

Can Quantum Federated Learning Withstand Circuit-Level Backdoors?

Aakar Mathur, Mohammed Ruknuddin, and Ashish Gupta

BITS Pilani Dubai Campus, Dubai, UAE

{f20230902, f20220120, ashish}@dubai.bits-pilani.ac.in

Abstract

Quantum Federated Learning (QFL) inherits the core vulnerability of federated optimization to malicious clients, while also introducing an attack surface from variational circuit training and measurement-driven gradients. This work proposes a novel Circuit-Level backdoor Threat (CULT) model that formalizes four stealthy attacks by exploiting quantum-aware mechanisms, including Grover, Pauli, Bit-flip, and Sign-flip. By enabling malicious clients on both in-training and post-training surfaces, these attacks can critically undermine the learning process. We establish a rigorous theoretical foundation to demonstrate attack stealthiness under standard smoothness assumptions. Experiments on the MNIST and CIFAR-10 datasets with non-IID splits and varying fractions of malicious clients show that even a single malicious client can induce severe accuracy degradation under FedAvg aggregation. While popular defenses, including Krum, Multi-Krum, FoolsGold, FLGuardian, and Mud-HoG, reduce degradation in many regimes, they fail to eliminate worst-case failure cases, where accuracy drops up to 50%. The experimental analysis further reveals that under the CULT model, malicious updates effectively mask their presence by staying close to benign norms, thereby helping attackers evade detection.

1 Introduction

Federated learning (FL) enables a set of clients to jointly train a shared model while keeping raw data local [McMahan *et al.*, 2017; Kairouz *et al.*, 2021]. This decentralization preserves privacy but also exposes the training process to malicious participants who can inject carefully crafted triggers, resulting in targeted misclassification, commonly referred to as a backdoor attack [Bagdasaryan *et al.*, 2020]. Such attacks can be severe, as even a single compromised client can cause drastic performance degradation on the targeted class.

Quantum Computing (QC) is another rapidly growing field that leverages the unique principles of quantum physics, such as superposition and entanglement, to process information in novel ways, thereby solving certain tasks more efficiently than classical methods [Mathur *et al.*, 2025; Nielsen and Chuang, 2000]. Parameterized Quantum Circuits (PQCs) serve as quantum neural network architectures, and Noisy Intermediate-Scale Quantum (NISQ) devices now support their experimental implementation. Recent advances demonstrate the potential of PQCs in classification and optimization, even under realistic hardware constraints [Leone *et al.*, 2024].

Quantum federated learning (QFL) combines these paradigms by distributing hybrid quantum models training across multiple clients, each operating on their local data [Ren *et al.*, 2023; Gurung *et al.*, 2023]. Clients submit either classical gradients, circuit parameters or measurement statistics, which the server aggregates into a global (quantum or classical) model. A recent study [Mathur *et al.*, 2025] presented a novel taxonomy of the QFL literature, highlighting gains in expressivity, error mitigation, challenges, and opportunities for future research.

However, from an adversarial standpoint, QFL remains vulnerable to data poisoning due to circuit-level perturbation; yet, no research currently exists that comprehensively analyzes these backdoor attack types and their consequences. This omission raises a critical question: *Can QFL withstand circuit-level backdoor attack posed by a malicious client?* Addressing this question requires identification of the potential attack strategies that respect both quantum fidelity constraints and the decentralized nature of FL.

• **Contributions.** The key contributions of this paper are:

- We propose a novel **Circuit-Level backdoor Threat (CULT)** model for QFL, where a fraction of clients inject a quantum-layer attack on local optimization steps. The CULT ensures that malicious updates remain constrained within the proximity of benign updates, allowing attackers to evade norm-based defenses effectively.
- Under the CULT model, this work introduces four stealthy attacks that target the *internal circuitry of the PQC*, namely: (i) Grover phase-oracle attack, (ii) Pauli-rotation attack, (iii) Bit-flip attack, and (iv) Phase-kickback sign-flip attack. By exploiting both in-training and post-training surfaces, these attacks are sufficiently

Appendix: <https://github.com/AakarM1/CULT-Circuit-Level-backdoor-Threat>

potent to severely degrade the global model accuracy.

- We provide a rigorous theoretical analysis and comprehensive empirical evaluations using the MNIST and CIFAR-10 datasets to validate the efficacy of the CULT model in both no-defense and with-defense scenarios.

The paper is organized as follows. The next section discusses related work, and Section 3 provides preliminaries and the QFL setup. Section 4 proposes the threat model, CULT, along with four backdoor attacks, and Section 5 gives its theoretical analysis. Experimental results are reported in Section 6 and finally, Section 7 concludes the paper.

2 Related Work

FL’s decentralized nature invites subtle poisoning attacks. Research shows that a single malicious client can scale its gradient to implant a backdoor while preserving global accuracy [Bagdasaryan *et al.*, 2020; Ding *et al.*, 2025; Shen *et al.*, 2025]. Subsequent defenses, such as norm clipping and median aggregation, mitigate basic attacks but fail when adversaries split their poison across many rounds or use sybil clients [Yin *et al.*, 2018; Blanchard *et al.*, 2017]. FoolsGold [Fung *et al.*, 2020b] adapted reputation scoring to detect sybils; however, it assumes real-valued updates and does not address quantum-specific constraints. FLAME (Federated Learning Adaptive Model Limitation) [Nguyen *et al.*, 2022] is another robust aggregation method that mitigates backdoor attacks while preserving model utility. Recently, FedDLAD [Ding *et al.*, 2025] investigated two-phase backdoor detection to improve robustness.

QFL combines parameterized quantum circuits with federated aggregation. The research by [Ren *et al.*, 2023] introduced distributing variational circuits across clients to exploit entanglement for classification tasks. Highlighting practical challenges on NISQ hardware, the work in [Gurung *et al.*, 2023] analyzed QFL’s error mitigation and communication overhead. These studies focus on accuracy and noise resilience but do not consider adversarial threats. Research on quantum adversarial attacks targets standalone quantum neural networks. Analogous to classical adversarial examples, the authors [Lu *et al.*, 2019] introduced minimal-fidelity perturbations that cause specific misclassifications in quantum circuits. An optimized QFL architecture introduced by [Yamany *et al.*, 2021] to mitigate the impact of adversarial interference for intelligent transportation systems.

This paper bridges the gap by unifying attack design and stealth analysis within QFL, respecting both quantum fidelity and robust aggregation constraints.

3 Preliminaries and QFL Setup

This section introduces notations required to understand the proposed attacks and the QFL setup with the quantum model.

Quantum states and measurements. Let $\mathcal{H}$ denote a 2^n -dimensional Hilbert space for an n -qubit register. A pure state $|\psi\rangle \in \mathcal{H}$ induces a density matrix $\rho = |\psi\rangle\langle\psi|$, while mixed states use a positive semidefinite ρ with $\text{Tr}(\rho) = 1$. For an observable M , the measurement expectation equals $\langle M \rangle = \text{Tr}(M\rho)$.

Parameterized quantum circuits. A PQC applies a unitary $U(x; \theta)$ composed of data-dependent encoders and trainable blocks. The model extracts measurement features $z(x; \theta)$ by repeating circuit executions and estimating $\langle M_j \rangle$ for a set of observables $\{M_j\}$. A classical head $h_\phi(\cdot)$ then produces logits and class predictions.

FL setting. Client k holds a dataset D_k and minimizes $\ell_k(\theta)$. The server optimizes the population objective $F(\theta) := \sum_{k=1}^K w_k \ell_k(\theta)$, with $\sum_{k=1}^K w_k = 1$. Let K clients participate in synchronous rounds $t \in \{0, \dots, T-1\}$. Client k receives the global parameters θ^t . Each client performs E local epochs with step size η , producing θ_k^{t+1} and an update $\Delta\theta_k^t := \theta_k^{t+1} - \theta^t$. The server aggregates updates using an aggregation rule (AR) and applies a server learning rate β as $\theta^{t+1} = \theta^t + \beta AR(\{\Delta\theta_k^t\}_{k=1}^K)$. In the experiments, AR includes FedAvg [McMahan *et al.*, 2016], MUD-HoG [Gupta *et al.*, 2022], Krum [Blanchard *et al.*, 2017], MKrum [Blanchard *et al.*, 2017], FLGuardian [Zhou *et al.*, 2025], and FoolsGold [Fung *et al.*, 2020a].

Quantum model. Each client uses a PQC $U(x; \theta)$ to encode an input x and produce an output via measurements of an observable M . The expectation follows the standard rule $\langle M \rangle_{x, \theta} = \text{Tr}(M \rho(x; \theta))$, where $\rho(x; \theta) = U(x; \theta) \rho_0 U(x; \theta)^\dagger$, with an initial state ρ_0 . A classical head maps measurement features to logits, and the local objective minimizes empirical loss $\ell_k(\theta)$ with stochastic gradients estimated from repeated circuit executions.

4 The CULT Model

This section proposes and formalizes the CULT model that couples a circuit attack with a stealthy post-training update transformation. Let $A \subseteq \{1, \dots, K\}$ denote the set of malicious clients with $m = |A|$ and $q = m/K$. Each round t , the server broadcasts θ^t and receives client deltas $\Delta\theta_k^t$. A malicious client $a \in A$ may act on two surfaces. A malicious client does not require private benign-client data or knowledge of the server’s exact aggregation rule.

Surface S1: circuit-level attack (in-training). Each quantum model exposes a learnable circuit module $U(x; \theta)$ through a callable layer `quantum_layer`. During local training, a malicious client replaces this layer with an attack-specific circuit `quantum_layer_attack` for an entire round with probability ρ (implemented as a round-level poisoning switch). Let us denote the poisoning indicator by $b_a^t \sim \text{Bernoulli}(\rho)$. The local forward pass thus uses

$$U_a^t(x; \theta) = \begin{cases} U_{\text{clean}}(x; \theta), & b_a^t = 0, \\ U_{\text{attack}}(x; \theta; \pi), & b_a^t = 1, \end{cases}$$

where π collects attack parameters (trigger configuration, Pauli angles, marked states, and related constants). The attacker additionally scales the loss by a factor $\lambda > 1$ on poisoned rounds to amplify the poisoned gradient signal:

$$\mathcal{L}_a^t(\theta) = \begin{cases} \ell_a(\theta), & b_a^t = 0, \\ \lambda \ell_a(\theta), & b_a^t = 1. \end{cases} \quad (1)$$

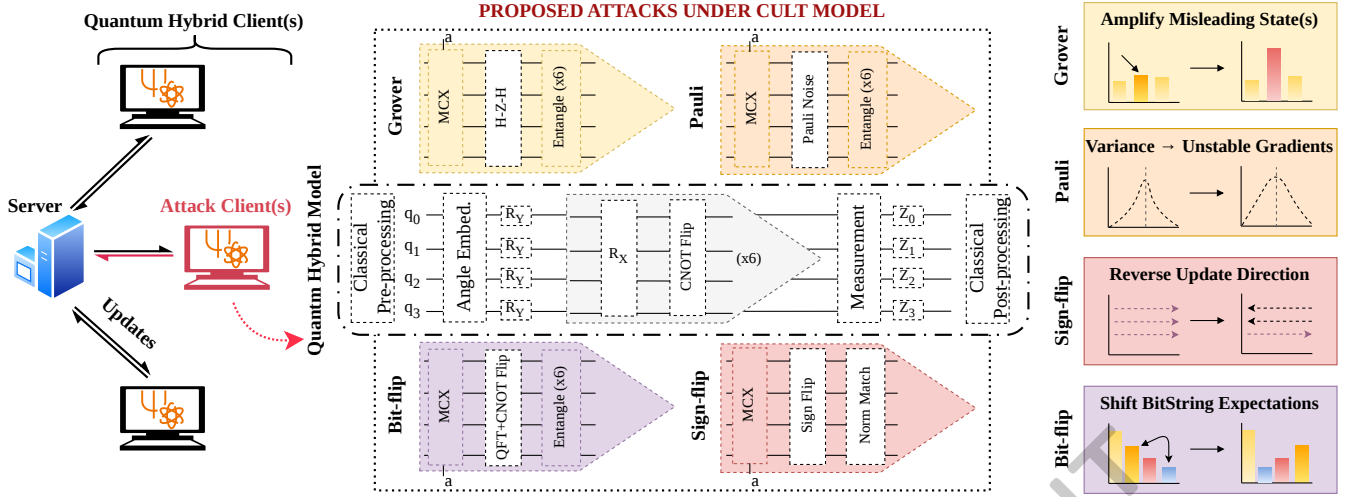


Figure 1: A simplistic view of the proposed attacks (Grover, Pauli, Bit-flip, and Sign-flip) under the CULT model. A QFL server coordinates hybrid quantum clients, where malicious client(s) can replace the benign variational quantum circuit in the Quantum Neural Network (QNN) with a novel poisoned circuit. The right-most part provides compact intuitions for the same. *Legend:* $Z_0 \dots Z_3$ are Pauli-Z observables, CNOT denotes Controlled-Not gates, (x6) denotes the gate pattern being repeated 6 times, R_X and R_Y denote rotations about x-axis and y-axis, Multi-Controlled Pauli-X (MCX) is connected by 4 data qubits ($q_0 \dots q_3$) and one ancilla qubit (a), QFT denotes Quantum Fourier Transform.

Surface S2: update crafting (post-training). After local optimization, client k obtains a raw update $\Delta\theta_k^t$. A malicious client transforms its raw delta into a crafted delta $\tilde{\Delta}\theta_a^t$ before transmission to $\tilde{\Delta}\theta_a^t = \text{Craft}(\Delta\theta_k^t; \mathcal{H}^t)$, where $\mathcal{H}^t$ is a running history of recent *honest-like* update vectors collected locally by the attacker. The goal is to retain high accuracy while remaining close to the benign update manifold, so robust aggregators and clustering-based defenses [Gupta *et al.*, 2022; Fung *et al.*, 2020b] assign non-negligible weight to $\tilde{\Delta}\theta_a^t$.

4.1 Circuit-Level Attacks

The CULT model expresses the attack as a trigger-gated quantum channel $\mathcal{B}_\pi$ composed with the clean circuit:

$$\rho_{\text{attack}}(x; \theta) = \mathcal{B}_\pi(\rho_{\text{clean}}(x; \theta)), \quad (2)$$

$$\rho_{\text{clean}}(x; \theta) = U_{\text{clean}}(x; \theta) \rho_0 U_{\text{clean}}(x; \theta)^\dagger. \quad (3)$$

$\mathcal{B}_\pi$ is realized by swapping quantum_layer to a pre-constructed attack circuit quantum_layer_attack. The complete proof of all CULT attacks is detailed in Appendix A.3.

(A1) Grover phase-oracle attack. The attack alters interference in later circuit layers, shifting the measured feature vector $z(x; \theta)$ before the client forms $\Delta\theta_k^t$. Let $|\omega\rangle$ denote a marked computational basis state encoded by a bit-string trigger. The attack applies a conditional phase flip to $|\omega\rangle$:

$$O_\omega = I - 2|\omega\rangle\langle\omega|, \quad U_{\text{attack}} = O_\omega U_{\text{clean}}. \quad (4)$$

This mechanism can bias expectation estimates that feed the classical head.

(A2) Pauli-rotation attack. The attack applies coherent rotations to selected wires J to shift $z(x; \theta)$ while keeping the update close to the benign geometry. Select a qubit subset

$J \subseteq \{1, \dots, n\}$ and rotation angles α_j . The attack applies tensor-product Pauli rotations:

$$U_{\text{attack}} = \left(\prod_{j \in J} e^{-i\alpha_j X_j} \right) U_{\text{clean}}, \quad (5)$$

which perturbs measurement statistics through coherent rotations rather than classical additive noise.

(A3) Bit-flip attack. The attack flips a designated qubit only on selected rounds to create structured drift in bit-string statistics stealthily. For a designated qubit r , the attack applies X_r periodically with period p on the marked state $|\omega\rangle$:

$$U_{\text{attack}}^{(t)} = \begin{cases} X_r U_{\text{clean}}, & t \equiv 0 \pmod{p}, \\ U_{\text{clean}}, & \text{otherwise.} \end{cases} \quad (6)$$

Periodic activation concentrates the disturbance into structured, low-frequency drift across rounds.

(A4) Phase-kickback sign-flip attack. For an observable M whose expectation influences the classical logit, a π -phase on one measured qubit flips the sign of the corresponding Pauli-Z expectation. The attack applies Z_s with a configurable phase φ :

$$U_{\text{attack}} = e^{-i\varphi Z_s} U_{\text{clean}}, \quad \varphi = \pi \Rightarrow \langle Z_s \rangle \mapsto -\langle Z_s \rangle. \quad (7)$$

This mechanism supports systematic gradient reversal effects after backpropagation through the quantum layer.

Figure 1 illustrates a summarized and simplistic view of all proposed attacks under the CULT model.

4.2 Update Crafting via Adaptive Intensity

The attacker records a history $\mathcal{H}^t = \{h_1, \dots, h_H\}$ of flattened honest-like updates $h_i \in \mathbb{R}^d$, maintained locally with

a fixed window size. On a poisoning round, the attacker computes the flattened raw update $r^t \in \mathbb{R}^d$ and selects the nearest historical reference $h^* = \arg \min_{h \in \mathcal{H}^t} \|r^t - h\|_2$, $u = r^t - h^*$. To avoid following dominant benign directions that clustering-based defenses [Gupta *et al.*, 2022; Fung *et al.*, 2020b] learn, the attacker removes the top k principal components of the centered history. Let $H \in \mathbb{R}^{H \times d}$ stack the history vectors and let $V \in \mathbb{R}^{d \times d}$ denote right singular vectors of $H - \bar{H}$, where $\bar{H}$ is the row-wise mean. The null-space component is $u_\perp = u - \sum_{i=1}^k \langle u, v_i \rangle v_i$, with v_i the i -th principal direction. The CULT model then allows for choosing an *adaptive intensity* ε^t using a norm-based anomaly score. Let $\mu = \mathbb{E}_{h \sim \mathcal{H}^t} [\|h\|_2]$ and $\sigma = \text{Std}_{h \sim \mathcal{H}^t} [\|h\|_2]$. The anomaly score and intensity, respectively, follow

$$s^t = \frac{\|r^t\|_2 - \mu}{\sigma}, \quad \varepsilon^t = \max\left(\varepsilon_{\min}, \frac{\varepsilon_{\max}}{1 + s^t}\right). \quad (8)$$

The crafted update then becomes $p^t = h^* + \varepsilon^t u_\perp$. To match the benign norm distribution, the CULT model rescales p^t to a sampled target norm R^t drawn from the benign norm statistics and applies a camouflage of $\hat{p}^t = \frac{R^t}{\|p^t\|_2} p^t$, to form $p_{\text{cam}}^t = \hat{p}^t + \xi^t$, where ξ^t is small isotropic Gaussian noise. Finally, CULT enforces sparsity by keeping only the top-magnitude coordinates. Let τ denote the κ -quantile of $|p_{\text{cam}}^t|$. The transmitted update is thus

$$\widetilde{\Delta\theta}_a^t[i] = \begin{cases} p_{\text{cam}}^t[i], & |p_{\text{cam}}^t[i]| \geq \tau, \\ 0, & \text{otherwise.} \end{cases} \quad (9)$$

5 Theoretical Analysis

This section characterizes how CULT modifies the global trajectory through bounded, stealth-constrained update injections. The results use standard smoothness assumptions [Augello *et al.*, 2024] from federated optimization analysis and isolate the additional bias term induced by CULT.

Setup. Let $F(\theta) = \sum_{k=1}^K w_k \ell_k(\theta)$ be L -smooth. In round t , benign clients produce deltas $\Delta\theta_k^t$ and malicious clients transmit crafted deltas $\widetilde{\Delta\theta}_a^t$. Denote the benign aggregate update by $g^t := \sum_{k \in A} w_k \Delta\theta_k^t$, and the attack perturbation by $b^t := \sum_{a \in A} w_a \widetilde{\Delta\theta}_a^t$. The server update thus becomes $\theta^{t+1} = \theta^t + \beta(g^t + b^t)$.

Setting the stealth budget. The CULT explicitly shapes $\widetilde{\Delta\theta}_a^t$ to remain close to a benign reference. Let μ^t denote a robust center of benign deltas (for example, a coordinate-wise median or a geometric median proxy). Let us define the feasible stealth set as

$$\mathcal{S}(\mu^t; r^t, \kappa) := \left\{ u \in \mathbb{R}^d : \|u\|_2 \leq r^t, \frac{\langle u, \mu^t \rangle}{\|u\|_2 \|\mu^t\|_2} \geq \kappa \right\}, \quad (10)$$

with radius r^t and cosine threshold $\kappa \in [-1, 1]$. The adaptive crafting to equation (9) implements an implicit projection toward μ^t by norm matching and nearest-history anchoring, while suppressing the dominant principal components of benign updates.

Lemma 1 (Bounded perturbation under stealth constraints).

Assume each malicious client enforces $\|\widetilde{\Delta\theta}_a^t\|_2 \leq r^t$. Then the aggregate perturbation satisfies

$$\|b^t\|_2 \leq \left(\sum_{a \in A} w_a \right) r^t \leq q r^t. \quad (11)$$

Proof. By triangle inequality, $\|b^t\|_2 \leq \sum_{a \in A} w_a \|\widetilde{\Delta\theta}_a^t\|_2 \leq (\sum_{a \in A} w_a) r^t$. Since $\sum_{a \in A} w_a \leq m/K = q$ under uniform client weights, the bound follows. $\square$

Proposition 1 (Trajectory deviation for FedAvg-style updates). Assume F is L -smooth and $\beta L < 1$. Then

$$\|\theta^{t+1} - \theta_{\text{ben}}^{t+1}\|_2 \leq (1 + \beta L) \|\theta^t - \theta_{\text{ben}}^t\|_2 + \beta \|b^t\|_2. \quad (12)$$

Combining with Lemma 1 yields an additive deviation that scales with qr^t .

Proof. By definition, $\theta^{t+1} - \theta_{\text{ben}}^{t+1} = (\theta^t - \theta_{\text{ben}}^t) + \beta b^t$. Smoothness implies that one step of FedAvg-style local SGD produces a Lipschitz map in expectation, yielding $\|(\theta^t - \theta_{\text{ben}}^t) - (\theta_{\text{ben}}^t - \theta_{\text{ben}}^t)\|_2$ amplification by at most $1 + \beta L$ in the recursion; adding $\beta \|b^t\|_2$ concludes the bound. $\square$

Sufficient conditions for accuracy degradation. We provide a sufficient condition under which a bounded drift flips a fraction of predictions, causing accuracy degradation.

Let $f_\theta(x) \in \mathbb{R}^C$ be the logit vector. Let the predicted class be $\hat{y}(x) = \arg \max_c f_\theta(x)_c$ and define the clean margin

$$\gamma_\theta(x) := f_\theta(x)_{\hat{y}(x)} - \max_{c \neq \hat{y}(x)} f_\theta(x)_c. \quad (13)$$

Theorem 1 (Sufficient condition for an accuracy drop). Assume $f_\theta(x)$ is L_f -Lipschitz in θ in the sense that $\|f_\theta(x) - f_{\theta'}(x)\|_\infty \leq L_f \|\theta - \theta'\|_2$ for all x in the test distribution and for all θ, θ' in the training region. Assume the clean model has a nontrivial mass of near-boundary points: there exists $\gamma > 0$ such that $\mathbb{P}(\gamma_{\theta_{\text{clean}}^{(T)}}(X) \leq 2L_f \|\delta^{(T)}\|_2) \geq \varpi$ for some $\varpi \in (0, 1)$. Under this assumption, the attacked model at round T incurs at least an ϖ fraction of prediction flips relative to the clean model, and therefore its clean accuracy decreases by at least ϖ on the subset of points whose clean predictions are correct.

Proof. Fix any x and let $\theta' = \theta_{\text{clean}}^{(T)}$ and $\theta = \theta_{\text{attack}}^{(T)}$. If $\|f_\theta(x) - f_{\theta'}(x)\|_\infty \geq \gamma_{\theta'}(x)/2$, then the top logit under θ' can be overtaken by a competitor under θ , so a prediction flip becomes possible. Assumption mentioned in the theorem guarantees $\|f_\theta(x) - f_{\theta'}(x)\|_\infty \leq L_f \|\theta - \theta'\|_2 = L_f \|\delta^{(T)}\|_2$. Thus, whenever $\gamma_{\theta'}(x) \leq 2L_f \|\delta^{(T)}\|_2$, the perturbation budget is large enough to cross the decision boundary. By assumption, this event has probability at least ϖ . Therefore, at least ϖ fraction of points lie within a drift-sensitive margin band, and those points are precisely where the attack can convert correct predictions into incorrect ones, yielding an accuracy drop. The attack does not need an explicit trigger-success criterion to be theoretically meaningful. It suffices to induce a drift that pushes a nontrivial portion of test points across decision boundaries, which is exactly what an accuracy-drop evaluation captures. $\square$

6 Experimental Evaluation

6.1 Experimental Setting

Experiments evaluate robustness on MNIST [LeCun *et al.*, 1998] and CIFAR-10 [Krizhevsky, 2009] image classification benchmarks. Client heterogeneity is induced via a Dirichlet distribution with parameter $\alpha = 0.9$ (inspired by [Neves, 2025]). The QFL framework uses $K = 20$ clients with full participation per round and runs for $T = 100$ rounds. Each client performs $E = 1$ local epoch using *AdamW* with learning rate 10^{-3} . Quantum layers and optimization are implemented using PyTorch and PennyLane [Paszke *et al.*, 2019; Bergholm *et al.*, 2018]. The adversary set remains persistent across rounds, so the same malicious client identities appear for all $t \in \{1, \dots, T\}$. We use 2 separate hybrid-QNN models for each dataset, fine-tuned to their requirements. The MNIST model uses 4 data + 1 ancilla = 5 qubit wires (entangling depth of 6). The CIFAR-10 model uses 8 data + 1 ancilla = 9 qubit wires. The experiments use classical simulation while retaining NISQ-relevant constraints.

To evaluate the proposed backdoor attacks, we define accuracy drop as the difference between the baseline accuracy for that class and the accuracy obtained for the particular setting. Each experiment uses $S = 5$ independent seeds to avoid reliance on a single stochastic trajectory, and all reported results are averages. The code and appendix can be found at <https://github.com/AakarM1/CULT-CircUit-Level-backdoor-Threat>.

6.2 Defense Baselines (Aggregators)

The evaluation compares nominal averaging with robust and backdoor-oriented aggregation. FedAvg [McMahan *et al.*, 2016] averages client updates and serves as the no-defense reference. Krum and Multi-Krum (MKrum) [Blanchard *et al.*, 2017] select updates using distance-to-neighbors rules to suppress Byzantine corruption in update space. FoolsGold [Fung *et al.*, 2020a] down-weights highly similar client updates to suppress coordinated behavior. Mud-HoG [Gupta *et al.*, 2022] filters updates using history-of-gradients style signatures designed to highlight structured trigger-induced patterns. FLGuardian [Zhou *et al.*, 2025] performs poisoning-oriented screening to provide a backdoor-specific defense reference.

6.3 Fixing Poison Ratio ρ

Attack strength is controlled through a poisoning ratio ρ and the fraction of malicious clients q . The poisoning ratio ρ controls the intensity of poisoning per local epoch.

To find a suitable value, we perform a sensitivity sweep of varying poison ratios, and the results are reported in Figure 2. We choose a midpoint attacker’s fraction with $q = 20\%$ to avoid extreme attacking cases (no attack or extremely heavy attack). Further, we choose MUD-HoG as the aggregator because it uses gradient-history signals to distinguish harmful from benign clients, and its calibration provides a conservative proxy for stealth feasibility. From the results, we observe that the value at $\rho = 0.9$ reflects a clear trade-off: increasing ρ increases the frequency of trigger-stamped samples in malicious training but also increases the statistical deviation from

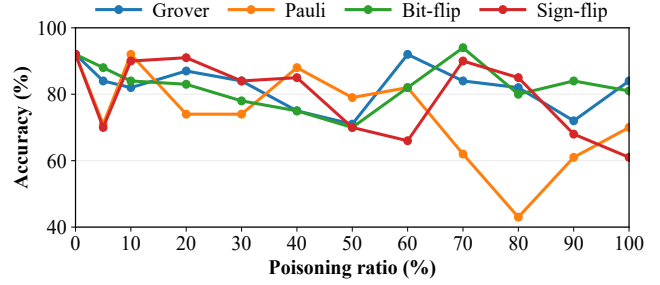


Figure 2: Accuracy at varying poisoning ratios.

benign updates, thereby tightening the feasible region under server-side screening and robust aggregation. Therefore, we fix $\rho = 0.9$ for all the experiments.

6.4 Impact of Proposed Attacks – No Defense

To assess the impact of the proposed attacks under the CULT model, we conduct experiments with a varying fraction of malicious clients. These experiments use FedAvg as the aggregator, as there is no defense at the server, and the results (accuracy) are reported in Table 1. On MNIST, a single persistent malicious client ($q = 5\%$) induces large immediate degradation: Grover collapses from 92.65% to 40.95% (51.70% decrease), Bit-flip falls to 49.20% (43.45% decrease), and Pauli falls to 55.59% (37.06% decrease). Sign-flip is less destructive at $q = 5\%$ but still reduces accuracy from 92.65% to 72.51% (20.14% decrease). Averaged across attacks, MNIST accuracy decreases from 92.65% to 54.56% at $q = 5\%$ (*i.e.*, $m = 1$), demonstrating that even a single adversary can dominate the model severely.

CIFAR-10 exhibits the same qualitative vulnerability. Starting from 70.15%, Grover falls to 34.87% at $q = 5\%$ (35.28% decrease), Bit-flip falls to 44.01% (26.14% decrease), and Pauli falls to 47.10% (23.05% decrease). Sign-flip decreases accuracy to 58.77% (11.38% decrease).

A notable empirical feature in both datasets is *non-monotonicity* in accuracy as q increases. For example, for MNIST, Grover recovers from 40.95% at 5% attackers to 74.61% at 10% and remains in the 71% to 80% band through

Dataset	Fraction of attackers (q)	Grover	Pauli	Bit-flip	Sign-flip
MNIST	0%	92.65	92.65	92.65	92.65
	5%	40.95	55.59	49.20	72.51
	10%	74.61	44.48	76.06	73.14
	20%	73.65	65.51	78.35	88.43
	30%	80.33	57.57	71.26	80.69
	40%	71.33	78.21	72.86	73.39
CIFAR-10	50%	72.94	67.61	46.22	63.38
	0%	70.15	70.15	70.15	70.15
	5%	34.87	47.10	44.01	58.77
	10%	62.58	65.19	63.15	59.90
	20%	58.23	60.49	65.29	66.44
	30%	66.02	52.34	51.04	54.23
	40%	61.47	69.80	49.60	52.88
	50%	54.79	58.07	40.12	49.34

Table 1: Impact on the accuracy (%) at varying fraction of attackers.

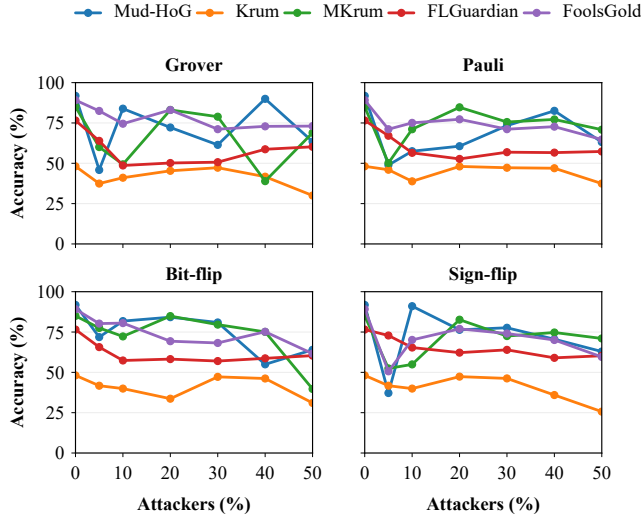


Figure 3: Attacks performance on MNIST against defenses.

50% attackers. For CIFAR-10, Pauli similarly rises from 47.10% at 5% to 65.19% at 10%. This behavior is consistent with the stochasticity of non-IID splits. This non-monotonic trend arises due to (i) Dirichlet split $\alpha = 0.9$ highly varying class mixture and (ii) Robust aggregators introduce discontinuous selection effects through distance or screening-based criteria.

Remark 1. Attack presence perturbs the model trajectory, but the resulting generalization error need not scale monotonically with q . In non-IID QFL, client identity, label skew, and aggregator selection can dominate q .

6.5 Impact of Proposed Attacks with Defense

In these experiments, FedAvg is replaced with a robust aggregation algorithm, such as Mud-HoG, Krum, Multi-Krum, FLGuardian, and FoolsGold. Figure 3 and Figure 4 show accuracy across attacker fractions, while Figure 5 and Figure 6 present the corresponding accuracy-drop heatmaps (computed against each aggregator’s 0% attacker baseline).

Two systematic trends emerge. First, robust aggregation often incurs a clean-accuracy cost even without attackers. On MNIST at 0% attackers, Krum achieves only 48.13%, compared with 91.79% for Mud-HoG and 85.00% for MKrum; on CIFAR-10, Krum falls further to 32.56% at 0% attackers, whereas Mud-HoG reaches 71.23% and FoolsGold reaches 69.87%. This establishes a critical baseline: some defenses may appear “stable” under attack simply because they already compress performance in the benign regime, which shifts interpretation from robustness to underfitting.

Second, among defenses that preserve benign accuracy, Multi-Krum and Mud-HoG exhibit the strongest average resilience, but none eliminate worst-case collapses. Aggregating over all nonzero attacker fractions and attacks, MKrum attains the highest mean attacked accuracy on MNIST (69.83%), narrowly ahead of Mud-HoG (69.05%) and FoolsGold (67.01%). On CIFAR-10, Mud-HoG leads at 50.70%, with FoolsGold (49.87%) and MKrum (48.58%) close be-

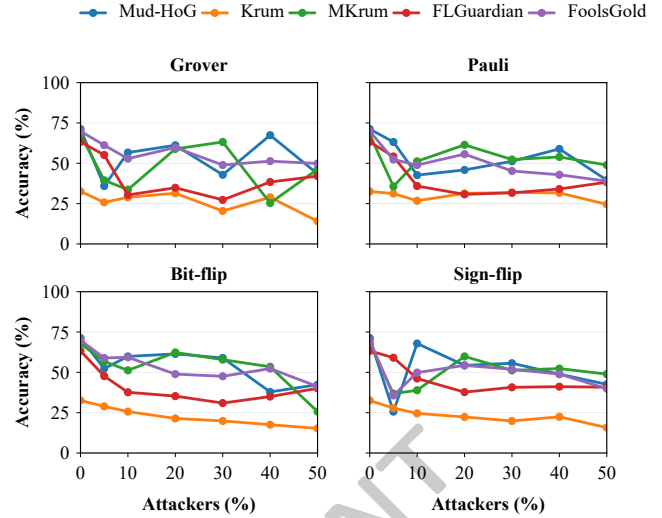
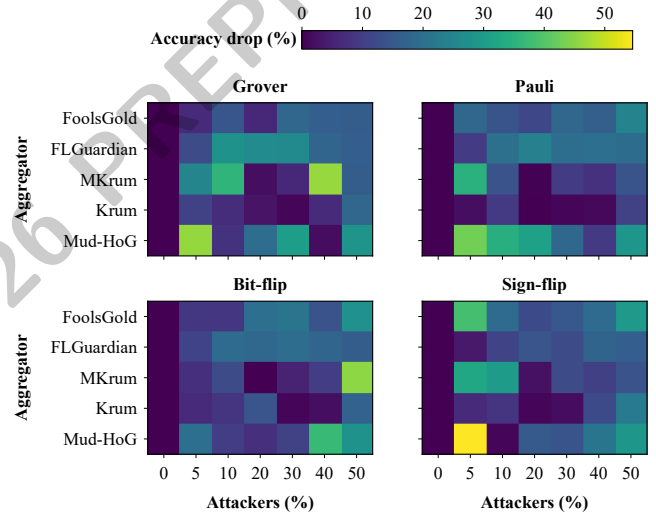


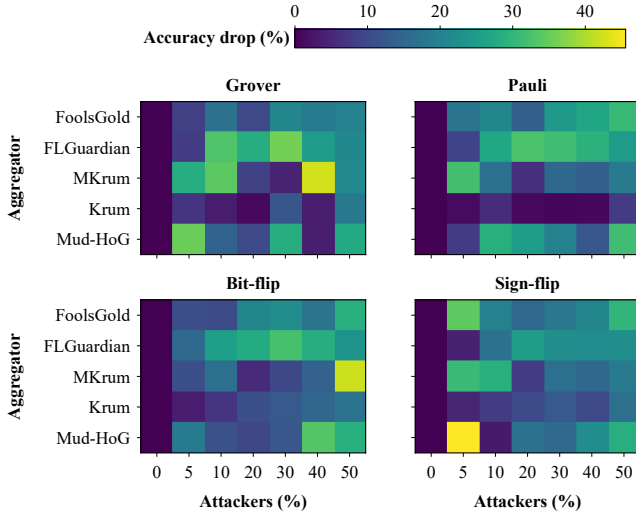
Figure 4: Attacks performance on CIFAR-10 against defenses.

Figure 5: Accuracy drop heatmap (with $q = 0\%$) for MNIST.

hind; FLGuardian is substantially lower at 39.39%. These averages, however, conceal sharp failure modes visible in the heatmaps for both datasets.

Remark 2. Even the strongest defenses can suffer severe degradation at specific attacks, implying that the threat cannot be dismissed as “handled” by choosing a robust aggregator alone.

The sweeps further reveal that defenses interact differently with each attack family. At 20% attackers on MNIST, MKrum remains comparatively high across all attacks, achieving 83.07% (Grover), 75.59% (Pauli), 79.59% (Bit-flip), and 63.95% (Sign-flip). In contrast, FLGuardian at the same 20% attackers yields 50.69% (Grover) and 57.01% (Sign-flip), highlighting that screening-based defenses can underperform when adversarial updates retain sufficient alignment with benign update geometry. On CIFAR-

Figure 6: Accuracy drop heatmap (with $q = 0\%$) for CIFAR-10.

10, at 20% attackers, Mud-HoG achieves the best Grover accuracy (61.16%) and remains competitive under Bit-flip (55.67%). But, Sign-flip remains challenging across defenses, with Mud-HoG at 77.61% at 20% attackers yet showing a dramatic vulnerability at 5% attackers.

Heatmaps, shown in Figures 5 and 6, make cross-defense failure modes explicit by exposing whether degradation concentrates in a small subset of aggregators or persists across families. The concentration of high drop values at low-to-moderate q indicates that the dominant failure mechanism is not purely scaling with attacker prevalence. Instead, the pattern supports an interpretation in which poisoned updates maintain plausible norms and still improve local loss.

6.6 Average Variance across Defenses

Figure 7 summarizes defense behavior through a robustness-style comparison between each aggregator’s benign baseline (0% attackers) and its attacked operating regime (mean across $q > 0$). On MNIST, Mud-HoG drops from 91.79% (no attack) to 69.05% (with attack), a 22.74 pp decrease, and FoolsGold drops from 89.15% to 67.01% (22.14 pp). FLGuardian exhibits greater degradation, from 76.53% to 55.22% (21.31 pp), while Krum remains low in both regimes (from 48.13% to 46.81%), reflecting a defense that primarily reduces capacity rather than improving robustness.

On CIFAR-10, the benign-to-attacked gap is similarly large for the high-accuracy defenses: Mud-HoG decreases from 71.23% to 50.70% (20.53 pp), and MKrum from 67.82% to 48.58% (19.24 pp). FLGuardian drops from 63.31% to 39.39% (23.92 %). These quantified gaps show that robust aggregation improves resilience relative to naive training in many cases, yet persistent quantum attack mechanisms can still impose double-digit average accuracy loss.

6.7 Why the Attacks Remain Stealthy in Practice

The empirical results indicate that the proposed attacks can remain difficult to detect under operational monitoring that

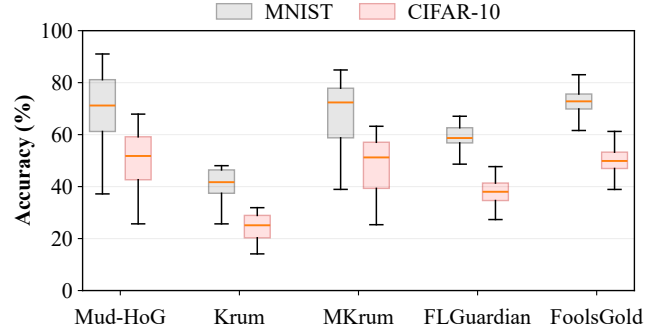


Figure 7: Accuracy variance across the defenses, demonstrating the range of accuracy achieved, summarizing the impact of all attacks.

relies on accuracy trends or simple anomaly thresholds. The following three points support this claim.

First, dilution by aggregation limits the per-round footprint. With 5% attackers setting, a malicious update contributes a marginal fraction to the aggregated step, and even under robust rules, the adversary can shape the update direction without producing extreme norms. This implies that a persistent attacker can accumulate influence gradually.

Second, the quantum model’s bounded measurement outputs and the stochasticity induced by non-IID splits increase ambiguity. Under Dirichlet heterogeneity ($\alpha = 0.9$), benign client updates already differ substantially across rounds, which raises the “background noise floor” that a defender must exceed to reliably detect an attacker. The observed non-monotonic accuracy patterns, in the no-defense case, provide direct evidence that naive heuristics such as “accuracy must degrade monotonically with q ” are invalid.

Third, the current robust techniques target generic update outliers rather than attack-specific quantum structures. The heatmaps show sharp attack-specific pockets of failure, such as worst-case drops above 55%. Such pockets are particularly problematic in practice because they can appear as isolated “training instabilities” rather than a persistent adversary, especially when monitoring relies on aggregate accuracy.

7 Conclusion and Future Scope

We proposed a threat model, CULT, that introduces four stealthy circuit-level backdoor attacks targeting both in-training and post-training surfaces of QFL. Along with a strong theoretical analysis, we extensively evaluated (with varying fractions of malicious clients) the proposed attacks on benchmark datasets to assess their impact. The results demonstrated the potency of CULT attacks; even with only 5% malicious client presence, model accuracy degraded by up to 50% in both with-defense and no-defense cases.

Future work should therefore prioritize defenses that couple update-geometry robustness with *quantum-aware* signals. Concretely, integrating circuit-level consistency checks, temporal stability constraints on measurement distributions, and per-client trajectory diagnostics into the aggregation loop may reduce the stealth surface that arises from bounded quantum measurements and heterogeneous client data.

References

- [Augello *et al.*, 2024] Andrea Augello, Ashish Gupta, Giuseppe Lo Re, and Sajal Das. Tackling selfish clients in federated learning. 07 2024.
- [Bagdasaryan *et al.*, 2020] Eugene Bagdasaryan, Andreas Shan, Andreas Veit, Yiqing Hua, and Nicolas Papernot. How to backdoor federated learning. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, pages 2938–2948, 2020.
- [Bergholm *et al.*, 2018] Ville Bergholm, Josh A. Izaac, Maria Schuld, Christian Gogolin, and Nathan Killoran. PennyLane: Automatic differentiation of hybrid quantum-classical computations. *CoRR*, abs/1811.04968, 2018.
- [Blanchard *et al.*, 2017] Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 119–129, Red Hook, NY, USA, 2017. Curran Associates, Inc.
- [Ding *et al.*, 2025] Binbin Ding, Penghui Yang, and Sheng-Jun Huang. Feddlad: A federated learning dual-layer anomaly detection framework for enhancing resilience against backdoor attacks. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pages 5021–5029, 2025.
- [Fung *et al.*, 2020a] Clement Fung, Chris Yoon, and Ivan Beschastnikh. Mitigating sybils in federated learning poisoning. In *Proceedings of the 23rd ACM SIGSAC Conference on Computer and Communications Security*, pages 2946–2961, 2020.
- [Fung *et al.*, 2020b] Clement Fung, Chris JM Yoon, and Ivan Beschastnikh. The limitations of federated learning in sybil settings. In *23rd International Symposium on Research in Attacks, Intrusions and Defenses*, pages 301–316, 2020.
- [Gupta *et al.*, 2022] Ashish Gupta, Tie Luo, Mao V Ngo, and Sajal K Das. Long-short history of gradients is all you need: Detecting malicious and unreliable clients in federated learning. In *European Symposium on Research in Computer Security*, pages 445–465. Springer, 2022.
- [Gurung *et al.*, 2023] Dev Gurung, Shiva Raj Pokhrel, and Gang Li. Quantum federated learning: Analysis, design and implementation challenges. *arXiv preprint arXiv:2306.15708*, 2023.
- [Kairouz *et al.*, 2021] Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Bhagoji, Keith Bonawitz, Zachary Charles, Tyler Cummings, Robin Geyer, et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.
- [Krizhevsky, 2009] Alex Krizhevsky. Learning multiple layers of features from tiny images. April 2009.
- [LeCun *et al.*, 1998] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- [Leone *et al.*, 2024] Lorenzo Leone, Salvatore F.E. Oliviero, Lukasz Cincio, and M. Cerezo. On the practical usefulness of the Hardware Efficient Ansatz. *Quantum*, 8:1395, July 2024.
- [Lu *et al.*, 2019] Sirui Lu, Lu-Ming Duan, and Dong-Ling Deng. Quantum adversarial machine learning. *arXiv preprint arXiv:2001.00030*, 2019.
- [Mathur *et al.*, 2025] Aakar Mathur, Ashish Gupta, and Sajal K. Das. When federated learning meets quantum computing: Survey and research opportunities. *IEEE Communications Surveys & Tutorials*, pages 1–1, 2025.
- [McMahan *et al.*, 2016] H. Brendan McMahan, Eider Moore, Daniel Ramage, and Blaise Agüera y Arcas. Federated learning of deep networks using model averaging. *CoRR*, abs/1602.05629, 2016.
- [McMahan *et al.*, 2017] H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pages 1273–1282, 2017.
- [Neves, 2025] Nuno Neves. Mingling with the good to backdoor federated learning, 2025.
- [Nguyen *et al.*, 2022] Thien Duc Nguyen, Phillip Rieger, Huili Chen, Hossein Yalame, Helen Möllering, Hossein Fereidooni, Samuel Marchal, Markus Miettinen, Azalia Mirhoseini, Shaza Zeitouni, et al. {FLAME}: Taming backdoors in federated learning. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 1415–1432, 2022.
- [Nielsen and Chuang, 2000] Michael A. Nielsen and Isaac L. Chuang. *Quantum Computation and Quantum Information*. Cambridge University Press, 2000.
- [Paszke *et al.*, 2019] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d’Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 8024–8035, 2019.
- [Ren *et al.*, 2023] Chao Ren, Rudai Yan, Huihui Zhu, Han Yu, Minrui Xu, Yuan Shen, Yan Xu, Ming Xiao, Zhao Yang Dong, Mikael Skoglund, Dusit Niyato, and Leong Chuan Kwek. Towards quantum federated learning. *arXiv preprint arXiv:2306.09912*, 2023.

- [Shen *et al.*, 2025] Wei Shen, Wenke Huang, Guancheng Wan, and Mang Ye. Label-free backdoor attacks in vertical federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 20389–20397, 2025.
- [Yamany *et al.*, 2021] Waleed Yamany, Nour Moustafa, and Benjamin Turnbull. Oqfl: An optimized quantum-based federated learning framework for defending against adversarial attacks in intelligent transportation systems. *IEEE Transactions on Intelligent Transportation Systems*, 24(1):893–903, 2021.
- [Yin *et al.*, 2018] Dong Yin, Ying Chen, Kannan Ramchandran, and Peter L. Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. In *Proceedings of the 35th International Conference on Machine Learning*, pages 5650–5659, 2018.
- [Zhou *et al.*, 2025] Xingjie Zhou, Xianzhang Chen, Shukan Liu, Xuehong Fan, Qiao Sun, Lin Chen, Meikang Qiu, and Tao Xiang. Flguardian: Defending against model poisoning attacks via fine-grained detection in federated learning. *IEEE Transactions on Information Forensics and Security*, 20:5396–5410, 2025.