HealSplit: Towards Self-Healing through Adversarial Distillation in Split Federated Learning

Yuhan Xie[1,2], **Chen Lyu**[1,2*]

[1]Shanghai University of Finance and Economics
[2]MoE Key Laboratory of Interdisciplinary Research of Computation and Economics,
Shanghai University of Finance and Economics
yhtse@stu.sufe.edu.cn, lyu.chen@sufe.edu.cn

Abstract

Split Federated Learning (SFL) is an emerging paradigm for privacy-preserving distributed learning. However, it remains vulnerable to sophisticated data poisoning attacks targeting local features, labels, smashed data, and model weights. Existing defenses, primarily adapted from traditional Federated Learning (FL), are less effective under SFL due to limited access to complete model updates. This paper presents HealSplit, the first unified defense framework tailored for SFL, offering end-to-end detection and recovery against five sophisticated types of poisoning attacks. HealSplit comprises three key components: (1) a topology-aware detection module that constructs graphs over smashed data to identify poisoned samples via topological anomaly scoring (TAS); (2) a generative recovery pipeline that synthesizes semantically consistent substitutes for detected anomalies, validated by a consistency validation student; and (3) an adversarial multi-teacher distillation framework trains the student using semantic supervision from a Vanilla Teacher and anomaly-aware signals from an Anomaly-Influence Debiasing (AD) Teacher, guided by the alignment between topological and gradient-based interaction matrices. Extensive experiments on four benchmark datasets demonstrate that HealSplit consistently outperforms ten state-of-the-art defenses, achieving superior robustness and defense effectiveness across diverse attack scenarios.

Introduction

Split Federated Learning (SFL) (Thapa et al. 2022) integrates the strengths of Federated Learning (FL) (Liu et al. 2024; Yazdinejad et al. 2024) and Split Learning (SL) (Vepakomma et al. 2018; Lin et al. 2024), offering enhanced privacy protection and reduced computational overhead. An SFL architecture comprises client-side and server-side models. Each client performs the forward pass locally using its client-side model, and then transmits the resulting smashed data (intermediate representations) along with corresponding labels to the server. The server-side model performs the rest of the forward and backward computations, producing gradients concerning the smashed data. These gradients are then returned to clients to update their local models. During this process, client-side updates are aggregated by the Fed server, while server-side updates are aggregated by the main server.

Although SFL is recognized as a robust and privacy-preserving learning paradigm (Chen and Zhang 2022), recent studies have revealed its susceptibility to various data poisoning attacks (Fang et al. 2020; Wu et al. 2024). These attacks aim to compromise the learning process by manipulating sophisticated malicious data or modifying model weights, ultimately degrading the performance of the global model or causing misclassifications. The collaborative and split nature of SFL introduces multiple potential attack surfaces (Ma et al. 2022), including local features (Tolpegin et al. 2020), labels (Gajbhiye, Singh, and Gupta 2022; Ismail and Shukla 2023), smashed data (Wu et al. 2024), and client-side model weights (Fang et al. 2020; Khan et al. 2022), each of which can be exploited to disrupt training or corrupt model integrity.

To defend against these sophisticated data poisoning attacks in SFL, existing defense strategies remain inadequate, primarily because they are adapted from traditional FL settings. Techniques such as Krum and Multi-Krum (MKRum) (Blanchard et al. 2017), Trimmed Mean and Median (Yin et al. 2018), and Bulyan (Guerraoui, Rouault et al. 2018) primarily rely on statistical aggregation to filter out anomalous or malicious gradients. More advanced defenses, including FLTrust (Cao et al. 2020), DnC (Shejwalkar and Houmansadr 2021), FedDMC (Mu et al. 2024), and ShieldFL (Ma et al. 2022), enhance robustness through trust-based scoring, dimensionality reduction, structural modeling, or encrypted similarity matching. However, these methods typically assume access to full model updates or raw gradients from all clients (Yazdinejad et al. 2024). Such an assumption does not hold in the SFL setting due to its architectural split and the transmission of smashed data. Consequently, their defensive effectiveness is significantly compromised in SFL. Furthermore, most defenses are designed to address isolated attack vectors, limiting their generalizability against the diverse and complex threat landscape inherent in SFL (Ma et al. 2022).

In this paper, we propose HealSplit, a unified defense framework that delivers end-to-end protection against diverse and sophisticated data poisoning attacks in SFL by seamlessly integrating detection and recovery mechanisms. In the architecture of SFL, the smashed data transmitted

*Chen Lyu is the corresponding author.

from clients to the server serves as the primary medium for poisoning attacks across various stages of the learning process (Wu et al. 2024). Accordingly, HealSplit focuses on securing the smashed data to defend against a broad spectrum of poisoning attacks. As shown in Fig. 1, to effectively detect poisoned samples that exhibit anomalous connectivity patterns in the smashed data, a topology-aware detection module is proposed. This module computes a topological anomaly score (TAS) using Personalized PageRank (PPR) (Gasteiger, Bojchevski, and Günnemann 2018) on a k-nearest neighbors (KNN) graph.

To further rectify the deviations induced by data poisoning attacks, HealSplit incorporates a GAN-based module (Huang et al. 2024) to generate high-quality substitute representations for detected poisoned samples. These synthetic representations are then validated for semantic consistency using a dedicated student model. The student model is trained via adversarial multi-teacher distillation (Sun et al. 2025), with supervision provided by two complementary sources: (1) the Anomaly-Influence Debiasing (AD) Teacher, which regulates inter-task information propagation through an inter-task influence matrix designed to mitigate anomaly-induced bias. This matrix combines the TAS with Gradient Interaction Scores (GIS), enabling the model to selectively propagate information along reliable and structurally consistent task-label paths; and (2) the Vanilla Teacher, which preserves semantic integrity by modeling the distribution of clean data. Contributions from both teachers are dynamically balanced using a momentum-adaptive optimization strategy. Additionally, we theoretically prove that HealSplit reduces gradient variance on the server side by improving gradient similarity.

To comprehensively evaluate the robustness and generalization of HealSplit, we conduct extensive experiments across multi-vector attacks (e.g., DP + SP), heterogeneous data distributions (i.e., IID $\leftrightarrow$ non-IID), model architectures (e.g., ResNet18 $\leftrightarrow$ VGG16), and adaptive attack strategies. The results show that HealSplit consistently outperforms state-of-the-art defenses, which often fail under dynamic and challenging real-world conditions. Even in the presence of adaptive attacks, it maintains quite a high accuracy, showcasing strong resilience and broad applicability in SFL.

Our contributions are summarized as follows:

- We propose the first unified defense framework for SFL that effectively tackles five challenging and diverse attack types: label poisoning, data poisoning, smashed data poisoning, weight poisoning, and multi-vector poisoning.
- We introduce a topology-aware detection mechanism that constructs a graph over smashed data and computes a TAS, enabling the detection of poisoned samples by capturing both local and global structural anomalies.
- We propose a consistency validation student to verify GAN-generated replacements and ensure semantic fidelity. It is optimized via a momentum-adaptive strategy under an adversarial distillation framework.
- Extensive experiments on four benchmark datasets demonstrate that HealSplit consistently outperforms ten classical and advanced baselines in both defense efficacy and robustness.

Related Work

Defenses of Split Federated Learning. Classic defenses adopt statistical aggregation to mitigate malicious gradients, including Krum and Mkrum (Blanchard et al. 2017), TrimMean and Median (Yin et al. 2018), and Bulyan (Guerraoui, Rouault et al. 2018). In contrast, advanced defenses emphasize robust aggregation and anomaly detection. FLTrust (Cao et al. 2020) assigns trust scores based on the alignment between client updates and the server model, and normalizes update magnitudes to reduce the impact of malicious inputs. DnC (Shejwalkar and Houmansadr 2021) identifies outliers via principal component analysis and leverages dimensionality reduction for efficiency. Feddmc (Mu et al. 2024) detects malicious clients through dimensionality reduction, noise-resistant clustering, and self-ensemble correction. ShieldFL (Ma et al. 2022) incorporates two-trapdoor homomorphic encryption for secure aggregation and employs cosine similarity to detect and suppress encrypted model poisoning.

However, these existing defenses are inherited from FL, and recent work (Wu et al. 2024) demonstrates their inconsistent effectiveness against poisoning attacks in SFL.

Adversarial Distillation. Knowledge distillation (Hinton, Vinyals, and Dean 2015) is an effective technique for transferring knowledge from a large teacher model to a smaller student model. Recently, adversarial distillation (Goldblum et al. 2020; Zhao, Wang, and Wei 2024) has gained traction for not only improving model compression but also enhancing robustness (Singh, Croce, and Hein 2023; Angarano et al. 2024) and fairness (Chai, Jang, and Wang 2022; Li et al. 2024) through adversarial training (Ganin and Lempitsky 2015). Among them, DTDBD (Li et al. 2024), which mitigates domain bias via a dual-teacher framework that balances an unbiased teacher for de-biasing and a clean teacher for domain knowledge transfer. Similarly, B-MTARD (Zhao, Wang, and Wei 2024) refines adversarial training by integrating a clean and a robust teacher, employing entropy-based and normalization loss balancing to enhance both accuracy and robustness.

Background

Problem Statement

Let the SFL system consist of N clients $\{c_i\}_{i=1}^{N}$, each holding a private dataset $\mathcal{D}_i = \mathcal{D}_i^{\text{tr}} \cup \mathcal{D}_i^{\text{te}} \sim \mathcal{P}_i$, where $\mathcal{P}_i$ denotes the local data distribution of client c_i. The training and test sets are $\mathcal{D}_i^{\text{tr}} = \{(x_j, y_j)\}_{j=1}^{m_i^{\text{tr}}}$ and $\mathcal{D}_i^{\text{te}} = \{(x_j, y_j)\}_{j=1}^{m_i^{\text{te}}}$, where x_j and $y_j \in \mathcal{Y}$ denote input features and labels, and m_i^{tr} and m_i^{te} denote the number of training and test samples, respectively. Clients $\mathcal{C} = \{c_i\}_{i=1}^{N}$ are partitioned into benign $\mathcal{C}_{\text{ben}}$ and malicious $\mathcal{C}_{\text{att}} = \mathcal{C} \setminus \mathcal{C}_{\text{ben}}$.

Each client maintains a local model $g_{\theta_{c_i}}$, which maps inputs to smashed data $z_j = g_{\theta_{c_i}}(x_j) \sim \mathcal{Q}_i$. These smashed data are sent to a server-side model $h_{\theta_{s_i}}$ for forward computation. In back-propagation, the resulting gradients are used

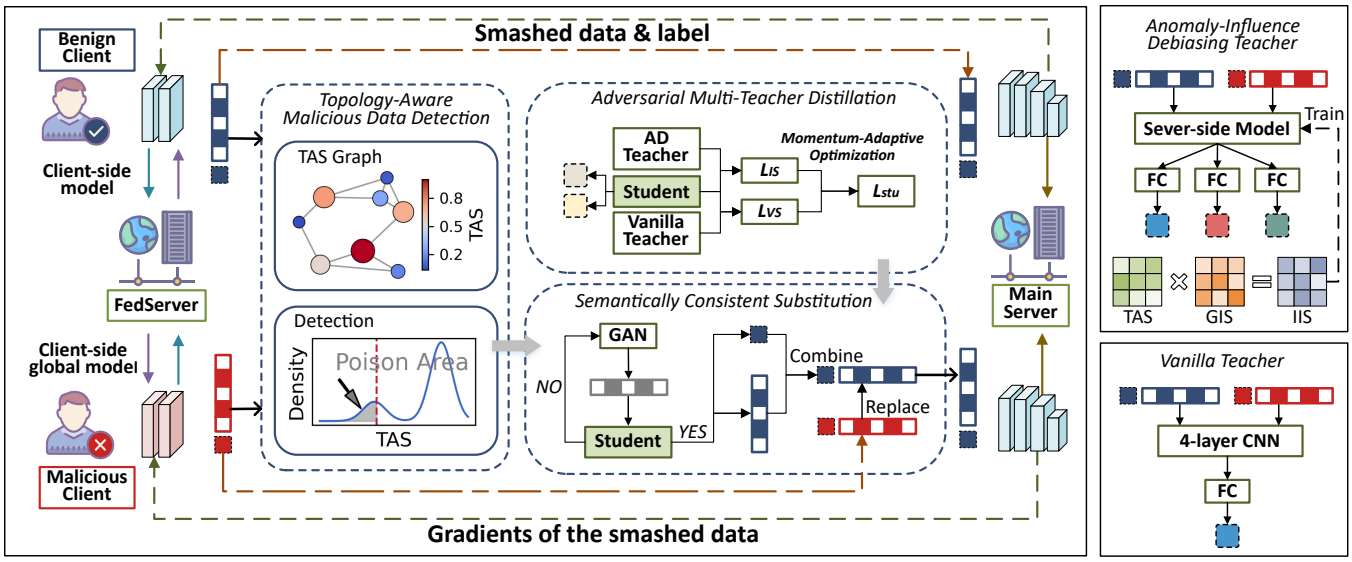


Figure 1: The framework of HealSplit. HealSplit first detects poisoned samples using a KNN-based TAS, and then employs a GAN to generate substitute smashed data. These substitutes are subsequently validated by a consistency validation student model, which is trained via adversarial multi-teacher distillation to ensure semantically consistent substitution.

to update θ_{c_i}. To preserve privacy, the updated client-side and server-side models are sent to the federated server and the main server for aggregation, respectively. The complete model for client c_i is $f_{\theta_i} = h_{\theta_{s_i}} \circ g_{\theta_{c_i}}$.

Defender Objective. The defender aims to balance robustness against diverse poisoning attacks and clean performance, formulated as a regularized optimization problem:

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}^{te}} \left[\ell(f_\theta(x), y)\right] + \mu\mathcal{R}_{\text{robust}}(\theta), \tag{1}$$

where $\ell(\cdot,\cdot)$ is a task-specific loss function, $\mathcal{R}_{\text{robust}}(\theta)$ measures the model's sensitivity to poisoning behaviors, and μ controls the trade-off between accuracy and robustness.

Threat Model

In SFL, attackers with varying levels of knowledge and capability can launch diverse and sophisticated data poisoning attacks targeting different stages of the learning pipeline. We categorize these attacks into five major types:

Label Poisoning (LP, $\mathcal{A}_1$): The ground-truth label is perturbed as $y'_j = (y_j + \delta_y) \bmod C$, where C is the number of classes and δ_y is the label shift.

Data Poisoning (DP, $\mathcal{A}_2$): The local dataset is modified as $\mathcal{D}'_k = \{(x'_j, y_j)\}_{j=1}^{m'}$, where $x'_j = x_j + \delta_x$ and δ_x denote the input perturbation.

Smashed Poisoning (SP, $\mathcal{A}_3$): The smashed data are modified as $z'_j = g_\phi(x_j) + \delta_z$, where $g_\phi(x_j)$ represents the smashed data and δ_z is the feature-level perturbation.

Weight Poisoning (WP, $\mathcal{A}_4$): The model parameters are manipulated before aggregation, expressed as $\theta' = \theta + \Delta_\theta$, where Δ_θ represents the weight perturbation.

Multi-Vector Poisoning: This is a composite attack strategy that integrates multiple poisoning techniques, which is defined as: $\mathcal{A}_{Multi} = \bigcup_{i=1}^{4} S_i\mathcal{A}_i$, $S \in \{0,1\}^4$, where S_i indicates whether the i-th attack $\mathcal{A}_i$ is applied.

Methodology

Topology-Aware Malicious Data Detection

Inspired by the graph propagation mechanisms in social networks (Zhu et al. 2024; Cui and Jia 2024), our detection framework exploits the topological properties of poisoned data in SFL. As shown in Fig. 2b, poisoned samples tend to form locally dense, yet globally isolated clusters in the feature space. This is characterized by (1) high feature similarity within malicious samples, and (2) weak connections to benign data. These observations suggest that topology-based detection can effectively identify poisoning patterns.

Graph Representation. Given smashed data and labels $\mathcal{D} = \{(z_k, y_k)\}_{k=1}^{K}$ obtained from SFL rounds, the weighted graph $G = (V, E)$ is represented by an adjacency matrix $\mathbf{W}$:

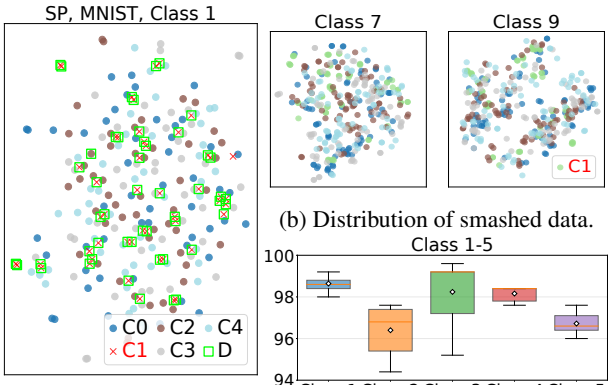
$$W_{kj} = \begin{cases} \exp\left(-\gamma\|z_k - z_j\|^2\right), & \text{if } z_j \in \mathcal{N}_k \text{ and } z_k \in \mathcal{N}_j \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where $\gamma = (2\sigma^2)^{-1}$ and σ is the median between all pairs of points in the KNN graph. $\mathcal{N}_k$ denotes the KNN set of z_k.

Topological Anomaly Score. To identify topologically anomalous nodes exhibiting deviations in propagation patterns, we compute the TAS r using PPR, capturing both local and global graph structures. TAS is initialized based on node degrees and updated iteratively at each propagation step t:

$$r_k^{(t+1)} = \mathbb{I}_{[t=0]} \cdot \frac{1}{d_k+\epsilon} + \mathbb{I}_{[t\geq 1]} \cdot \left(\alpha \sum_{w\in\mathcal{N}(k)} \frac{r_w^{(t)}}{d_w} + (1-\alpha)v_k\right), \tag{3}$$

where $\mathbb{I}[\cdot]$ denotes the indicator function, d_w is the degree of node w, v_k is the personalized teleportation vector, ϵ is a small constant to prevent division by zero, and α controls the trade-off between local and global propagation.



(a) Identification of poisoned data.

(b) Distribution of smashed data.

(c) Detection rate across classes.

Figure 2: Topology-aware malicious data detection: (a) the detection performance, (b) the distribution of smashed data, and (c) detection statistics across different classes. The red crosses denote the smashed data transmitted by malicious client c_1, while D denotes the detected malicious samples.

Adaptive Threshold. For automatic anomaly detection, we apply kernel density estimation to the TAS:

$$\hat{f}(r) = \frac{1}{Kh} \sum_{k=1}^{K} \mathcal{K}\left(\frac{r - r_k}{h}\right), \tag{4}$$

where $\mathcal{K}(\cdot)$ is the Gaussian kernel function, r denotes the evaluation point, and h is the bandwidth. The detection threshold T is defined as:

$$T = \min\left(\underset{r}{\operatorname{argmin}} \hat{f}(r), Q_\rho(\{r_k\})\right), \tag{5}$$

where $Q_\rho(\{r_k\})$ denotes the ρ-percentile of the score set $\{r_k\}$. Data with scores below the adaptive threshold T are marked as poisoned. Detection results are shown in Fig. 2.

Semantically Consistent Substitution

To replace detected malicious smashed data, we train a vanilla GAN using the identified clean smashed data:

$$\begin{aligned} \mathcal{L}_{\mathrm{D}} &= -\mathbb{E}_{\mathbf{z}}[\log D(\mathbf{z})] - \mathbb{E}_{\tilde{\mathbf{z}}}[\log(1 - D(\tilde{\mathbf{z}}))] \\ \mathcal{L}_{\mathrm{G}} &= -\mathbb{E}_{\tilde{\mathbf{z}}}[\log D(\tilde{\mathbf{z}})] \end{aligned} \tag{6}$$

where $\mathbf{z}$ represents clean smashed features, $\tilde{\mathbf{z}}$ are generated features, and $D(\cdot)$ is the discriminator network.

HealSplit synchronizes with SFL at appropriate intervals, using only the smashed data from the current update round to train the GAN. Although the generator aims to approximate the distribution of clean features, limited training data may lead to semantically inconsistent outputs. To ensure reliability, each generated sample is evaluated by a consistency validation student, and only those with high confidence and label consistency are selected to replace poisoned data.

Anomaly-Influence Debiasing Teacher

In SFL, malicious clients often employ similar attack patterns (Alsaheel et al. 2021; Luo et al. 2022), leading to biased training that disrupts global aggregation (He et al. 2024; Alber et al. 2025). To capture these patterns, we utilize TAS and GIS to train an AD teacher model. This model dynamically adjusts label influence by amplifying those that facilitate poisoning detection and attenuating those that obscure it, thereby guiding the learning process toward more effective classification.

Gradient Interaction Score. In multi-task learning, gradients from different tasks propagate through shared parameters, resulting in inter-task interactions. (Yu et al. 2020; Sun et al. 2025). Inspired by this, we define the GIS to model two types of inter-task interactions: one between poisoning patterns and client identity, and the other between poisoning patterns and category semantics.

Let $\mathcal{T} = \{a, b, c\}$ represent the tasks, where a corresponds to poisoning identification, b to client identification, and c to category classification. Each task $t \in \mathcal{T}$ has a corresponding label set $\mathcal{Y}_t$. For a task pair $p = (t_x, t_y) \in \{(a, b), (a, c)\}$, the GIS between individual labels $y_{t_x} \in \mathcal{Y}_{t_x}$ and $y_{t_y} \in \mathcal{Y}_{t_y}$ is represented as a matrix $\mathbf{G}_p$:

$$\mathbf{G}_p = \begin{bmatrix} \cos\left(\nabla_{t_x}(y_{t_x}^{(1)}), \nabla_{t_y}(y_{t_y}^{(1)})\right) & \cdots & \cos\left(\nabla_{t_x}(y_{t_x}^{(1)}), \nabla_{t_y}(y_{t_y}^{(|\mathcal{Y}_{t_y}|)})\right) \\ \vdots & \ddots & \vdots \\ \cos\left(\nabla_{t_x}(y_{t_x}^{(|\mathcal{Y}_{t_x}|)}), \nabla_{t_y}(y_{t_y}^{(1)})\right) & \cdots & \cos\left(\nabla_{t_x}(y_{t_x}^{(|\mathcal{Y}_{t_x}|)}), \nabla_{t_y}(y_{t_y}^{(|\mathcal{Y}_{t_y}|)})\right) \end{bmatrix}, \tag{7}$$

Where cosine similarity $\cos(\cdot, \cdot)$ measures the alignment between task gradients. The magnitude of $\mathbf{G}_p(\cdot, \cdot)$ reflects the degree of gradient alignment between tasks: larger values denote cooperative interactions conducive to unbiased optimization, while smaller values indicate conflicting objectives that may hinder learning.

Loss Function of AD Teacher. To mitigate anomaly-induced bias, we control inter-task interference by designing an inter-task influence score matrix $\mathbf{M}_p$. Specifically, $\mathbf{G}_p$ serves as a structural prior to guide the construction of a label-aware transition matrix that integrates task-level and label-level relationships. For each task pair p, $\mathbf{M}_p$ is computed by combining the TAS matrix $\mathbf{R}$ with the corresponding GIS matrix $\mathbf{G}_p$:

$$\mathbf{M}_p = (1 - \beta)\mathbf{E}^\top \left(\mathbf{I}_K - \beta \cdot \left(\mathrm{RowNorm}\left(\mathbf{R} \odot (\mathbf{E}\mathbf{G}_p\mathbf{F}^\top)\right)\right)\right)^{-1} \mathbf{F} \tag{8}$$

where $\beta \in (0, 1)$ is the restart probability controlling the range of information propagation, $\mathbf{E} \in \{0, 1\}^{K \times |\mathcal{Y}t_x|}$ and $\mathbf{F} \in \{0, 1\}^{K \times |\mathcal{Y}t_y|}$ are the node-to-label mapping matrices for tasks t_x and t_y, respectively, $\odot$ denotes the Hadamard product, $\mathbf{I}_K$ is the $K \times K$ identity matrix, and K, $|\mathcal{Y}t_x|$, and $|\mathcal{Y}t_y|$ denote the number of nodes and the sizes of the label sets $\mathcal{Y}t_x$ and $\mathcal{Y}_{t_y}$, respectively.

The final loss function of AD Teacher is:

$$\begin{aligned} \mathcal{L}_{\mathrm{AD}} = \sum_{k=1}^{K} (&\mathcal{L}_a(\hat{\mathbf{y}}_k^a, \mathbf{y}_k^a) + \lambda_b[\mathbf{M}_{(a,b)}]_{\mathbf{y}_k^a, \mathbf{y}_k^b} \mathcal{L}_b(\hat{\mathbf{y}}_k^b, \mathbf{y}_k^b) \\ &+ \lambda_c[\mathbf{M}_{(a,c)}]_{\mathbf{y}_k^a, \mathbf{y}_k^c} \mathcal{L}_c(\hat{\mathbf{y}}_k^c, \mathbf{y}_k^c)), \end{aligned} \tag{9}$$

where $\mathcal{L}_a$, $\mathcal{L}_b$, and $\mathcal{L}_c$ denote the loss functions for poisoning identification, client identification, and category classification, respectively, with λ_b and λ_c balancing the latter two

tasks. The term $[\mathbf{M}_p]_{\mathbf{y}_k^a,\mathbf{y}_k^b}$ quantifies the influence between the label sets of tasks a and b for sample k, where $\mathbf{y}_k^a$ and $\mathbf{y}_k^b$ are their respective labels. Similarly, $[\mathbf{M}_{(a,c)}]_{\mathbf{y}_k^a,\mathbf{y}_k^c}$ captures the influence between tasks a and c.

Consistency Validation Student

Inspired by adversarial (Sauer et al. 2024b,a) and multi-teacher distillation (Wen et al. 2024; Ma et al. 2024), we integrate both to enhance smashed data consistency, improving performance and reducing bias (Li et al. 2024). Two teachers capture complementary data aspects in an adversarial setup, while a momentum-adaptive design enables the student to learn more robust and generalized representations.

Adversarial Multi-Teacher Distillation. The Vanilla Teacher is trained with the identified smashed data, the training loss is:

$$\mathcal{L}_{\text{Van}} = \sum_{k=1}^{K} \mathcal{L}_a(\hat{\mathbf{y}}_k^a, \mathbf{y}_k^a). \tag{10}$$

The adversarial distillation loss for transferring knowledge from both the Vanilla Teacher model $h_{T_{\text{van}}}$ and the AD Teacher model $h_{T_{\text{AD}}}$ to the student model h_S is:

$$\mathcal{L}_{\text{VS}} = \tau^2 \cdot \text{KL}(\text{LogSoftmax}\left(h_{T_{\text{van}}}(z_i)/\tau\right), \text{Softmax}\left(h_S(z_i)/\tau\right)), \tag{11}$$

$$\mathcal{L}_{\text{IS}} = \tau^2 \cdot \text{KL}(\text{LogSoftmax}\left(h_{T_{\text{AD}}}(z_i)/\tau\right), \text{Softmax}\left(h_S(z_i)/\tau\right)), \tag{12}$$

where τ is the temperature parameter.

The Vanilla and AD Teachers serve complementary roles in adversarial distillation: the Vanilla Teacher captures clean semantics to identify valid features, while the AD Teacher focuses on anomalies to guide deviation detection. This synergy enables the student to integrate semantic clarity with anomaly awareness, ensuring robust evaluation of GAN-generated features for consistency and label alignment.

Momentum-Adaptive Optimization. In order to balance the contributions of the AD Teacher and the Vanilla Teacher and to prevent either from dominating. We design a momentum-adaptive optimization scheme. The total loss of the consistency validation student is defined as:

$$\mathcal{L}_{\text{Stu}} = \sum_{k=1}^{K} (\mathcal{L}_a(\hat{\mathbf{y}}_k^a, \mathbf{y}_k^a) + \lambda_b \mathcal{L}_b(\hat{\mathbf{y}}_k^b, \mathbf{y}_k^b) + \mu \mathcal{L}_{\text{VS}} + \eta \mathcal{L}_{\text{IS}}), \tag{13}$$

where μ, η weight the contributions from the Vanilla and AD Teachers, respectively. We update these weights at each iteration via momentum-based rules:

$$\mu_t = m \cdot \mu_{t-1} + (1-m) \cdot \sigma\left(\kappa \cdot \frac{\mathcal{L}_{\text{VS}} - \mathcal{L}_{\text{IS}}}{\mathcal{L}_{\text{VS}} + \mathcal{L}_{\text{IS}} + \epsilon}\right), \tag{14}$$

$$\eta_t = m \cdot \eta_{t-1} + (1-m) \cdot \sigma\left(\kappa \cdot \frac{\mathcal{L}_{\text{IS}} - \mathcal{L}_{\text{VS}}}{\mathcal{L}_{\text{IS}} + \mathcal{L}_{\text{VS}} + \epsilon}\right), \tag{15}$$

where $m \in (0, 1]$ is the momentum parameter, κ is a scaling factor, $\sigma(\cdot)$ is the sigmoid function, and ϵ is a small constant. This dynamic adjustment mechanism enables a smooth and gradual balance between the contributions of the two teachers, thereby enhancing model stability and overall performance during training. The momentum term facilitates steady updates to μ and η, preventing abrupt shifts.

Parameter	Value	Parameter	Value
Number of clients	10	Learning rate	1×10^{-4}
Malicious client ratio	20	Training epochs	100
Krum parameter	10	Participation rate	100
Trimmed parameter	10	Attack Method	DP+SP
Sparsefed paprameter	60	Batch size	64
Bottleneck dimension	3	Local epochs	1
Coordinate updates	30,000	Attack epochs	100
Attack learning rate	0.01	Base Model	Resnet-18
Dataset	MNIST	AD Teacher	Sever-side model
Vanilla Teacher	4-layer CNN	Student Model	4-layer CNN

Table 1: Experimental parameter settings.

Theoretical Foundations of HealSplit

In SFL, where models are divided into client-side models $g_{\theta_{c_i}}$ and server-side models $h_{\theta_{s_i}}$, we extend the concept of Inter-client Gradient Variance (CGV) (Kairouz et al. 2021; Karimireddy et al. 2020) to introduce Inter-Server Gradient Variance (SGV) and establish its upper bound (Kairouz et al. 2021; Woodworth, Patel, and Srebro 2020):

Definition 1. ***Inter-Server Gradient Variance (SGV):*** *SGV*$(F, \theta_s) = \mathbb{E}_{(z,y)\sim\mathcal{Q}_n} \left\|\nabla_{\theta_s} f_n(\theta_s; z, y) - \nabla_{\theta_s} F(\theta_s)\right\|^2$. *SGV is assumed to be upper-bounded, i.e., there exists a constant* σ *such that SGV*$(F, \theta_s) \leq \sigma^2$.

HealSplit enhances smashed data quality by replacing anomalous smashed data identified through topological detection with reliable substitutes that align with the global distribution $\mathcal{Q}$. This process not only improves generalization but also effectively reduces SGV.

To formalize this effect, consider a training round T where a client c_n transmits smashed data composed of m_n clean samples and $\widehat{m}_n$ poisoned samples drawn from its local training set $\mathcal{D}_n^{\text{tr}}$. Across all clients, the total number of clean and poisoned samples satisfies $\sum_{n=1}^{N} m_n = M$ and $\sum_{n=1}^{N} \widehat{m}_n = \widehat{M}$, respectively. Suppose a fraction α of the clients are malicious. Under the SGV framework, after performing semantically consistent substitution via HealSplit, the following theorem establishes how the objective function constrains gradient dissimilarity:

Theorem 1. *Under the* ***SGV*** *framework (Definition 1), if the ratio of clean samples in a client's dataset satisfies:* $\frac{m_n}{m_n+\widehat{m}_n} = \frac{M}{M+\widehat{M}}$, *then the robust objective* $\widehat{F}(\theta_s)$ *bounds the gradient dissimilarity:* *SGV*$(\widehat{F}, \theta_s) = \frac{\alpha^2 M^2}{(M+\widehat{M})^2} \left\|\nabla_{\theta_s} f_n(\theta_s; z, y) - \nabla_{\theta_s} F(\theta_s)\right\|^2 \leq$ *SGV*(F, θ_s).

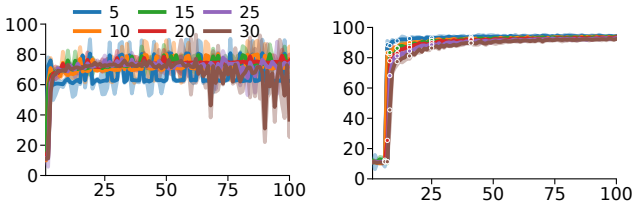
Experiment

Experimental Setup

Baselines and Metrics. We assess four attack strategies (DP, WP, SP, and LP) and their combinations (DP + SP, WP + SP, and LP + SP). For comparison, we evaluate ten defense methods: FedAvg (McMahan et al. 2017), Trim-Mean, Median, Sparsefed (Panda et al. 2022), Krum, Bulyan, FLTrust,

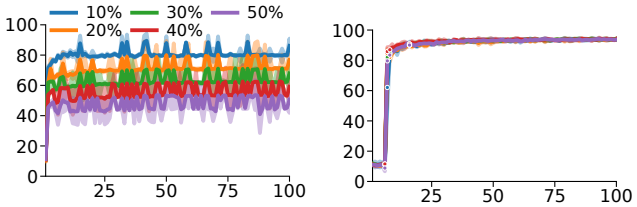
Defense Method	W/o A	DP	WP	SP	LP	DP+SP	WP+SP	LP+SP
FedAvg	96.90±0.01	10.12±0.85	44.74±19.73	**96.90±0.12**	79.23±0.73	9.19±0.18	68.22±0.48	64.82±2.42
Trim-Mean	**97.61±0.58**	11.12±1.25	64.82±10.17	94.88±0.45	79.71±2.32	10.98±0.28	70.31±2.25	68.12±7.47
Median	93.84±1.36	46.42±3.39	62.90±9.43	68.57±3.18	80.44±1.81	11.45±2.34	46.64±8.29	59.86±7.67
Sparsefed	96.65±0.54	9.51±1.19	13.23±1.14	20.06±10.16	74.50±2.24	9.35±0.48	75.24±0.85	69.75±2.87
Krum	96.66±1.18	76.77±3.71	15.91±4.86	71.62±0.91	82.95±0.71	70.48±1.38	76.20±0.53	70.68±3.22
Bulyan	96.85±0.62	10.64±1.19	21.90±1.97	96.82±2.62	77.26±2.23	10.81±2.41	73.00±2.56	69.23±1.72
FLTrust	96.52±1.03	76.48±1.20	48.70±6.11	94.42±1.18	55.56±3.99	73.39±1.44	11.33±4.89	32.41±4.50
DnC	97.27±1.00	80.58±1.32	82.18±2.58	95.33±0.98	80.43±1.86	76.34±2.33	78.82±5.65	75.33±2.69
Feddmc	92.48±2.39	75.80±2.17	31.96±4.18	90.79±1.65	62.53±4.51	75.23±1.50	30.42±5.80	11.51±0.84
ShieldFL	97.58±0.84	83.73±0.90	84.24±1.91	96.35±2.56	78.18±2.57	75.54±2.39	75.16±4.30	12.97±2.17
HealSplit	97.17±1.27	**96.86±0.77**	**95.99±1.69**	96.75±0.86	**96.72±0.64**	**93.88±0.60**	**92.44±0.73**	**93.88±1.38**

Table 2: Results of each defense method under different attack types ("W/o A" denotes the absence of attacks).



(a) DnC (b) HealSplit

Figure 3: Defense efficacy across varying client numbers. The circles represent the number of update rounds for the defense model.

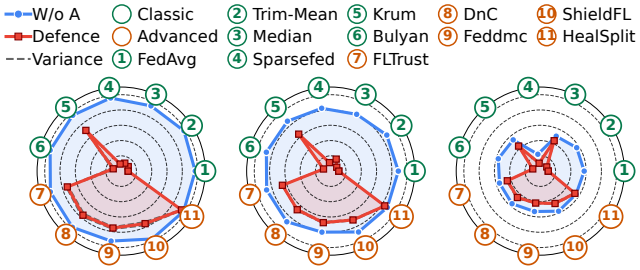


(a) DnC (b) HealSplit

Figure 4: Defense efficacy across varying proportions of malicious clients.

DnC, Feddmc (Mu et al. 2024), and ShieldFL. The primary evaluation metric is the reduction in accuracy across all task test sets, which reflects the effectiveness of each defense.

Datasets and Models. We evaluate HealSplit on four image datasets: MNIST (LeCun, Cortes, and Burges 1998), F-MNIST (Xiao, Rasul, and Vollgraf 2017), CIFAR-10 (Krizhevsky, Hinton et al. 2009), and HAM10000 (Tschandl, Rosendahl, and Kittler 2018). Among these, HAM10000 follows a non-IID distribution, while the others are IID. To evaluate HealSplit's robustness under non-IID conditions, we construct a Non-IID MNIST variant M-q, with larger q indicating greater non-IIDness. We experiment with three commonly used architectures: ResNet-18 (R18) (He et al. 2016), ResNet-152 (R152), and VGG16 (Simonyan and Zisserman 2014).



(a) MNIST (b) F-MNIST (c) CIFAR-10

Figure 5: Defense efficacy across datasets.

SFL System Settings. Unless otherwise specified, each dataset is partitioned among 10 clients, with 20% acting as malicious participants. The SFL framework is trained for 100 epochs using FedAvg as the default aggregation strategy, under a combined DP and SP attack scheme. Default system configurations are listed in Table 1.

Experimental Results

Robustness under Diverse Threats. Our first set of experiments evaluates HealSplit's robustness against varying attack strategies, as illustrated in Table 2.

HealSplit demonstrates consistently strong defense performance, maintaining over 92% accuracy across all attack scenarios with minimal degradation. It significantly outperforms advanced baselines, which often exhibit large accuracy fluctuations under different attack types. Unlike conventional defenses tailored to specific threats, HealSplit remains robust even against challenging composite attacks.

Notably, state-of-the-art methods such as FLTrust fail under combined attacks like WP+SP, with accuracy dropping to 11.33%, exposing critical vulnerabilities in the SFL setting. In contrast, HealSplit is inherently attack-agnostic: it requires no prior knowledge of the attack type and avoids fixed defense assumptions.

Moreover, HealSplit detects anomalous client behavior in real time and adaptively adjusts decision thresholds during training, eliminating the need for manual hyperparameter tuning required by other methods such as Krum's neighbor count or SparseFed's norm clipping threshold.

Component	MNIST	F-MNIST	CIFAR-10	HAM10k
HealSplit	**93.88±0.10**	**84.11±0.47**	**53.87±1.11**	**72.27±0.52**
w/o Vanilla Teacher	90.99±1.16	80.64±1.49	51.27±1.35	69.64±1.51
w/o AD Teacher	87.34±0.16	75.17±0.39	46.40±0.60	63.20±0.45
w/o Distillation	74.38±3.20	69.65±3.79	42.75±2.66	59.61±2.49
w/o Adversarial	92.74±1.68	82.59±1.37	51.55±1.67	70.40±1.25

Table 3: Results of ablation study.

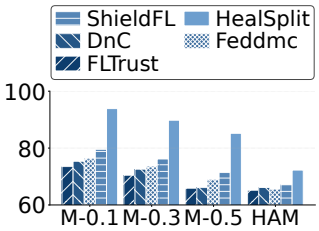


Figure 6: Generalization across non-IID datasets

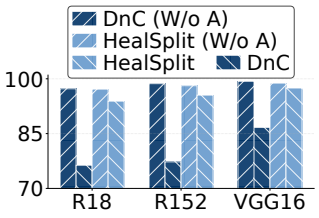


Figure 7: Generalization across models.

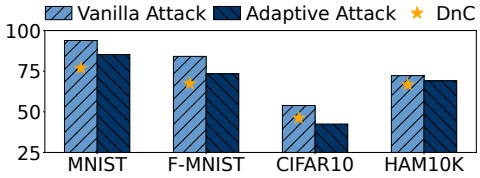


Figure 8: Robustness of HealSplit under adaptive attack.

Ablation Study. Our second set of experiments conducts the ablation study under the strong composite attack DP+SP to assess the contribution of each component in HealSplit. The results are presented in Table 3.

Among all components, the AD Teacher and the distillation mechanism contribute most significantly to HealSplit's overall robustness. Removing the AD Teacher causes a substantial drop in robustness, indicating its role in mitigating model bias through real-time behavioral adjustment. The distillation mechanism is equally critical, as its removal consistently lowers accuracy across tasks, reflecting its effectiveness in integrating multi-task knowledge and enhancing generalization. Excluding the Vanilla Teacher destabilizes training and increases sensitivity to noisy updates, emphasizing its function as a clean semantic reference. Finally, disabling the adversarial mechanism considerably weakens defense performance, highlighting its importance in strengthening resistance to strong attacks.

Defense Under System Variations. Our third set of experiments evaluates the impact of client configurations, including the number of clients and the proportion of adversaries. The results are presented in Fig. 3 and Fig. 4.

HealSplit achieves consistently better performance than the state-of-the-art method DnC after a few rounds of fine-tuning, demonstrating superior robustness across all settings. As the number of clients increases, HealSplit maintains stable, high accuracy, while DnC suffers significant degradation with noticeable variance. Similarly, when the proportion of adversaries rises, HealSplit shows only a mild decline in performance, in contrast to the sharp accuracy drop observed in DnC, highlighting its stronger resilience to adversarial participation.

Defense Generalization Across Data. Our fourth set of experiments evaluates HealSplit's performance under data heterogeneity. The results are presented in Fig. 5 and Fig. 6.

HealSplit demonstrates strong generalization, consistently outperforming all baselines under both IID and non-IID conditions. On the MNIST dataset, as the data becomes increasingly non-IID, HealSplit maintains stable accuracy above 85%, while baseline methods suffer significant performance degradation due to increased distribution heterogeneity. On the HAM dataset, which reflects a real-world distributional shift, HealSplit continues to perform robustly, further highlighting its effectiveness across diverse and challenging data environments.

Defense Generalization Across Model. Our fifth set of experiments evaluates HealSplit's generalization ability across different model architectures. The results are presented in Fig. 7.

Across all tested model architectures, HealSplit consistently outperforms the state-of-the-art baseline DnC, demonstrating strong robustness across varying network structures. Under challenging multi-vector attacks such as DP + SP, it maintains significantly higher accuracy, further underscoring its resilience to architectural variations.

Adaptive Attack. Our sixth set of experiments evaluates HealSplit's performance under an adaptive attack to assess its robustness. The results are presented in Fig. 8.

Specifically, the attacker minimizes the divergence in TAS between poisoned and clean smashed data to evade detection by the topology-aware malicious data detection. These stealthy anomalies subsequently propagate to downstream stages such as GAN training and consistency validation student training, ultimately compromising the overall system.

Although the presence of a more serious and more adaptive threat leads to a noticeable performance drop for HealSplit, the method remains highly competitive. It continues to outperform the strongest existing defenses across multiple datasets, demonstrating superior robustness and generalization even under intensified attack scenarios.

Conclusion

This paper presents HealSplit, the first comprehensive defense framework against diverse and sophisticated data poisoning attacks in SFL. Unlike prior defenses that address isolated attack types or require full model access, HealSplit mitigates a broad spectrum of poisoning attacks by securing the smashed data, which serves as the primary conduit for adversarial manipulation in SFL. It seamlessly integrates topology-aware detection, adversarial multi-teacher distillation, and generative recovery into a unified end-to-end system. Extensive experiments demonstrate that HealSplit consistently outperforms existing defenses in both robustness and effectiveness, offering a practical and generalizable solution to enhance the security of SFL.

Acknowledgments

This work was supported by the National Key R&D Program of China (2023YFA1009500).

References

Alber, D. A.; Yang, Z.; Alyakin, A.; Yang, E.; Rai, S.; Valliani, A. A.; Zhang, J.; Rosenbaum, G. R.; Amend-Thomas, A. K.; Kurland, D. B.; et al. 2025. Medical large language models are vulnerable to data-poisoning attacks. *Nature Medicine*, 1–9.

Alsaheel, A.; Nan, Y.; Ma, S.; Yu, L.; Walkup, G.; Celik, Z. B.; Zhang, X.; and Xu, D. 2021. {ATLAS}: A sequence-based learning approach for attack investigation. In *30th USENIX security symposium (USENIX security 21)*, 3005–3022.

Angarano, S.; Martini, M.; Navone, A.; and Chiaberge, M. 2024. Domain generalization for crop segmentation with standardized ensemble knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5450–5459.

Blanchard, P.; El Mhamdi, E. M.; Guerraoui, R.; and Stainer, J. 2017. Machine learning with adversaries: Byzantine tolerant gradient descent. *Advances in neural information processing systems*, 30.

Cao, X.; Fang, M.; Liu, J.; and Gong, N. Z. 2020. Fltrust: Byzantine-robust federated learning via trust bootstrapping. *arXiv preprint arXiv:2012.13995*.

Chai, J.; Jang, T.; and Wang, X. 2022. Fairness without demographics through knowledge distillation. *Advances in Neural Information Processing Systems*, 35: 19152–19164.

Chen, J.; and Zhang, A. 2022. Fedmsplit: Correlation-adaptive federated multi-task learning across multimodal split networks. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, 87–96.

Cui, C.; and Jia, C. 2024. Propagation tree is not deep: Adaptive graph contrastive learning approach for rumor detection. In *Proceedings of the AAAI Conference on artificial intelligence*, volume 38, 73–81.

Fang, M.; Cao, X.; Jia, J.; and Gong, N. 2020. Local model poisoning attacks to Byzantine-Robust federated learning. In *29th USENIX security symposium (USENIX Security 20)*, 1605–1622.

Gajbhiye, S.; Singh, P.; and Gupta, S. 2022. Data poisoning attack by label flipping on splitfed learning. In *International Conference on Recent Trends in Image Processing and Pattern Recognition*, 391–405. Springer.

Ganin, Y.; and Lempitsky, V. 2015. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, 1180–1189. PMLR.

Gasteiger, J.; Bojchevski, A.; and Günnemann, S. 2018. Predict then propagate: Graph neural networks meet personalized pagerank. *arXiv preprint arXiv:1810.05997*.

Goldblum, M.; Fowl, L.; Feizi, S.; and Goldstein, T. 2020. Adversarially robust distillation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 3996–4003.

Guerraoui, R.; Rouault, S.; et al. 2018. The hidden vulnerability of distributed learning in byzantium. In *International conference on machine learning*, 3521–3530. PMLR.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

He, P.; Xu, H.; Xing, Y.; Liu, H.; Yamada, M.; and Tang, J. 2024. Data poisoning for in-context learning. *arXiv preprint arXiv:2402.02160*.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Huang, N.; Gokaslan, A.; Kuleshov, V.; and Tompkin, J. 2024. The gan is dead; long live the gan! a modern gan baseline. *Advances in Neural Information Processing Systems*, 37: 44177–44215.

Ismail, A. T. Z.; and Shukla, R. M. 2023. Analyzing the vulnerabilities in splitfed learning: Assessing the robustness against data poisoning attacks. *arXiv preprint arXiv:2307.03197*.

Kairouz, P.; McMahan, H. B.; Avent, B.; Bellet, A.; Bennis, M.; Bhagoji, A. N.; Bonawitz, K.; Charles, Z.; Cormode, G.; Cummings, R.; et al. 2021. Advances and open problems in federated learning. *Foundations and trends® in machine learning*, 14(1–2): 1–210.

Karimireddy, S. P.; Kale, S.; Mohri, M.; Reddi, S.; Stich, S.; and Suresh, A. T. 2020. Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning*, 5132–5143. PMLR.

Khan, M. A.; Shejwalkar, V.; Houmansadr, A.; and Anwar, F. M. 2022. Security analysis of splitfed learning. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*, 987–993.

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

LeCun, Y.; Cortes, C.; and Burges, C. J. C. 1998. MNIST Handwritten Digit Database. Available: http://yann.lecun.com/exdb/mnist.

Li, J.; Feng, X.; Gu, T.; and Chang, L. 2024. Dual-Teacher De-biasing Distillation Framework for Multi-domain Fake News Detection. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*, 3627–3639. IEEE.

Lin, Z.; Zhu, G.; Deng, Y.; Chen, X.; Gao, Y.; Huang, K.; and Fang, Y. 2024. Efficient parallel split learning over resource-constrained wireless edge networks. *IEEE Transactions on Mobile Computing*, 23(10): 9224–9239.

Liu, Y.; Kang, Y.; Zou, T.; Pu, Y.; He, Y.; Ye, X.; Ouyang, Y.; Zhang, Y.-Q.; and Yang, Q. 2024. Vertical federated learning: Concepts, advances, and challenges. *IEEE Transactions on Knowledge and Data Engineering*, 36(7): 3615–3634.

Luo, C.; Lin, Q.; Xie, W.; Wu, B.; Xie, J.; and Shen, L. 2022. Frequency-driven imperceptible adversarial attack on semantic similarity. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 15315–15324.

Ma, Z.; Dong, J.; Ji, S.; Liu, Z.; Zhang, X.; Wang, Z.; He, S.; Qian, F.; Zhang, X.; and Yang, L. 2024. Let all be whitened: Multi-teacher distillation for efficient visual retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 4126–4135.

Ma, Z.; Ma, J.; Miao, Y.; Li, Y.; and Deng, R. H. 2022. ShieldFL: Mitigating model poisoning attacks in privacy-preserving federated learning. *IEEE Transactions on Information Forensics and Security*, 17: 1639–1654.

McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; and y Arcas, B. A. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, 1273–1282. PMLR.

Mu, X.; Cheng, K.; Shen, Y.; Li, X.; Chang, Z.; Zhang, T.; and Ma, X. 2024. Feddmc: Efficient and robust federated learning via detecting malicious clients. *IEEE Transactions on Dependable and Secure Computing*.

Panda, A.; Mahloujifar, S.; Bhagoji, A. N.; Chakraborty, S.; and Mittal, P. 2022. Sparsefed: Mitigating model poisoning attacks in federated learning with sparsification. In *International Conference on Artificial Intelligence and Statistics*, 7587–7624. PMLR.

Sauer, A.; Boesel, F.; Dockhorn, T.; Blattmann, A.; Esser, P.; and Rombach, R. 2024a. Fast high-resolution image synthesis with latent adversarial diffusion distillation. In *SIGGRAPH Asia 2024 Conference Papers*, 1–11.

Sauer, A.; Lorenz, D.; Blattmann, A.; and Rombach, R. 2024b. Adversarial diffusion distillation. In *European Conference on Computer Vision*, 87–103. Springer.

Shejwalkar, V.; and Houmansadr, A. 2021. Manipulating the byzantine: Optimizing model poisoning attacks and defenses for federated learning. In *NDSS*.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Singh, N. D.; Croce, F.; and Hein, M. 2023. Revisiting adversarial training for imagenet: Architectures, training and generalization across threat models. *Advances in Neural Information Processing Systems*, 36: 13931–13955.

Sun, Y.; Liu, Z.; Hooi, B.; Yang, Y.; Fathony, R.; Chen, J.; and He, B. 2025. Multi-Label Node Classification with Label Influence Propagation. In *The Thirteenth International Conference on Learning Representations*.

Thapa, C.; Arachchige, P. C. M.; Camtepe, S.; and Sun, L. 2022. Splitfed: When federated learning meets split learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 8485–8493.

Tolpegin, V.; Truex, S.; Gursoy, M. E.; and Liu, L. 2020. Data poisoning attacks against federated learning systems. In *Computer security–ESORICs 2020: 25th European symposium on research in computer security, ESORICs 2020, guildford, UK, September 14–18, 2020, proceedings, part i 25*, 480–501. Springer.

Tschandl, P.; Rosendahl, C.; and Kittler, H. 2018. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1): 1–9.

Vepakomma, P.; Gupta, O.; Swedish, T.; and Raskar, R. 2018. Split learning for health: Distributed deep learning without sharing raw patient data. *arXiv preprint arXiv:1812.00564*.

Wen, H.; Pan, L.; Dai, Y.; Qiu, H.; Wang, L.; Wu, Q.; and Li, H. 2024. Class incremental learning with multi-teacher distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 28443–28452.

Woodworth, B. E.; Patel, K. K.; and Srebro, N. 2020. Minibatch vs local sgd for heterogeneous distributed learning. *Advances in Neural Information Processing Systems*, 33: 6281–6292.

Wu, X.; Yuan, H.; Li, X.; Ni, J.; and Lu, R. 2024. Evaluating Security and Robustness for Split Federated Learning Against Poisoning Attacks. *IEEE Transactions on Information Forensics and Security*.

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

Yazdinejad, A.; Dehghantanha, A.; Karimipour, H.; Srivastava, G.; and Parizi, R. M. 2024. A robust privacy-preserving federated learning model against model poisoning attacks. *IEEE Transactions on Information Forensics and Security*.

Yin, D.; Chen, Y.; Kannan, R.; and Bartlett, P. 2018. Byzantine-robust distributed learning: Towards optimal statistical rates. In *International conference on machine learning*, 5650–5659. Pmlr.

Yu, T.; Kumar, S.; Gupta, A.; Levine, S.; Hausman, K.; and Finn, C. 2020. Gradient surgery for multi-task learning. *Advances in neural information processing systems*, 33: 5824–5836.

Zhao, S.; Wang, X.; and Wei, X. 2024. Mitigating accuracy-robustness trade-off via balanced multi-teacher adversarial distillation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhu, J.; Gao, C.; Yin, Z.; Li, X.; and Kurths, J. 2024. Propagation Structure-Aware Graph Transformer for Robust and Interpretable Fake News Detection. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 4652–4663.