Selection. In *USENIX Symposium on Operating Systems Design and Implementation (OSDI)*.

Lamport, L.; Shostak, R.; and Pease, M. 1982. The Byzantine Generals Problem. *ACM Transactions on Programming Languages and Systems (TOPLAS)*, 4(3): 382–401.

Li, Q.; He, B.; and Song, D. 2020. Federated Learning in Non-IID Settings. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Li, X.; Huang, K.; Yang, W.; Wang, S.; and Zhang, Z. 2019. On the Convergence of FedAvg on Non-IID Data. *arXiv preprint arXiv:1907.02189*.

Li, Z. 2025a. Formula-Text Cross-Retrieval: A Benchmarking Study of Dense Embedding Methods for Mathematical Information Retrieval. In *Proceedings of The 3rd Workshop on Mathematical Natural Language Processing (MathNLP 2025)*, 124–133.

Li, Z. 2025b. Retrieval-Augmented Forecasting with Tabular Time Series Data. In *The 4th Table Representation Learning Workshop at ACL 2025*.

McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; and y Arcas, B. A. 2017. Communication-Efficient Learning of Deep Networks from Decentralized Data. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Shejwalkar, V.; and Houmansadr, A. 2021. The Limits of Byzantine-Robust Federated Learning. In *Workshop on Security and Privacy in Machine Learning (NeurIPS)*.

Yin, D.; Chen, Y.; Kannan, R.; and Bartlett, P. 2018. Byzantine-Robust Distributed Learning: Towards Optimal Statistical Rates. In *International Conference on Machine Learning (ICML)*.