International Journal of Allied Practice, Research and Review
Website: www.ijaprr.com (ISSN 2350-1294)

StealthDistill: A Coordinated and Adaptive Distributed Backdoor Attack Framework for Knowledge Distillation-Based Federated Learning

Riaz Ahmed
Associate Professor,
Department of Computer Applications,
Government Degree College, Rajouri, Jammu and Kashmir, India.

Abstract – Federated learning (FL) systems increasingly utilize knowledge distillation (KD) to condense client updates, minimize communication cost, and permit diverse on-device models. Although KD-based FL (KD-FL) enhances efficiency, the soft-label exchange technique creates an attack surface that has not been examined before. We propose StealthDistill a coordinated and adaptive distributed backdoor attack system that utilizes the distillation channel to inject persistent backdoors bypassing state-of-the-art Byzantine-robust and anomaly-based defences. StealthDistill breaks down a single global trigger into spatially and semantically disjoint sub-triggers spread over colluding clients. We design an adaptive temperature-scaling controller that perturbs soft-label distributions sufficiently to bias the student model but without generating statistically detectable gradient or logic anomalies. A stealth scheduler based on reinforcement learning dynamically modifies the participation rate, trigger strength, and distillation temperature of each malicious client according to the observed aggregation behaviour of the server. We analyze StealthDistill on CIFAR-10, CIFAR-100 and a Non-IID FEMNIST partition, under FedAvg, FedProx and four representative defences (Krum, Multi-Krum, FoolsGold and norm-clipping with differential privacy noise). StealthDistill attains an average attack success rate (ASR) of 93.4% with the main-task accuracy degradation less than 1.2% and the detection rate reduced by 61.7% compared with the standard distributed backdoor attack (DBA). We further examine the cost of stealth in terms of convergence rounds and present two candidate mitigations, namely distillation-aware trigger forensics and temperature-anomaly monitoring.

Keywords: Federated Learning, Knowledge Distillation, Backdoor Attack, Adversarial Machine Learning, Distributed Trust, Robust Aggregation

I. Introduction

Federated learning is a collaborative model training framework without centralizing raw data over remote clients. To mitigate the high communication cost and architectural heterogeneity of cross-device FL, recent systems combine knowledge distillation, where clients

share compact soft-label or logit summaries instead of full gradient updates, with a server-side or peer-side student model that learns to mimic the aggregated teacher ensemble. We refer to this paradigm as KD-FL, which is promising for edge deployments but has not been scrutinized for security when compared to gradient-based FL.

Existing distributed backdoor attacks (DBA) divide a global trigger pattern into local pieces that are injected by colluding clients, and use the federated aggregation stage to reassemble the whole trigger in the global model. However, DBA and its variants are developed for gradient averaging aggregators, and are not optimal for the soft-label exchange and temperature-scaled training that characterizes KD-FL. We demonstrate that a naive transfer of gradient-domain DBA to the distillation domain results in logit-distribution anomalies that are observable, and inferior attack persistence in the presence of standard robust-aggregation countermeasures.

We propose StealthDistill, and make three contributions. First, we add trigger fragmentation in the soft-label space, where cooperating clients each bias a discontinuous subset of class-similarity connections instead of a shared pixel-domain pattern, making the aggregate manipulation harder to pinpoint to any single client's contribution. Secondly, we propose an adaptive temperature-scaling controller to adjust the distillation temperature in each round to keep the malicious logit perturbation within the normal variance of benign clients. Third, we offer a reinforcement-learning stealth scheduler that models defence evasion as a sequential decision problem, learning when, how much to disrupt, and how to throttle activity based on observed server-side filtering signals.

We compare StealthDistill with four countermeasures, including robust aggregation (Krum, Multi-Krum), similarity based filtering (FoolsGold), and noise based mitigation (norm-clipping with differential privacy) on three datasets and two aggregation procedures. Our results demonstrate that distillation-domain coordination yields significant improvements in both stealth and persistence over pixel-domain DBA baselines.

II. Related Work

2.1 Backdoor Attacks in Federated Learning

Early works on FL backdoor showed that a single malicious client can inject backdoor by model replacement, which still works under FedAvg because to the difficulties to differentiate malicious update and benign client drift under Non-IID data. The line of work on distributed backdoor attacks showed that distributing a global trigger among numerous colluding clients boosts attack success while keeping each particular client's contribution minimal, escaping defences that detect large-norm or high-similarity updates from a single source. Our study generalizes this decomposition technique from the gradient domain to the soft-label domain in KD-FL.

2.2 Knowledge Distillation in Federated Learning

KD-FL methods include ensemble distillation (for model fusion) and federated distillation (for communication-efficient heterogeneous FL), which replace gradient exchange with logit or soft-label interchange, commonly employing a public or synthetic proxy dataset. This decreases communication costs and facilitates client model variability, but switches the attack surface from gradient statistics to logit and temperature statistics, which conventional FL defences were not built to monitor.

2.3 Defences

Byzantine-robust aggregators such as Krum and Multi-Krum pick updates that are closest to a majority cluster in terms of pair wise distance. FoolsGold can penalize customers that consistently have similar update directions over rounds, which is a hallmark of Sybil-style cooperation. Differential-privacy noise injection and norm clipping limit the effect of every single update. These defences were mostly proven on gradient domain attacks, and our investigations quantify their effectiveness when the modified signal is instead a soft-label distribution.

III. Threat Model

We consider a typical cross-silo KD-FL scenario with N clients and a central server that aggregates soft labels from clients over a common proxy dataset to train or update a global student model. A percentage f of clients (f <= 20% in our studies) are conspiring adversaries that control their local training data, local logit computation and distillation temperature, but cannot directly manipulate the server's aggregation code or see other clients' raw updates. The adversaries can see the global model after each round (a standard assumption in the FL security literature) and utilize this feedback to change their approach in the next rounds. The attacker aims to misclassify any input with a composite trigger into an attacker-chosen target class for the global student model, while maintaining clean-data accuracy and evading detection by server-side defences.

IV. StealthDistill Methodology

4.1 Overview

StealthDistill has three coordinated components: (i) Distillation-Domain Trigger Fragmentation (DDTF) to spread the backdoor objective over the soft-label contributions of colluding clients, instead of a shared pixel pattern; (ii) an Adaptive Temperature Controller (ATC) to tune each malicious client’s distillation temperature so as to restrict the statistical deviation of its logics from the benign population; and (iii) a Reinforcement-Learning Stealth Scheduler (RLSS) to select per-round participation and perturbation intensity so as to maximize attack success under a detection-risk budget.

4.2 Distillation-Domain Trigger Fragmentation

Instead of embedding one visual trigger pattern for each malevolent client, we split the trigger into K spatially and semantically distinct sub-patterns and assign one or more sub-patterns to each cooperating client. Each malicious client k trains its local teacher on τ_k- and y_t-poisoned data, but instead of uploading raw gradients, it uploads soft-label vectors computed at a higher temperature on the shared proxy set. The server-side student is trained to predict an aggregate of these soft labels. Therefore, the composite trigger only emerges in the student's decision boundary when sufficient fragments are present in the proxy-evaluation neighbourhood. This reflects the pixel-domain logic of DBA but is entirely in the output-distribution space. This decouples the attack from the gradient norm or update direction of any single client, which is the main signal used by most existing countermeasures.

4.3 Adaptive Temperature Controller

The softness of a client's output distribution is determined by the distillation temperature T. The attacker estimates the benign population's per-round logit entropy as a Gaussian with mean μ_b and standard deviation σ_b, based on publicly observable global-model behavior and involvement in earlier rounds. For each malicious client and round, the ATC determines the temperature T^* that maximizes the KL-divergence-based backdoor signal injected into the aggregate soft label, while keeping the resulting entropy in the range of $\mu_b \pm 1.5\sigma_b$. This makes each malevolent contribution statistically indistinguishable from benign variation to simple entropy- or variance-based outlier tests.

4.4 Reinforcement-Learning Stealth Scheduler

We model the attack round by round control as a Markov Decision Process. The state includes the clean accuracy of the current global model, the estimated attack success rate on a held-out trigger set, the recent selection rate of malicious clients by the aggregator, and an estimated defense-pressure signal extracted from how often previous malicious contributions were down-weighted or excluded. The action space consists of participation decision, trigger sub-pattern intensity and the temperature offset for each malicious client. Reward is a combination of attack-success gain, accuracy-preservation and a stealth term that penalizes deviation from the benign logit-statistics envelope. We train the scheduler with Proximal Policy Optimization (PPO) in a simulated FL setting, and then apply the learned policy to the real target system, enabling the attack to generalize across defensive configurations without per-defence retraining.

4.5 Algorithm Summary

In each communication round: (1) The RLSS observes the current state and decides participation and intensity actions for each malicious client. (2) Participating malicious clients locally compute poisoned soft labels using their assigned trigger fragment and the temperature selected by the ATC. (3) All clients upload soft labels to the server, which aggregates them under the active defense and updates the global student via distillation loss. (4) The attacker estimates

the new state from observable model outputs and updates the RLSS policy via the stored reward signal. The cycle continues until a goal ASR or round budget is reached.

V. Experimental Setup

5.1 Datasets and Models

We evaluate on CIFAR-10 and CIFAR-100 (ResNet-18 teacher, ResNet-8 student) and a Non-IID FEMNIST partition (CNN teacher, MLP student), with 100 simulated clients sampled at 10% participation per round under a Dirichlet($\alpha = 0.5$) data partition to emulate realistic Non-IID heterogeneity. The adversarial fraction f is varied over {5%, 10%, 20%}.

5.2 Aggregation Protocols and Defenses

We test two base aggregation protocols, FedAvg and FedProx ($\mu = 0.01$), each combined with one of four defences: Krum, Multi-Krum ($m = 5$), FoolsGold, and norm-clipping with Gaussian DP noise ($\sigma = 0.01$, clip threshold $C = 1.0$). All defences are adapted to operate on soft-label vectors rather than gradients, following the same distance and similarity metrics as their gradient-domain formulations, applied to the distillation outputs.

5.3 Baselines

We compare StealthDistill against three baselines: (i) a single-client backdoor attack (BadNets-style) ported to KD-FL, (ii) standard pixel-domain DBA ported to KD-FL without temperature adaptation, and (iii) a static distillation attack using fixed temperature and fixed participation rate, to isolate the contribution of the RL scheduler.

5.4 Metrics

- Attack Success Rate (ASR): fraction of triggered test inputs classified as the target label.
- Main Task Accuracy (MTA) Drop: clean-accuracy degradation relative to a non-attacked baseline.
- Detection Rate (DR): fraction of malicious contributions flagged or down-weighted by the active defense.
- Rounds-to-Convergence Overhead (RCO): additional communication rounds required to reach the same MTA as a benign-only run.

VI. Experimental Results

6.1 Overall Attack Effectiveness

Table 1 reports average results across CIFAR-10, CIFAR-100, and FEMNIST at f = 10% adversarial clients, averaged over the four defences and both aggregation protocols.

Method	ASR (%)	MTA Drop (%)	Detection Rate (%)
BadNets (single-client, ported)	41.2	0.6	78.3
Pixel-domain DBA (ported)	76.5	1.1	54.9
Static Distillation Attack	81.7	1.4	47.2
StealthDistill (full)	93.4	1.2	20.9

Table 1. Average attack *success, accuracy cost, and detection rate across all datasets and defenses (f = 10%).*

6.2 Per-Defense Breakdown

Table 2 isolates StealthDistill's performance under each individual defence on CIFAR-10 (f = 10%, FedAvg), showing that similarity-based filtering (FoolsGold) remains the most effective defense, though detection rate still drops substantially relative to the DBA baseline.

Defense	DBA Baseline ASR (%)	StealthDistill ASR (%)	StealthDistill DR (%)
Krum	58.1	89.7	24.6
Multi-Krum	63.4	91.2	22.8
FoolsGold	44.9	85.0	33.5
Norm-clip + DP noise	70.2	95.8	12.4

Table 2. Per-defense attack success rate (ASR) and detection rate (DR), CIFAR-10, FedAvg, f = 10%.

6.3 Ablation Study

We ablate each component of StealthDistill on CIFAR-10 under Multi-Krum to attribute the source of stealth gain. Removing the RLSS and employing a set participation schedule reduces ASR by 9.3 points and raises detection rate by 14.1 points, demonstrating that the adaptive scheduling is the major contributor to defence evasion. Replacing the ATC with a fixed high temperature leads to a dramatic increase in detection rate (to 41.6%), establishing temperature calibration as the main mechanism for bypassing the entropy-based outlier checks. We find that removing DDTF and resorting to a shared pixel-domain trigger reduces ASR by 17.8 points, indicating that distillation-domain fragmentation is required to achieve high attack success when any single client contribution is hampered by clipping or robust aggregation.

Configuration	ASR (%)	Detection Rate (%)
Full StealthDistill	91.2	22.8
− RLSS (fixed schedule)	81.9	36.9
− ATC (fixed high T)	85.4	41.6
− DDTF (shared pixel trigger)	73.4	29.7

Table 3. Ablation of StealthDistill components, CIFAR-10, Multi-Krum, f = 10%.

6.4 Effect of Adversarial Fraction

Raising adversarial percentage from 5% to 20% enhances ASR from 79.6% to 97.1%, but increases detection rate from 14.2% to 33.8%. This is consistent with the expected trade-off between attack power and the statistical footprint left by a larger colluding population. StealthDistill shows a good stealth-to-success ratio against all baselines across the whole range.

6.5 Convergence Overhead

We see that StealthDistill adds a small overhead of convergence, needing on average 6.4 additional communication cycles compared to a benign-only run, to achieve similar main-task accuracy, against 11.2 additional rounds for the static distillation attack baseline. This implies that the adaptive throttling of the RL scheduler also has a side benefit of less interference to benign convergence dynamics, which helps the attack stay inconspicuous to round-level accuracy monitoring.

VII. Discussion and Candidate Mitigations

Our results suggest that defenses against gradient-domain attacks do not translate well to the soft-label domain employed in KD-FL, primarily because temperature scaling gives an additional degree of flexibility for an adversary to conceal distributional shifts. Based on this research, we propose two candidate mitigations, not claiming a perfect answer, but hoping to inspire future defence design.

7.1 Distillation-Aware Trigger Forensics

StealthDistill breaks down the trigger across clients in the output space. If the server groups clients based on pairwise softlabel agreement on out-of-distribution or synthetic proxy inputs, it may be able to identify coordinated sub-pattern contributions that are hidden when comparing each client's logits individually with benign-population statistics.

7.2 Temperature-Anomaly Monitoring

Rather than simply measuring their entropy at a single round, monitoring the implied temperature of soft labels submitted by each client across multiple rounds could expose adaptive

temperature manipulation; honest clients typically show consistent temperature patterns related to the difficulty of their local data, not deliberate manipulation from round to round.

7.3 Limitations

We evaluate in a simulated cross-silo setting, with idealized assumptions on the attacker's observ ability. In real deployments, the attacker's visibility to the aggregation outcomes may be limited, which would lower the effectiveness of the RL scheduler. We also compare to off-the-shelf defences not specially tailored to KD-FL but standard in the literature.

VIII. Conclusion

We proposed StealthDistill, a coordinated and adaptive networked backdoor attack technique against federated learning based on knowledge distillation. By fragmenting triggers in the soft-label domain, adaptively calibrating distillation temperature, and learning a reinforcement-learning-based participation and intensity policy, StealthDistill achieves significantly higher attack success and lower detection rates than ported gradient-domain baselines on multiple datasets and defences. Our findings inspire the development of distillation-aware countermeasures, such as trigger forensics over soft-label clustering and temperature-anomaly monitoring, as critical complements to the existing Byzantine-robust aggregation in KD-FL deployments.

IX. References

[1] Bagdasaryan, E., Veit, A., Hua, Y., Estrin, D., & Shmatikov, V. (2020). How to backdoor federated learning. AISTATS.

[2] Xie, C., Huang, K., Chen, P.-Y., & Li, B. (2020). DBA: Distributed backdoor attacks against federated learning. ICLR.

[3] Blanchard, P., El Mhamdi, E. M., Guerraoui, R., & Stainer, J. (2017). Machine learning with adversaries: Byzantine tolerant gradient descent. NeurIPS.

[4] Fung, C., Yoon, C. J. M., &Beschastnikh, I. (2020). The limitations of federated learning in sybil settings. RAID.

[5] Li, D., & Wang, J. (2019). FedMD: Heterogenous federated learning via model distillation. NeurIPS Workshop.

[6] Lin, T., Kong, L., Stich, S. U., & Jaggi, M. (2020). Ensemble distillation for robust model fusion in federated learning. NeurIPS.

[7] McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. AISTATS.

[8] Wang, H., Sreenivasan, K., Rajput, S., Vishwakarma, H., Agarwal, S., Sohn, J.-y., Lee, K., &Papailiopoulos, D. (2020). Attack of the tails: Yes, you really can backdoor federated learning. NeurIPS.