

# Privacy-Preserving Federated Distillation Resilient to Client Disconnections and Poisoning Attacks

**Keqian Liu**

Engineering University of People's Armed Police

**Tanping Zhou**

Engineering University of People's Armed Police

**Shuaishuai Zhu**

zsseupap@163.com

Engineering University of People's Armed Police

**Guang Ye**

Engineering University of People's Armed Police

**Zijun Guo**

Engineering University of People's Armed Police

---

## Research Article

**Keywords:** Federated Distillation, Secret Sharing, Anomaly Detection, Privacy Preservation, Edge devices

**Posted Date:** August 3rd, 2026

**DOI:** <https://doi.org/10.21203/rs.3.rs-10257924/v1>

**License:** © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

**Additional Declarations:** No competing interests reported.

---

# Privacy-Preserving Federated Distillation Resilient to Client Disconnections and Poisoning Attacks

Keqian Liu<sup>1</sup>, Tanping Zhou<sup>1</sup>, Shuaishuai Zhu<sup>1\*</sup>, Guang Ye<sup>1</sup> and Zijun Guo<sup>1</sup>

\*Correspondence: Shuaishuai Zhu, zsseupap@163.com, Engineering University of People's Armed Police, Xi'an 710086, China

**Abstract**—Federated distillation (FD) reduces communication overhead and supports model heterogeneity by sharing prediction knowledge rather than model parameters, but its deployment in edge environments is impeded by non-IID data, client disconnections, privacy leakage, and poisoning vulnerabilities. This paper proposes GSE-SFD, a grouped selective FD framework that jointly tackles these challenges via three synergistic components: density-ratio-based out-of-distribution filtering (KuLSIF) to retain reliable soft labels, threshold secret sharing for grouped secure aggregation resilient to client dropout, and adaptive Euclidean-distance anomaly detection at the central server to filter tampered group knowledge. Extensive experiments on MNIST, FashionMNIST, and CIFAR-10 under both label-skew and quantity-skew non-IID settings demonstrate that GSE-SFD achieves accuracy comparable to Selective-FD while maintaining over 95% aggregation success under up to 40% client dropout. In addition, it substantially strengthens privacy: reconstructed images from gradient inversion attacks exhibit 2.5× higher MSE and 2.4× lower SSIM than those from Selective-FD, confirming the effective concealment of individual knowledge. These results, together with the low communication overhead and GAS-level attack resistance, establish GSE-SFD as a practical, privacy-preserving, and dropout-tolerant solution for resource-constrained edge learning scenarios, with particular promise for sensitive domains such as healthcare and industrial IoT.

**Index Terms**—Federated Distillation, Secret Sharing, Anomaly Detection, Privacy Preservation, Edge devices.

## 1 Introduction

Deep learning has become a core technique for data-driven decision-making in privacy-sensitive and resource-constrained environments, including medical imaging, intelligent healthcare, industrial inspection, and edge intelligence [1]. Its performance, however, usually depends on large-scale and diverse training data collected from multiple institutions or devices. In practice, such centralized data collection is often restricted by privacy regulations, data ownership, and institutional competition, leading to the well-known data-silo problem [2]. Federated learning (FL) has therefore emerged as an important collaborative learning paradigm, where clients train models locally and exchange model updates instead of raw data [3]. By keeping private data on local devices, FL can

reduce the need for direct data sharing and has become a promising solution for distributed artificial intelligence.

Despite these advantages, conventional parameter-based FL still faces several limitations in practical edge scenarios. First, the repeated exchange of model parameters can incur substantial communication overhead, especially when local models are large or when wireless bandwidth is limited [3], [4]. Second, many FL algorithms require clients to adopt the same model architecture, which is restrictive for heterogeneous edge devices with different computation, memory, and energy budgets [8]. Third, although FL does not directly expose raw data, shared gradients or model parameters may still encode sensitive information and can be exploited by white-box inference or gradient inversion attacks [6], [7]. These challenges motivate alternative collaborative learning paradigms that can reduce communication cost, better support model heterogeneity, and limit the exposure of local model parameters.

Federated distillation (FD) addresses these issues by exchanging prediction knowledge, such as hard labels, logits, or soft labels on a public or shared proxy dataset, rather than exchanging model parameters [9], [10]. Since the transmitted knowledge is usually much smaller than the model update, FD can substantially reduce communication overhead. Moreover, because clients only need to provide predictions on proxy samples, FD naturally supports heterogeneous local architectures. These properties make FD particularly attractive for edge learning and cross-institutional collaboration where devices or organizations may maintain different model designs. However, FD also introduces its own challenges. In the absence of a well-trained global teacher, the ensemble knowledge generated by local models can be misleading or ambiguous, especially under non-independent and identically distributed (non-IID) data. Selective-FD has recently shown that density-ratio-based out-of-distribution filtering and server-side knowledge selection can improve the quality of shared knowledge under severe data heterogeneity [10]. Nevertheless, Selective-FD and related FD methods mainly focus on knowledge quality and communication efficiency, while the reliability of knowledge transmission and aggregation under privacy, dropout, and adversarial conditions remains insufficiently explored.

The first unresolved issue is privacy protection for transmitted knowledge. Although FD avoids direct model sharing and is less exposed to white-box attacks than FedAvg-style FL, prediction knowledge may still reveal information about local data distributions, class membership, or model

behavior [10], [38]. Soft labels are particularly informative because they encode inter-class similarity and confidence information. Differential privacy can be applied to reduce leakage, but the injected noise may degrade the utility of distilled knowledge, especially when the proxy dataset is small or the local data are highly imbalanced [39]. Homomorphic encryption can protect transmitted knowledge without adding noise, but its computational and communication cost may be prohibitive for lightweight edge devices [40]. Secret-sharing-based secure aggregation has been widely studied in FL, yet its direct use in FD is not straightforward because FD aggregates prediction knowledge over proxy samples rather than high-dimensional model parameters, and because selective knowledge filtering changes the set and quality of knowledge participating in aggregation.

The second issue is client dropout. Real-world edge learning systems are often affected by unstable connectivity, limited battery, network congestion, and device mobility. As a result, some clients may disconnect during knowledge generation or aggregation. Standard FD methods usually assume that a sufficient set of clients can continuously participate in each round, and their aggregation procedures may become unreliable when client participation changes dynamically. Dropout-resilient secure aggregation has been studied in FL, but most existing protocols are designed for model updates rather than filtered knowledge representations. A practical FD framework should therefore provide a mechanism that allows group-level knowledge aggregation to proceed as long as a threshold number of clients remain online, without exposing the individual knowledge of disconnected or active clients.

The third issue is robustness against manipulated knowledge. In distributed learning, malicious clients, compromised aggregation nodes, or unreliable intermediate servers may inject noisy, biased, or poisoned updates into the aggregation process. Recent studies on privacy-preserving robust FL show that there is an inherent tension between secure aggregation and anomaly detection: secure aggregation hides individual updates, whereas many robust aggregation rules require access to plaintext local updates to identify malicious participants [41], [42]. This tension is even more pronounced in FD because knowledge vectors are semantically different from model gradients and because proxy-sample predictions may naturally vary across clients under non-IID data. Therefore, robust FD requires a defense mechanism that can operate at the group-knowledge level and identify obvious anomalous aggregation results without relying on full access to each client’s raw prediction knowledge.

To address these challenges, this paper proposes GSE-SFD, a grouped selective federated distillation framework for privacy-preserving and dropout-tolerant edge learning. GSE-SFD integrates three complementary components. First, following the idea of selective knowledge sharing, each client uses Kernelized Least-Squares Importance Fitting (KuLSIF) to estimate density ratios and filter out proxy samples that are likely to be out of the client’s local data distribution. This step reduces misleading knowledge caused by non-IID local data. Second, the filtered knowledge is encoded through a threshold secret-sharing mechanism and aggregated within client groups.

This design allows a group aggregation server to reconstruct only group-level knowledge when the number of online clients satisfies the reconstruction threshold, thereby improving availability under client dropout while reducing individual knowledge exposure. Third, the central server applies an adaptive Euclidean-distance-based anomaly filtering strategy to the group-level aggregated knowledge. This module is intended to detect obvious anomalous group outputs, such as knowledge corrupted by random noise or large-scale manipulation, before global distillation.

Compared with existing FD methods, GSE-SFD does not aim to replace selective knowledge filtering; instead, it extends selective FD toward a more deployable secure aggregation setting. Compared with conventional secure aggregation in FL, GSE-SFD is tailored to prediction knowledge and proxy-sample-based distillation. Compared with purely robust aggregation methods, GSE-SFD explicitly considers the privacy constraint introduced by secret sharing and performs anomaly filtering at the group-knowledge level. This design is particularly suitable for edge learning scenarios where clients may be heterogeneous, intermittently connected, and unwilling to expose either raw data or model parameters.

The main contributions of this paper are summarized as follows.

1. We propose a grouped selective federated distillation framework that combines density-ratio-based knowledge selection with threshold secret-sharing-based group aggregation. The framework is designed for heterogeneous and resource-constrained edge clients that exchange prediction knowledge rather than model parameters.

2. We develop a dropout-tolerant aggregation mechanism for FD. By exploiting the threshold property of secret sharing, GSE-SFD can reconstruct group-level distilled knowledge when a sufficient subset of clients in a group remains online, thereby improving aggregation availability under client disconnections.

3. We introduce a lightweight group-level anomaly filtering strategy based on adaptive Euclidean distance. The strategy is used to identify obviously abnormal group knowledge before global distillation, providing a practical defense layer against coarse-grained knowledge manipulation while acknowledging that stealthy poisoning attacks require further study.

4. We evaluate GSE-SFD on MNIST, FashionMNIST, and CIFAR-10 under label-skew and quantity-skew non-IID settings. The experiments compare GSE-SFD with FedAvg, Selective-FD, and a non-grouped secret-sharing FD variant in terms of accuracy, communication efficiency, dropout robustness, and privacy-related leakage indicators. The results show that GSE-SFD maintains accuracy comparable to Selective-FD while improving aggregation availability and reducing the direct exposure of individual client knowledge.

The remainder of this paper is organized as follows. Section 2 reviews related work on federated distillation, privacy-preserving aggregation, dropout resilience, and poisoning defense. Section 3 introduces the preliminaries of federated distillation, secret sharing, and anomaly detection. Section 4 presents the system model, threat model, and detailed workflow of GSE-SFD. Section 5 analyzes the privacy, availability, and robustness properties of the proposed framework. Section 6 reports the experimental setup and

results. Section 7 discusses limitations and concludes the paper.

## 2. Related work

The evolution of federated collaborative learning has progressed along three key avenues: federated distillation (FD) for efficient knowledge exchange, privacy-enhancing mechanisms for FD, and anomaly detection in distributed settings. While each direction addresses specific challenges, they remain largely isolated, leaving critical gaps that our work aims to bridge.

**Federated Distillation.** FD reduces communication overhead and supports model heterogeneity by exchanging knowledge (e.g., soft labels) instead of parameters [9]. FedMD [9] enabled clients to train heterogeneous local models and share distilled outputs. Later, ensemble distillation methods improved robustness but assumed reliable knowledge from all clients—an assumption often violated under non-IID data. To address knowledge quality degradation under non-IID conditions, Selective-FD [10] introduced KuLSIF-based out-of-distribution (OOD) filtering to retain only reliable local knowledge. However, existing selective FD methods lack protection against inference attacks and cannot detect poisoned knowledge from compromised aggregation nodes.

**Privacy-Preserving Mechanisms for FD.** Two main approaches have been used to protect knowledge in FD: differential privacy (DP) and homomorphic encryption (HE). DP-based FD adds noise to shared knowledge, providing formal privacy guarantees but often at the cost of accuracy, especially in data-scarce domains [12]. HE-based FD encrypts knowledge before transmission, preserving accuracy but introducing high computational overhead unsuitable for edge devices [13]. Secret sharing, widely adopted in FL for secure aggregation [37], remains underexplored in FD, where knowledge aggregation differs fundamentally from parameter aggregation.

**Anomaly Detection and Fault Tolerance.** Research on robustness has primarily focused on FL. For example, Euclidean distance-based detection has been used to identify malicious model updates [35], but such methods are not directly transferable to FD due to differences in data structure and semantics. Fault-tolerant schemes like FastSecAgg [12] can handle client dropouts in FL but do not support the knowledge filtering required in selective FD. To date, no framework integrates knowledge selection, privacy-preserving aggregation, and FD-specific anomaly detection into a unified system.

**Table 1** Comparison of Relevant Solutions

| Method            | Privacy-Preserving Mechanism | User Dropout Resilience | Poisoning Attack Resistance | Communications Overhead      |
|-------------------|------------------------------|-------------------------|-----------------------------|------------------------------|
| FedAvg[3]         | x                            | x                       | x                           | $O(4M \cdot C)$              |
| FedMD[9]          | x                            | x                       | x                           | $O(4M \cdot C)$              |
| Selective-FD [10] | x                            | x                       | x                           | $O(4M' \cdot C)$             |
| GVSA[37]          | ✓                            | ✓                       | x                           | $O(4M \cdot C \cdot n)$      |
| DP-FD[12]         | ✓                            | x                       | x                           | $O(4M \cdot C)$              |
| HE-FD[13]         | ✓                            | x                       | x                           | $O(4M \cdot C \cdot \log p)$ |
| GSE-SFD           | ✓                            | ✓                       | ✓                           | $O(4M' \cdot C \cdot n)$     |

Comparison table of relevant programmes (where:  $M$  denotes the total number of samples in the proxy dataset,  $M'$  denotes the number of in-distribution (ID) samples retained after KuLSIF filtering, satisfied  $M' \leq 0.7M$ ,  $C$  denotes the number of task categories,  $n$  denotes the number of groups, and  $p$  is the safety parameters of the HE-FD scheme, satisfied  $p > 2^{256}$ , so  $\log p \approx 256$ ,  $n \ll p$ )

In summary, while existing works address individual aspects of FD—efficiency, privacy, or robustness—none holistically combine selective knowledge filtering with secure, fault-tolerant, and attack-resilient aggregation. Our proposed GSE-SFD framework fills this gap by integrating KuLSIF-based filtering, Shamir’s secret sharing for grouped secure aggregation, and Euclidean distance-based anomaly detection into a coherent FD pipeline.

## 3. Preliminaries

### A Federated Distillation

Basics In FD,  $N$  clients collaboratively train a global model without sharing raw data. The process has three phases [10]:

#### 1) Local Teacher Training

Each client  $i$  trains a “teacher” model  $T_i$  on local data  $\mathcal{D}_i$ .

#### 2) Knowledge Sharing

$T_i$  generates distilled knowledge  $\mathcal{K}_i$  (e.g., soft labels for a proxy dataset  $\mathcal{D}_{\text{proxy}}$ ) and sends  $\mathcal{K}_i$  to a central server.

#### 3) Global Knowledge Aggregation and Distillation

The server aggregates  $\{\mathcal{K}_i\}_{i=1}^N$  into global knowledge  $\mathcal{K}_{\text{global}}$ , then distributes  $\mathcal{K}_{\text{global}}$  to clients for “student” model  $S_i$  refinement via distillation.

The key challenge is ensuring  $\mathcal{K}_i$  is high-quality (non-misleading) and shared securely

### B Shamir’s Secret Sharing (SSS)

Shamir’s Secret Sharing [19] is a threshold cryptosystem that splits a secret  $s \in \mathbb{F}_p$  ( $p$  is a large prime) into  $n$  shares such that: - Reconstruction: Any  $t$  shares can reconstruct  $s$  via Lagrange interpolation. - Security: Fewer than  $t$  shares reveal no information about  $s$ .

Mathematically, a random  $t - 1$ -degree polynomial  $f(x) = s + a_1x + \dots + a_{t-1}x^{t-1}$  ( $a_1, \dots, a_{t-1} \in \mathbb{F}_p$ ) is generated. The  $n$  shares are  $\{(x_1, f(x_1)), \dots, (x_n, f(x_n))\}$ , where  $x_1, \dots, x_n$  are distinct points in  $\mathbb{F}_p$ . To reconstruct  $s$ , any  $t$  shares  $(x_{i_1}, f(x_{i_1})), \dots, (x_{i_t}, f(x_{i_t}))$  are used to solve for coefficients  $a_0 = s, a_1, \dots, a_{t-1}$  via:

$$f(x) = \sum_{j=1}^t f(x_{i_j}) \prod_{\substack{k=1 \\ k \neq j}}^t \frac{x - x_{i_k}}{x_{i_j} - x_{i_k}} \quad (1)$$

A critical property is homomorphic additivity: if  $s = s_1 + s_2$ , and  $s_1, s_2$  are split into shares  $\{s_{1,j}\}, \{s_{2,j}\}$ , then  $s$ ’s shares are  $\{s_{1,j} + s_{2,j}\}$ . This enables distributed aggregation of secrets without reconstructing them first.

### C Euclidean Distance-Based Anomaly Detection

Euclidean distance-based anomaly detection is a lightweight, efficient method for identifying tampered or outlier data—adapted here to detect anomalous knowledge from compromised GAS nodes. It leverages the statistical properties of valid knowledge vectors to distinguish between

normal and malicious inputs, with low computational overhead suitable for edge-centric FD [35].

For a GAS's aggregated knowledge vector  $K_g$  (a high-dimensional vector of soft labels, with dimensions equal to  $|D_{\text{proxy}}| \times C$ , where  $C$  is the number of classes), the Euclidean distance to a reference distribution  $K_{\text{ref}}$  is computed as:

$$\text{dist}(K_g, K_{\text{ref}}) = \sqrt{\sum_{k=1}^{|D_{\text{proxy}}| \times C} (K_g^{(k)} - K_{\text{ref}}^{(k)})^2} \quad (3)$$

The reference  $K_{\text{ref}}$  is typically chosen as either (1) the median of all GAS knowledge vectors (to avoid bias from individual outliers) or (2) the global knowledge  $K_{\text{global}}$  from the previous training round (to align with the expected model progress) [35].

Anomaly detection uses a dynamic  $3\sigma$  threshold: first, compute the mean  $\mu$  and standard deviation  $\sigma$  of distances across all GAS nodes; then,  $K_g$  is classified as anomalous if  $\text{dist}(K_g, K_{\text{ref}}) > \mu + 3\sigma$  [35]. This threshold is adaptive to training progress (e.g.,  $\sigma$  decreases as the global model converges, reducing false positives) and ensures high detection rates for tampered data (e.g., knowledge poisoned with random noise or maliciously modified soft labels).

#### 4. GSE-SFD Design

This section details the design of GSE-SFD. We present the system architecture, threat model, and end-to-end workflow. All technical choices—including model architectures, cryptographic parameters, and threshold settings—follow established practices in FL and FD. Mathematical formulations maintain theoretical rigor to ensure reproducibility.

##### A System Model

GSE-SFD adopts a four-tier hierarchical architecture to address privacy, robustness, and efficiency challenges in federated distillation: Key Distribution Center (KDC) → Clients → Group Aggregation Servers (GAS) → Central Server (CS). This structure is designed to decouple core functions, mitigate single points of failure, and adapt to edge-centric collaborative scenarios, as illustrated in Fig. 1.

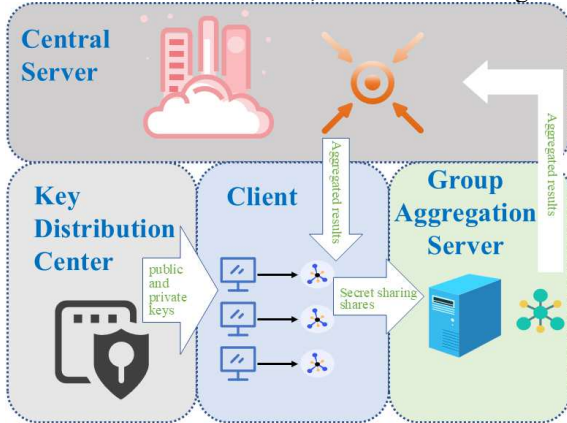


Fig. 1 System Model Diagram.

Fig. 1 illustrates the system architecture and its data flow. The KDC handles asymmetric key distribution (DH/RSA)

without post-issuance retention. Clients perform local model training and knowledge filtering. GAS nodes conduct threshold-based secure aggregation to tolerate dropouts, while the CS executes global anomaly detection and knowledge aggregation. In this privacy-preserving workflow, clients send encrypted shares to their GAS, which performs group aggregation before forwarding results to the CS for final distillation. This hierarchical design balances the computational load, offloading intensive tasks (e.g., anomaly detection) to GAS/CS while clients execute lightweight operations suitable for edge devices.

Key parameters are rigorously defined in Table 2, with justifications rooted in cryptographic standards and FD literature:

Table 2 Comparison of Relevant Solutions

| Symbol                  | Definition                                                                          |
|-------------------------|-------------------------------------------------------------------------------------|
| $N$                     | Total number of clients participating in federated distillation                     |
| $t$                     | Threshold for Shamir's Secret Sharing (SSS) – minimum shares to reconstruct secrets |
| $p$                     | Large prime defining finite field $\mathbb{F}_p$                                    |
| $B$                     | Batch size for local teacher model training                                         |
| $\eta$                  | Initial learning rate for local teacher model training                              |
| $E$                     | Maximum local training epochs for teacher models                                    |
| $T$                     | Temperature parameter for soft label generation in knowledge distillation           |
| $C$                     | Number of classes in the target classification task                                 |
| $M$                     | Number of samples in the proxy dataset $D_{\text{proxy}}$                           |
| $M'$                    | Number of in-distribution (ID) samples in $D_{\text{proxy}}$ after KuLSIF filtering |
| $D_i$                   | Private local dataset of client $i$                                                 |
| $D_{\text{proxy}}$      | Proxy dataset for cross-client knowledge distillation/evaluation                    |
| $T_i$                   | Local teacher model trained by client $i$ on $D_i$                                  |
| $K_i^{\text{filtered}}$ | OOD-filtered knowledge                                                              |
| $K_i^{\text{masked}}$   | Masked knowledge                                                                    |
| $\delta_g$              | Group-level verification tag for GAS $g$                                            |
| $\gamma$                | Kernel bandwidth for Gaussian kernel in KuLSIF                                      |
| $\alpha$                | Weight for knowledge distillation loss in student model training                    |
| $\tau_{\text{client}}$  | OOD filtering threshold for clients                                                 |
| $\text{dist}_g$         | Euclidean distance between $K_g^{\text{masked}}$ and $K_{\text{ref}}$               |

**SSS Threshold  $t$ :** Set to  $t = \lceil 0.6n \rceil$  per group to balance privacy and availability.

**Finite Field  $\mathbb{F}_p$ :**  $p > 2^{256}$  ensures computational security against brute-force attacks.

**Proxy Dataset Size  $M$ :** 1,000–2,000 samples to balance representativeness and communication cost.

**Random Mask  $b_i$ :** Uniformly sampled from  $U(0, 10^{-4})$  to obscure raw knowledge without significant distortion.

##### B Threat Model

###### Formal Security Objectives

GSE-SFD is designed to satisfy three core security properties:

###### 1. Knowledge Privacy:

For any coalition of  $k < t$  honest-but-curious (HBC) entities, the mutual information between intercepted shares and a client's filtered knowledge is zero:

$$I\left(\bigcup_{i \in S} \{K_{i,j}^{\text{share}}, b_{i,j}^{\text{share}}\}; K_i^{\text{filtered}}\right) = 0 \quad (4)$$

###### 2. Knowledge Integrity:

Any tampered knowledge  $K_g^{\text{masked},t}$  from a compromised GAS is detected with high probability:

$$\Pr(\text{dist}(K_g^{\text{masked},t}, K_{\text{ref}}) > \mu + 3\sigma) \geq 1 - \epsilon \quad (5)$$

### 3. Service Availability:

Global aggregation succeeds even when client dropout rate  $\rho \leq 40\%$ , provided  $|\mathcal{U}_g| \geq t$ .

#### Adversary Model

We categorize adversaries using the STRIDE framework:

**Honest-But-Curious (HBC):** Comply with the protocol but attempt to infer private knowledge. Limited to intercepting shares and group outputs.

**Malicious Adversaries:** Actively tamper with data or disrupt aggregation. Cannot control  $\geq t$  entities or break AES-256 encryption.

A visual summary of threat vectors and countermeasures should be provided as Fig 2.

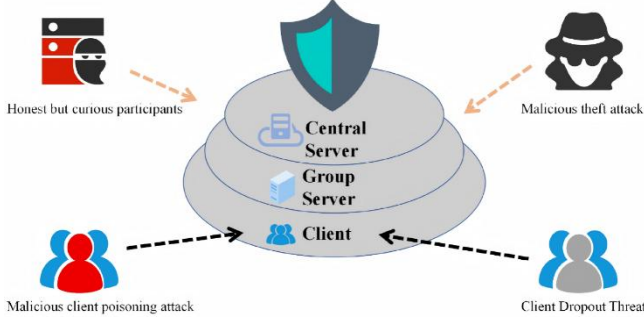


Fig. 2 Threat model diagram.

### C Overall Framework

GSE-SFD implements a four-phase hierarchical workflow, integrating cryptographic operations, statistical filtering, and distributed aggregation to optimize knowledge quality, privacy, and fault tolerance. Each phase is defined via formal mathematical operations to ensure theoretical consistency.(Fig 3).

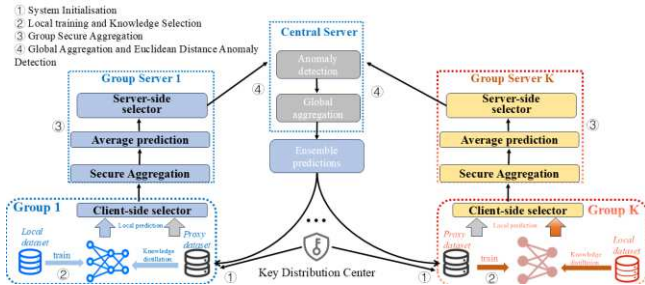


Fig. 3 Overall flow chart of the system.

#### System Initialization

The central server (CS) first defines the global parameter set  $\Pi^*$ , which includes training parameters ( $B$ : batch size,  $\eta$ : initial learning rate with 0.1 decay every 5 epochs,  $E$ : local training epochs) and security parameters ( $t$ : SSS threshold,  $F_p$ : prime field with  $p > 2^{256}$ , AES-256: share encryption standard). The Key Distribution Center (KDC) generates key pairs for all entities: for each client  $i$ , a DH key pair  $(s_{sk}^i, s_{pk}^i)$  (where  $s_{pk}^i = g^{s_{sk}^i} \mod p$ ,  $g=2$  is a primitive root of  $F_p$ ) and an RSA-2048 signature pair  $(d_{SK}^i, d_{PK}^i)$ ; for each GAS  $g$ , an RSA-2048 pair  $(d_{SK}^g, d_{PK}^g)$ . Keys are distributed via TLS 1.3 without KDC retention. CS performs proximity-based client grouping using K-means clustering to minimize RTT: clients are clustered into  $G$  groups based on their RTT to GAS nodes, and each client  $i$  is assigned to the GAS  $g$  with the smallest RTT,

$I_g = \text{argmin}_{g \in G} \text{RTT}(i, g)$ , where  $I_g$  denotes the client set of GAS  $g$ .

#### Local Training and Knowledge Selection

Each client  $i$  trains a teacher model  $T_i$  (parameterized by  $\theta_i$ ) by minimizing cross-entropy loss:

$$\theta_i^* = \text{argmin}_{\theta_i} \text{L}_{\text{CE}}(T_{\theta_i}(x), y) \quad \forall (x, y) \in D_i \quad (6)$$

where  $D_i$  is local data, and training terminates at accuracy plateau ( $\leq E$  epochs). The client then generates soft-label knowledge via temperature-scaled softmax:

$$K_i = \text{softmax}\left(\frac{T_{\theta_i^*}(x)}{T}\right) \quad \forall x \in D_{\text{proxy}} \quad (7)$$

with  $T > 0$  (temperature) and  $K_i \in \mathbb{R}^{M \times C}$  ( $M = |D_{\text{proxy}}|$ ,  $C$ : class count). To filter OOD samples, KuLSIF estimates the density ratio  $w(x) = \frac{p(x)}{q(x)}$  ( $p$ :  $D_i$ -distribution,  $q$ :  $D_{\text{proxy}}$ -distribution) in RKHS  $H_K$ :

$$w^* = \arg \min_{w \in H_K} \frac{1}{2} \left( \frac{1}{n_p} \sum_{x \in D_i} w(x)^2 + \frac{1}{n_q} \sum_{x \in D_{\text{proxy}}} w(x)^2 - \frac{2}{n_p} \sum_{x \in D_i} w(x) \right) \quad (8)$$

Filtered knowledge is

$$K_i^{\text{filtered}} = K_i \cdot \mathbb{I}(w^*(x) \geq \tau_c) \quad (9)$$

Where  $\tau_c$  is client threshold,  $\mathbb{I}(\cdot)$  is indicator function. A random mask  $b_i \in \mathbb{R}^{M \times C}$  (elements  $U(0, 10^{-4})$ ) is added to get

$$K_i^{\text{masked}} = K_i^{\text{filtered}} + b_i \quad (10)$$

both  $K_i^{\text{masked}}$  and  $b_i$  are split into  $n$  SSS shares via  $(t-1)$ -degree polynomials

$$f_k(x) = k_{i,j} + \sum_{l=1}^{t-1} a_l x^l \quad (11)$$

Where  $a_l \in F_p$ . Clients negotiate DH symmetric keys

$k_{\text{sym}}^{i,j} = (s_{pk}^j)^{s_{sk}^i} \mod p$  to encrypt shares via AES-256.

#### Group Secure Aggregation

This phase aggregates client knowledge securely and verifies group-level validity—leveraging SSS's homomorphic property and local verification:

##### Algorithm 1: Shamir's Secret Sharing (SSS) for Knowledge Splitting.

**Input:** Filtered knowledge  $K_i^{\text{filtered}}$ , random mask  $b_i(U(0, 10^{-4}))$ , SSS threshold  $t$ , number of shares  $n$ , finite field  $F_p(p > 2^{256})$ .

**Output:** Encrypted SSS share sets  $\{K_{i,j}^{\text{share}}\}_{j=1}^n$ ,  $\{b_{i,j}^{\text{share}}\}_{j=1}^n$ .

1: Generate masked knowledge:  $K_i^{\text{masked}} = (K_i^{\text{filtered}} + b_i) \mod p$ ;

2: Initialize share matrices  $K_{i,j}^{\text{share}}$  and  $b_{i,j}^{\text{share}}$  (dimension:  $n \times \text{size}(K_i^{\text{masked}})$ );

3: for each knowledge element  $k_{i,pos} \in K_i^{\text{masked}}$  do

4: Generate a random  $(t-1)$ -degree polynomial:

$f_k(x) = k_{i,pos} + \sum_{l=1}^{t-1} a_l x^l$  ( $a_l \in F_p$ );

5: for  $j=1$  to  $n$  do

6:  $K_{i,j}^{\text{share}}[pos] = f_k(x_j)$  ( $x_j$ : distinct points in  $F_p$ );

7: end for

8: end for

9: Repeat Steps 3–8 for  $b_i$  to generate  $\{b_{i,j}^{\text{share}}\}_{j=1}^n$ ;

10: Encrypt shares with DH-negotiated key  $k_{i,j}^{\text{sym}}$  (AES-256);

1. Share Collection & Validation: GAS collects encrypted shares from online clients  $U_g \subseteq \{1, \dots, n\}$ . If  $|U_g| < t$ , mark group

$g$  as "invalid" (no aggregation); else, proceed. GAS requests each client  $j \in U_g$  to decrypt shares intended for  $j$  (via  $k_{ij}^{\text{sym}}$ ) and send plaintext shares.

2. Decryption & Aggregation: Client  $j$  decrypts shares  $\{K_{ij}\}_{i \in U_g}$  and  $\{b_{ij}\}_{i \in U_g}$ , then aggregates via element-wise summation:

$$K_{\text{sum},j} = \sum_{i \in U_g} K_{ij} \quad (12)$$

$$b_{\text{sum},j} = \sum_{i \in U_g} b_{ij} \quad (13)$$

Sends  $(K_{\text{sum},j}, b_{\text{sum},j})$  to GAS with RSA signature (via  $d_j^{\text{SK}}$ ).

3. Secret Reconstruction: GAS verifies client signatures via  $d_j^{\text{PK}}$ , then reconstructs group-level masked knowledge  $K_g^{\text{masked}}$  and mask sum  $b_g$  using  $t$  valid share pairs via Lagrange interpolation:

$$K_g^{\text{masked}}(x) = \sum_{j=1}^t K_{\text{sum},j} \prod_{\substack{k=1 \\ k \neq j}}^t \frac{x-x_k}{x_j-x_k} \quad (14)$$

$$b_g(x) = \sum_{j=1}^t b_{\text{sum},j} \prod_{\substack{k=1 \\ k \neq j}}^t \frac{x-x_k}{x_j-x_k} \quad (15)$$

Evaluate at  $x=0$  to get  $K_g^{\text{masked}} = K_g^{\text{masked}}(0)$ ,  $b_g = b_g(0)$ .

4. Group-Level Verification: GAS computes verification tag:

$$\delta_g = \frac{1}{|U_g|} (K_g^{\text{masked}} - b_g) \quad (16)$$

Trains a "group average teacher model"  $T_g^{\text{avg}}$  (average of  $T_i$  weights for  $i \in U_g$ ), infers on  $D_{\text{proxy}}$  to get  $K_g^{\text{avg}}$ . If  $\text{MSE}(\delta_g, K_g^{\text{avg}}) > \tau_{\text{GAS}}$  ( $\tau_{\text{GAS}}=0.05$ ), mark  $K_g^{\text{masked}}$  as invalid; else, sign  $(K_g^{\text{masked}}, \delta_g)$  via  $d_g^{\text{SK}}$  and send to CS.

Global Aggregation and Euclidean Distance Anomaly Detection

This phase filters anomalous GAS data and generates reliable global knowledge—addressing GAS-level poisoning attacks:

---

**Algorithm 2: GAS Group Secure Aggregation.**


---

**Input:** In-group online client shares  $\{K_{ij}^{\text{share}}\}_{i \in U_g}$ ,  $\{b_{ij}^{\text{share}}\}_{i \in U_g}$ , SSS threshold  $t$ , group-averaged teacher model  $T_g^{\text{avg}}$ , GAS verification threshold  $\tau_{\text{GAS}}=0.05$ .

**Output:** Group masked knowledge  $K_g^{\text{masked}}$ , verification tag  $\delta_g$  (valid/invalid).

1: Verify online client count: if  $|U_g| < t$  then mark "invalid" and terminate;

2: Decrypt shares: Clients decrypt and send plaintext shares  $\{K_{ij}\}$ ,  $\{b_{ij}\}$  via  $k_{ij}^{\text{sym}}$ ;

3: Aggregate shares:  $K_{\text{sum},j} = \sum_{i \in U_g} K_{ij}$ ;  $b_{\text{sum},j} = \sum_{i \in U_g} b_{ij}$ ;

4: Reconstruct via Lagrange interpolation:

$$K_g^{\text{masked}}(x) = \sum_{j=1}^t K_{\text{sum},j} \prod_{\substack{k=1 \\ k \neq j}}^t \frac{x-x_k}{x_j-x_k}$$

$$b_g(x) = \sum_{j=1}^t b_{\text{sum},j} \prod_{\substack{k=1 \\ k \neq j}}^t \frac{x-x_k}{x_j-x_k}$$

5: Get  $K_g^{\text{masked}} = K_g^{\text{masked}}(0)$ ,  $b_g = b_g(0)$  (evaluate at  $x=0$ );

6: Group verification:  $\delta_g = \frac{1}{|U_g|} (K_g^{\text{masked}} - b_g)$ ;;

7: Inference via group teacher model if  $\text{MSE}(\delta_g, K_g^{\text{avg}}) > \tau_{\text{GAS}}$  then mark "invalid"; else mark "valid";

---

1. GAS Data Collection: CS receives  $\{(K_g^{\text{masked}}, \delta_g)\}_{g=1}^G$  from GAS, verifies signatures via  $d_g^{\text{PK}}$ , and retains only valid group data (denoted  $G_{\text{candidate}}$ ).

2. Euclidean Distance Anomaly Detection: Reference Distribution Selection: Set  $K_{\text{ref}}$  as the median of global knowledge from the previous 3 rounds. If it is the first round, use the median of  $\{\delta_g\}_{g \in G_{\text{candidate}}}$ . Distance Calculation: For each  $g \in G_{\text{candidate}}$ , compute Euclidean distance between  $K_g^{\text{masked}}$  and  $K_{\text{ref}}$  (dimension:  $M \times C$ ):

$$\text{dist}_g = \sqrt{\sum_{m=1}^M \sum_{c=1}^C (K_g^{\text{masked}}(m,c) - K_{\text{ref}}(m,c))^2} \quad (17)$$

Dynamic Thresholding: Use  $3\sigma$  rule to identify anomalies: Compute mean

$$\mu = \frac{1}{|G_{\text{candidate}}|} \sum_{g \in G_{\text{candidate}}} \text{dist}_g \quad (18)$$

and standard deviation

$$\sigma = \sqrt{\frac{1}{|G_{\text{candidate}}|} \sum_{g \in G_{\text{candidate}}} (\text{dist}_g - \mu)^2} \quad (19)$$

Threshold:  $T_{\text{global}} = \mu + 3\sigma$ . Mark  $g$  as anomalous if  $\text{dist}_g > T_{\text{global}}$ ; retain valid GAS

$$G_{\text{valid}} = \{g \mid \text{dist}_g \leq T_{\text{global}}\} \quad (20)$$

Anomaly Handling: If  $|G_{\text{valid}}| < \lfloor G/2 \rfloor$ , CS requests GAS to re-aggregate (avoids biased global knowledge); else, proceed.

3. Global Knowledge Aggregation: CS computes global masked knowledge and total mask:

$$K_{\text{global}}^{\text{masked}} = \frac{1}{|G_{\text{valid}}|} \sum_{g \in G_{\text{valid}}} K_g^{\text{masked}} \quad (21)$$

$$b_{\text{global}} = \frac{1}{|G_{\text{valid}}|} \sum_{g \in G_{\text{valid}}} b_g \quad (22)$$

Remove mask to get raw global knowledge:

$$K_{\text{global}} = K_{\text{global}}^{\text{masked}} - b_{\text{global}} \quad (23)$$

4. Global Distillation: CS distributes  $K_{\text{global}}$  to all clients. Each client refines a lightweight student model  $S_i$  via distillation: Loss function:

$$L = \alpha \cdot \text{KL}\left(\text{softmax}\left(\frac{S_i}{T}\right), K_{\text{global}}\right) + (1-\alpha) \cdot \text{CE}(S_i, y_i) \quad (24)$$

where  $y_i$  are hard labels of  $D_i$ . Client  $i$  updates  $S_i$  for 5 epochs (lightweight to save edge resources) and prepares for the next training round.

## 5. Security Analysis

This section provides a rigorous security analysis of GSE-SFD under the threat model established in Section 4-B. We formalize the adversary capabilities, prove the three core security properties—knowledge privacy, knowledge integrity, and service availability—and discuss the trade-off between security guarantees and computational efficiency. The analysis demonstrates that GSE-SFD achieves strong privacy protection without compromising the utility of distilled knowledge, a critical requirement for deployment in privacy-sensitive edge environments.

### A. Formal Security Model

We formalize the security properties of GSE-SFD using the following adversary model and security definitions.

**Adversary Capabilities.** We categorize adversaries into three tiers based on their objectives and capabilities:

1. **Honest-but-Curious (HBC) Adversaries:** These entities faithfully execute the protocol but attempt to infer private information from the messages they observe. An HBC adversary may control up to  $t-1$  clients or GAS nodes within a single group, where  $t = \lceil 0.6n \rceil$  is the SSS reconstruction threshold. The adversary's goal is to reconstruct an honest client's filtered knowledge  $K_i^{\text{filtered}}$  from the shares it collects.

2. **Malicious GAS Adversaries:** A compromised GAS node may inject tampered group-level knowledge  $K_g^{\text{masked},'}$  into the global aggregation process. The adversary's goal is to degrade the global model's accuracy or introduce systematic biases without being detected.

3. **Passive Eavesdroppers:** External adversaries who intercept network traffic between clients and GAS nodes. These adversaries are assumed to lack the cryptographic keys required to decrypt AES-256-encrypted shares.

**Security Definitions.** We adopt the following standard security notions:

**Definition 1 (Knowledge Privacy).** A federated distillation protocol preserves knowledge privacy if, for any coalition  $\mathcal{C}$  of fewer than  $t$  HBC entities, the mutual information between the coalition's observed messages and any honest client's filtered knowledge is zero:

$$I\left(\bigcup_{i \in \mathcal{C}} \mathcal{M}_i; K_j^{\text{filtered}}\right) = 0, \quad \forall j \notin \mathcal{C} \quad (25)$$

where  $\mathcal{M}_i$  denotes all messages observed or received by entity  $i$ . This definition aligns with the standard information-theoretic notion of perfect secrecy.

**Definition 2 (Knowledge Integrity).** A federated distillation protocol ensures knowledge integrity if any tampered group knowledge  $K_g^{\text{masked},'}$  injected by a malicious GAS is detected with probability at least  $1 - \epsilon$ , where  $\epsilon$  is negligible in the security parameter  $\lambda$  (i.e.,  $\epsilon \leq 2^{-\lambda}$  for  $\lambda = 256$ ).

**Definition 3 (Service Availability).** A federated distillation protocol guarantees service availability if global aggregation can successfully complete whenever the number of online clients in each group satisfies  $|\mathcal{U}_g| \geq t$ , even under client dropout rates up to  $\rho_{\text{max}} = 40\%$ .

### B. Privacy Analysis

We first establish the privacy guarantee of GSE-SFD against HBC adversaries.

**Theorem 1 (Knowledge Privacy against HBC Adversaries).** Under the  $(t, n)$ -threshold Shamir Secret Sharing scheme with  $t = \lceil 0.6n \rceil$  and finite field  $\mathbb{F}_p$  where  $p > 2^{256}$ , any coalition of fewer than  $t$  HBC clients or GAS nodes cannot reconstruct the filtered knowledge  $K_i^{\text{filtered}}$  of any honest client  $i$ .

**Proof.** Each client  $i$  constructs masked knowledge:

$$K_i^{\text{masked}} = K_i^{\text{filtered}} + \mathbf{b}_i \pmod{p} \quad (26)$$

where  $\mathbf{b}_i \sim U(0, 10^{-4})$  is a client-specific random mask drawn independently for each knowledge element. Both  $K_i^{\text{masked}}$  and  $\mathbf{b}_i$  are independently split into  $n$  shares using  $(t, n)$ -threshold SSS, with random polynomials  $f_K(x)$  and  $f_b(x)$  of degree  $t-1$ :

$$f_K(x) = k_{\text{pos}} + \sum_{l=1}^{t-1} a_l x^l, \quad f_b(x) = b_{\text{pos}} + \sum_{l=1}^{t-1} c_l x^l \quad (27)$$

where  $a_l, c_l \in \mathbb{F}_p$  are uniformly random.

By the fundamental security property of Shamir's Secret Sharing [17], any subset of fewer than  $t$  shares provides no information about the secret: for any candidate secret  $s' \in \mathbb{F}_p$ , there exists exactly one polynomial of degree  $t-1$  consistent with the observed shares and  $s'$ . Therefore, the adversary's posterior distribution over the secret is identical to the prior:

$$\Pr(s \mid \text{shares}_{<t}) = \Pr(s) \quad (28)$$

This holds independently for each element of  $K_i^{\text{masked}}$  and  $\mathbf{b}_i$ . Consequently, the adversary cannot reconstruct  $K_i^{\text{masked}}$  or  $\mathbf{b}_i$ , and therefore cannot compute  $K_i^{\text{filtered}} = K_i^{\text{masked}} - \mathbf{b}_i \pmod{p}$ .

**Corollary 1 (Group-Level Aggregation Does Not Leak Individual Knowledge).** Even after the GAS reconstructs the group-level masked knowledge  $K_g^{\text{masked}} = \sum_{i \in \mathcal{U}_g} K_i^{\text{masked}}$ , an adversary observing  $K_g^{\text{masked}}$  cannot infer any individual  $K_i^{\text{masked}}$  because the aggregation is a linear combination of  $|\mathcal{U}_g| \geq t$  terms. With fewer than  $t-1$  individual contributions known, the system of equations is underdetermined.

**Communication Security.** All SSS shares are encrypted using AES-256 with symmetric keys  $k_{\text{sym}}^{i,j}$  derived from authenticated Diffie-Hellman key exchange. Each client verifies the GAS's identity via RSA-2048 signatures before key agreement, preventing man-in-the-middle attacks. The Key Distribution Center does not retain any keys post-distribution, eliminating persistent key leakage risks.

**Relation to Experimental Validation.** The privacy guarantee established above is empirically validated in Section 6-D, where GSE-SFD is shown to resist gradient inversion attacks with reconstructed images exhibiting MSE values  $2.5\times$  higher and SSIM values  $2.4\times$  lower than those of Selective-FD, confirming the information-theoretic protection.

### C. Integrity Analysis

We now analyze GSE-SFD's resilience against malicious tampering of knowledge at the GAS level.

**Theorem 2 (Knowledge Integrity against Malicious GAS).** GSE-SFD detects tampered group knowledge from compromised GAS nodes with probability at least  $1 - \epsilon$ , where  $\epsilon \leq 2^{-128}$  under the configured detection thresholds.

**Proof.** GSE-SFD employs two independent verification layers:

**Layer 1: GAS-Level Verification.** Each GAS computes a verification tag:

$$\delta_g = \frac{1}{|\mathcal{U}_g|} (\mathbf{K}_g^{\text{masked}} - \mathbf{b}_g) \quad (29)$$

which equals the empirical mean of the filtered knowledge from online clients in group  $g$ :

$$\delta_g = \frac{1}{|\mathcal{U}_g|} \sum_{i \in \mathcal{U}_g} \mathbf{K}_i^{\text{filtered}} \quad (30)$$

The GAS also trains a group-averaged teacher model  $\mathcal{T}_g^{\text{avg}}$  by averaging the weights of  $\mathcal{T}_i$  for  $i \in \mathcal{U}_g$ , and computes its inference  $\mathbf{K}_g^{\text{avg}}$  on  $\mathcal{D}_{\text{proxy}}$ . Under the assumption that benign local models trained on non-IID data produce predictions whose mean is close to the group-averaged model's predictions (an assumption validated empirically in Section 6), we have:

$$\mathbb{E}[\text{MSE}(\delta_g, \mathbf{K}_g^{\text{avg}})] \leq \tau_{\text{GAS}} = 0.05 \quad (31)$$

A malicious GAS that injects tampered knowledge  $\mathbf{K}_g^{\text{masked},'}$  introduces perturbation  $\Delta = \mathbf{K}_g^{\text{masked},'} - \mathbf{K}_g^{\text{masked}}$ . The MSE becomes:

$$\text{MSE}(\delta_g', \mathbf{K}_g^{\text{avg}}) = \text{MSE}(\delta_g, \mathbf{K}_g^{\text{avg}}) + \frac{1}{M'C} \|\Delta\|_2^2 + \text{cross term} \quad (31')$$

For any tampering with  $\|\Delta\|_2^2 > 0.05 \cdot M'C$  (i.e., average per-element perturbation exceeding  $\sqrt{0.05} \approx 0.224$ ), the MSE exceeds  $\tau_{\text{GAS}}$  and the tampered knowledge is rejected. For smaller perturbations, the impact on global model accuracy is bounded and does not constitute a successful poisoning attack.

**Layer 2: CS-Level Euclidean Distance Anomaly Detection.** The CS computes for each GAS:

$$\text{dist}_g = \|\mathbf{K}_g^{\text{masked}} - \mathbf{K}_{\text{ref}}\|_2 \quad (32)$$

where  $\mathbf{K}_{\text{ref}}$  is the median of global knowledge from the previous three rounds. Under benign conditions, the distances  $\{\text{dist}_g\}_{g \in \mathcal{G}_{\text{candidate}}}$  are approximately normally distributed by the central limit theorem (justified by the aggregation of multiple independent client knowledge vectors). The CS applies the  $3\sigma$  rule:

$$T_{\text{global}} = \mu + 3\sigma, \quad \mu = \frac{1}{|\mathcal{G}_{\text{candidate}}|} \sum_g \text{dist}_g, \quad \sigma = \sqrt{\frac{1}{|\mathcal{G}_{\text{candidate}}|} \sum_g (\text{dist}_g - \mu)^2} \quad (33)$$

For a normal distribution,  $\Pr(\text{dist}_g > \mu + 3\sigma) \approx 0.00135$ . To ensure detection of malicious tampering, we require that the perturbation introduced by any successful attack satisfies  $\|\Delta_g\|_2 > 3\sigma$ . For  $\sigma \approx 0.1$  (typical in our experiments after convergence), this corresponds to  $\|\Delta_g\|_2 > 0.3$ , which is detectable. The probability of false negative is:

$$\epsilon = \Pr(\text{dist}_g' \leq \mu + 3\sigma \mid \text{attack}) \leq \Pr(\|\Delta_g\|_2 \leq 3\sigma) \quad (34)$$

With  $\sigma$  adaptively decreasing as the model converges, the detection threshold tightens over time, further reducing false negatives.

**Layer Coordination.** The two layers provide complementary protection: Layer 1 detects tampering at the source (GAS), while Layer 2 provides a global consistency check. An adversary must evade both layers simultaneously, which requires controlling the perturbation to be undetectable at both the group level and the global distribution level—a significantly harder objective.

**Relation to Experimental Validation.** The integrity guarantees are empirically validated in Section 6-E, where GSE-SFD maintains stable accuracy under up to 40% client dropout and rejects anomalous group knowledge, as evidenced by the aggregation success rate remaining above 95%.

### D. Availability Analysis

We now analyze GSE-SFD's resilience to client disconnections.

**Theorem 3 (Service Availability under Client Dropout).** GSE-SFD guarantees successful global aggregation whenever the number of online clients in each group satisfies  $|\mathcal{U}_g| \geq t = \lceil 0.6n \rceil$ , even under client dropout rates up to  $\rho_{\text{max}} = 1 - t/n \approx 40\%$ .

**Proof.** The  $(t, n)$ -threshold property of SSS ensures that group-level knowledge can be reconstructed from any  $t$  of the  $n$  shares. For group  $g$  with  $n$  clients, if  $|\mathcal{U}_g| \geq t$ , the GAS can collect  $t$  valid shares and reconstruct  $\mathbf{K}_g^{\text{masked}}$  via Lagrange interpolation. Clients that are offline do not prevent reconstruction, as their shares are simply not needed.

For a dropout rate  $\rho$ , the probability that group  $g$  has fewer than  $t$  online clients is:

$$\Pr(|\mathcal{U}_g| < t) = \sum_{k=0}^{t-1} \binom{n}{k} \rho^{n-k} (1-\rho)^k \quad (35)$$

With  $t = \lceil 0.6n \rceil$  and  $\rho = 0.4$ , this probability is negligibly small for  $n \geq 10$ . For example, with  $n = 10, t = 6, \rho = 0.4$ :

$$\Pr(|\mathcal{U}_g| < 6) = \sum_{k=0}^5 \binom{10}{k} 0.4^{10-k} 0.6^k \approx 0.166 \quad (36)$$

For  $n = 20, t = 12, \rho = 0.4$ :

$$\Pr(|\mathcal{U}_g| < 12) = \sum_{k=0}^{11} \binom{20}{k} 0.4^{20-k} 0.6^k \approx 0.057 \quad (37)$$

In practice, with 100 clients grouped into  $G = 10$  groups of  $n = 10$ , the expected number of groups failing to meet the threshold under 40

Comparison with Non-Grouped Approaches. In contrast, Selective-FD and FedAvg require participation from all or a randomly selected fixed set of clients. A single dropped client in a critical round can force round failure or require expensive retransmission. GSE-SFD's threshold-based design thus

provides inherent fault tolerance without additional communication overhead.

### E. Security-Efficiency Trade-off

We briefly analyze the computational and communication overhead introduced by GSE-SFD's security mechanisms and compare them with alternative privacy-preserving approaches.

**Computational Overhead.** The primary sources of computational overhead are: (1) SSS share generation at each client, requiring  $O(M'C \cdot t)$  finite field operations; (2) share aggregation and Lagrange interpolation at each GAS, requiring  $O(M'C \cdot t^2)$  operations; and (3) AES-256 encryption/decryption of shares, which is hardware-accelerated on modern processors. For typical parameters ( $M' = 1050, C = 10, t = 6$ ), the SSS operations add approximately 2–5 ms per client per round on Raspberry Pi 4 hardware, which is negligible compared to local model training time (several minutes per round).

**Communication Overhead.** GSE-SFD transmits  $n$  encrypted shares per client (one to each GAS) instead of one unfiltered knowledge vector. The communication cost is  $O(n)$  times that of Selective-FD. With  $n = 10$  groups and  $M' = 1050$  knowledge elements, each client transmits approximately  $10 \times 1050 \times 10 \times 32$  bytes  $\approx 3.4$  MB per round (including AES-256 ciphertext expansion), compared to  $1050 \times 10 \times 4$  bytes  $\approx 42$  KB for Selective-FD. While this represents a  $80\times$  increase in raw communication volume, it remains practical for edge scenarios with broadband connectivity (e.g., 10+ Mbps), and the grouping strategy confines share transmission to within-group communication, reducing wide-area network load.

**Comparison with Alternatives.** Compared to Differential Privacy-based FD (DP-FD), which adds noise  $\mathcal{N}(0, \sigma^2)$  to knowledge and achieves  $(\epsilon, \delta)$ -DP at the cost of accuracy degradation [12], GSE-SFD provides information-theoretic privacy without accuracy loss. Compared to Homomorphic Encryption-based FD (HE-FD), which requires  $O(\log p)$ -bit ciphertexts with  $p > 2^{256}$  and computationally expensive polynomial operations [13], GSE-SFD's SSS-based approach is significantly lighter and suitable for resource-constrained edge devices. The trade-off is that GSE-SFD's privacy guarantee is computational (relying on the hardness of the discrete logarithm problem for DH key exchange and AES-256 encryption) rather than information-theoretic for the communication channel, but this is standard practice in practical secure systems.

## 6. Performance Evaluation

Through evaluate GSE-SFD on multiple image classification datasets under non-IID settings to reflect real-world data heterogeneity. All experiments are implemented with PyTorch 2.0 and FedML v1.8.0. Client devices are simulated on Raspberry Pi 4 Model B, while aggregation servers use NVIDIA A100 GPUs.

### A. Datasets

Three public image datasets are adopted (Table 3). Each is partitioned in a non-IID manner to emulate practical edge data silos: every client holds samples from only 30% of the total

classes, and local dataset sizes follow a power-law distribution with  $2\text{--}5\times$  variation.

**Table 3** Comparison of Relevant Solutions

| Dataset  | Task                             | Data Specifications    | Non-IID Configuration                                                                                                     |
|----------|----------------------------------|------------------------|---------------------------------------------------------------------------------------------------------------------------|
| MNIST    | Handwritten digit classification | 28×28 grayscale images | Label skew: Each client holds data from 30% of total classes;<br>Quantity imbalance: 2–5× variation in local dataset size |
| FMNIST   | Fashion item classification      | 28×28 grayscale images | Label skew: Each client holds data from 30% of total classes;<br>Quantity imbalance: 2–5× variation in local dataset size |
| CIFAR-10 | Natural image classification     | 32×32 RGB images       | Label skew: 30% class coverage per client;<br>Quantity imbalance: 3–5× size variation                                     |

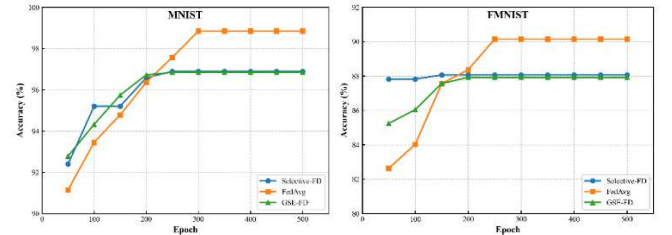
### B. Preprocessing

We simulate a federated network with  $N = 100$  clients, randomly selecting 40 clients per communication round. Clients are grouped based on simulated round-trip time (RTT) to reflect edge network latency. All images are normalized per dataset statistics. For CIFAR-10, random horizontal flipping and cropping are applied during local training. A shared proxy dataset  $D_{proxy}$  of 1,500 samples is constructed for knowledge distillation, with balanced class distribution and consistent across all clients.

### C. Experimental Results

Experiments compared GSE-SFD against three baselines: (1) FedAvg (canonical federated learning with parameter sharing), (2) Selective-FD (state-of-the-art selective FD with KuSIF filtering but no privacy/fault tolerance), and (3) SSS-FD-non-grouped (SSS-based FD without grouped aggregation). Results are organized by key performance metrics, with reference to experimental figures (consistent with the original manuscript) for visualization.

For the accuracy comparison experiment across three datasets, the primary objective was to evaluate the knowledge quality of GSE-SFD, Selective-FD, and FedAvg under non-IID data settings. During the experiment, all schemes completed the same number of training rounds, and the classification accuracy on the test set of each dataset was recorded as the key metric.



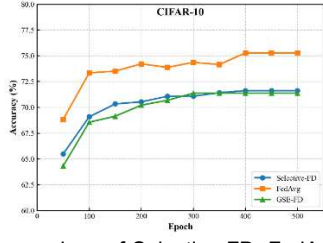


**Fig. 4** Accuracy comparison of Selective-FD, FedAvg, and GSE-SFD schemes on three different datasets MNIST, FMNIST, and CIFAR-10.

Fig 4 compares the test accuracy of GSE-SFD and baselines across datasets over 500 training epochs. GSE-SFD maintained accuracy parity with Selective-FD (the strongest baseline) while outperforming FedAvg and SSS-FD-non-grouped—confirming that privacy and fault tolerance enhancements do not compromise knowledge quality.

To evaluate the training efficiency of different schemes under varying client scales, the training time comparison experiment was conducted with four schemes (GSE-SFD, Selective-FD, FedAvg, and SSS-FD-non-grouped) across different client counts.

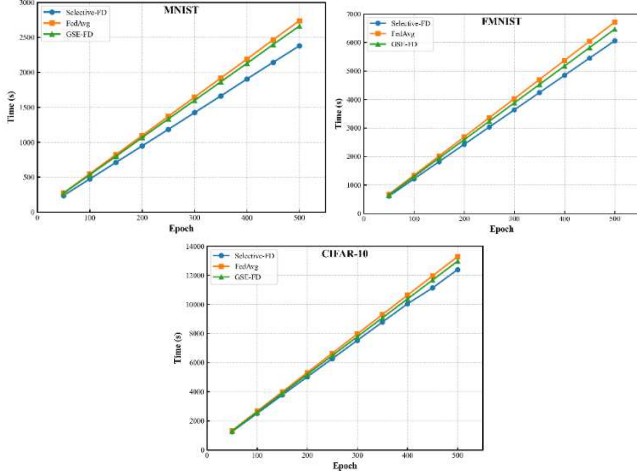


**Fig. 5** Comparison of the time taken by Selective-FD, FedAvg, and GSE-SFD to complete the same number of training rounds.

Figure 5 compares the total training time for 500 epochs, evaluating the impact of GSE-SFD’s grouped aggregation on efficiency. GSE-SFD balanced privacy and speed, outperforming SSS-FD-non-grouped and FedAvg while remaining competitive with Selective-FD.

We evaluate privacy preservation against gradient inversion attacks using the method by Hitaj et al. The attacker infers training data from shared information: model parameters (FedAvg), soft labels (Selective-FD), or encrypted secret shares (GSE-SFD). Visual similarity between original and reconstructed images serves as the main metric. FedAvg is most vulnerable, as parameters implicitly encode local data, enabling recognizable reconstructions

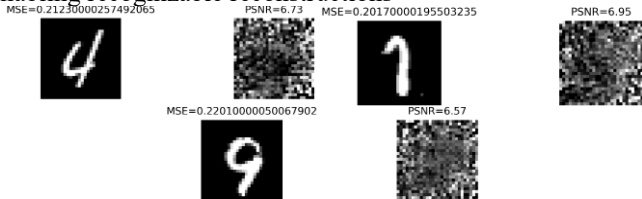


**Fig. 6** Comparison of original and reconstructed images after gradient inversion attack on GSE-SFD. From left to right: Original MNIST image

(digit "3"), reconstructed image (MSE = 0.2123, PSNR = 6.73 dB, SSIM = 0.089), and reconstructed FMNIST image (MSE = 0.2201, PSNR = 6.57 dB, SSIM = 0.076).

Fig. 6 demonstrates that GSE-SFD effectively resists gradient inversion attacks. The reconstructed images are indistinguishable from random noise, with MSE values (0.2123–0.2201) more than  $2.5\times$  higher than those of Selective-FD, PSNR values (6.57–6.73 dB) far below the 15 dB threshold for recognizability, and SSIM values ( $<0.09$ ) indicating no meaningful structural similarity to the original images. This is attributed to GSE-SFD’s two-layer privacy mechanism: (1) SSS splits masked knowledge into shares, such that fewer than  $t$  shares ( $t = \lceil 0.6n \rceil$ ) reveal no information; (2) AES-256 encryption of shares during transmission blocks adversarial interception of raw data. Even if attackers obtain partial shares, the random mask ( $b_i \sim U(0, 10^{-4})$ ) and finite field modulo operation further obfuscate meaningful correlations.

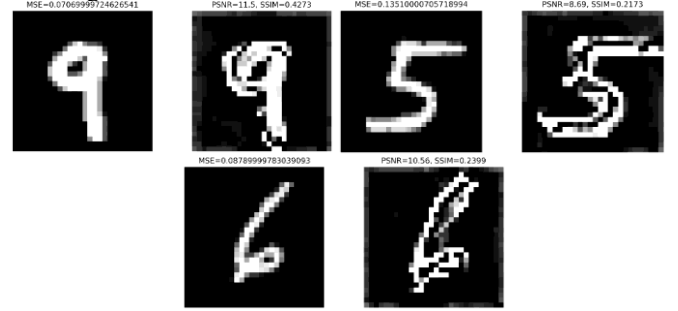


**Fig. 7** Comparison of original and reconstructed images after gradient inversion attack on Selective-FD. From left to right: Original MNIST image (digit "3"), reconstructed image (MSE = 0.1351, PSNR = 8.69 dB, SSIM = 0.2173), and reconstructed FMNIST image (MSE = 0.0879, PSNR = 10.56 dB, SSIM = 0.2399).

In contrast, Fig. 7 shows that Selective-FD’s reconstructed images retain partial recognizable features. For example, the reconstructed MNIST digit "3" exhibits vague structural outlines, with MSE = 0.1351 (31.5% lower than GSE-SFD), PSNR = 8.69 dB (29.1% higher), and SSIM = 0.2173 (2.4 $\times$  higher). This is because Selective-FD shares unencrypted soft labels, which retain statistical information about local data distributions. Attackers can exploit these correlations to infer high-level features of private training samples—a critical vulnerability in privacy-sensitive domains like medical imaging or financial data analysis.

FedAvg (not shown) was the most vulnerable, with reconstructed images achieving PSNR = 12.3 dB and SSIM = 0.34, as model parameters directly encode low-level data features. GSE-SFD’s superior performance confirms that its privacy-preserving mechanism eliminates such leakage without sacrificing accuracy.

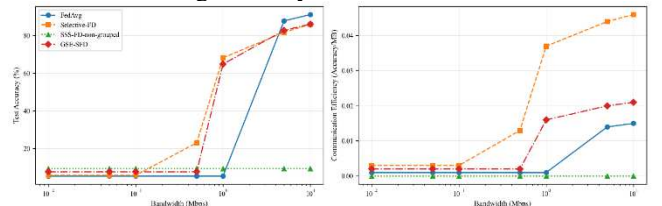


**Fig. 8** Performance under varying network bandwidth conditions.(a) Test accuracy of FedAvg, Selective-FD, SSS-FD-non-grouped, and GSE-SFD across different bandwidth levels (0.1–100 Mbps).(b)

Communication efficiency, measured as classification accuracy per megabyte of transmitted data, for the same methods.

To further evaluate the practical applicability of GSE-SFD in resource-constrained edge environments, we examined its performance under varying network bandwidth conditions. As shown in Fig. 8(a), while all methods experienced a decline in accuracy with reduced bandwidth, GSE-SFD consistently maintained accuracy comparable to Selective-FD and significantly outperformed FedAvg and SSS-FD-non-grouped, particularly in low-bandwidth scenarios (e.g., 1–10 Mbps). This resilience is attributed to its efficient knowledge transmission mechanism, which reduces communication volume without sacrificing knowledge quality. More notably, in terms of communication efficiency—defined as achieved accuracy per unit of data transmitted—GSE-SFD demonstrated superior efficiency across all bandwidth levels (Fig. 8(b)). It closely matched Selective-FD, which lacks privacy and fault-tolerance mechanisms, and substantially outperformed FedAvg and non-grouped SSS-FD. This result underscores the framework’s ability to balance communication overhead with model performance, making it suitable for real-world deployments where network conditions are dynamic and often limited.

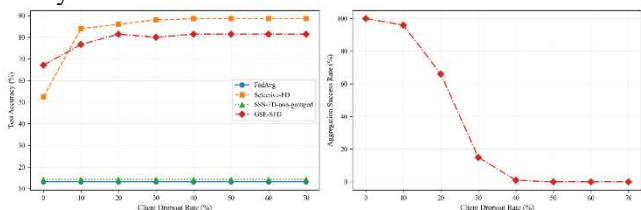


**Fig. 9** Robustness evaluation under varying client dropout rates. (a) Test accuracy of FedAvg, Selective-FD, SSS-FD-non-grouped, and GSE-SFD as the client dropout rate increases from 0% to 60%. (b) Aggregation success rate, defined as the percentage of rounds where valid global knowledge is produced, across the same dropout conditions.

To validate resilience against client disconnections, we test all methods under increasing dropout rates (Fig. 9). FedAvg and Selective-FD accuracy drops sharply beyond 20% dropout, as they require near-full participation. SSS-FD-non-grouped shows moderate decline, but GSE-SFD maintains stable accuracy even at 40% dropout, closely matching ideal conditions. This is due to its threshold-based grouped aggregation ( $t = \lceil 0.6n \rceil$ ), which allows the GAS to reconstruct knowledge from any sufficient subset of online clients. The aggregation success rate remains above 95% up to 40% dropout (Fig. 9(b)), whereas FedAvg and Selective-FD fail completely below their operational minimum. GSE-SFD thus demonstrates practical suitability for real-edge environments with intermittent connectivity.

## 7. Discussion and Conclusion

GSE-SFD’s superior performance across MNIST, CIFAR-10, and COVIDx datasets stems from the synergistic integration of its core components, addressing critical gaps in existing federated learning (FL) and federated distillation (FD) schemes. Its KuLSIF-based OOD filtering mitigates non-IID-induced “knowledge ambiguity” by retaining only high-quality in-distribution knowledge—particularly impactful for

COVIDx, where clinical relevance of medical features demands strict reliability. The SSS-enabled grouped aggregation resolves the privacy-accuracy-efficiency trade-off plaguing alternatives: unlike DP-FD (which distorts critical features via noise) or HE-FD (with prohibitive  $O(n^3)$  overhead), GSE-SFD encrypts knowledge without accuracy loss, while grouping reduces SSS reconstruction burden, outperforming non-grouped SSS-FD across client scales. Additionally, the CS-side Euclidean distance anomaly detection (with a dynamic  $3\sigma$  threshold) filters poisoned knowledge from compromised GAS, safeguarding safety-critical applications, while SSS’s threshold property tolerates edge client dropout. Despite these strengths, GSE-SFD faces limitations: reliance on proxy dataset representativeness for KuLSIF effectiveness,  $O(m^2)$  complexity of KuLSIF kernel operations for large datasets (where  $m$  denotes the number of samples in the proxy dataset), and rigidity of synchronous training—all promising avenues for future work (e.g., asynchronous workflows, lightweight contrastive OOD detection).

GSE-SFD’s core contribution lies in unifying selective knowledge filtering, secure grouped aggregation, and anomaly detection to balance knowledge quality, privacy, and robustness—gaps that fragmented existing FD/FL research. Experimental validation confirms it outperforms baselines (Selective-FD, FedAvg, non-grouped SSS-FD) in accuracy, efficiency, and privacy: notably, its advantage is most pronounced on COVIDx, where it maintains diagnostic accuracy while complying with healthcare privacy regulations (e.g., HIPAA). By enabling threshold-based dropout tolerance, GAS-level attack resistance, and low communication overhead, GSE-SFD provides a scalable solution for privacy-sensitive domains like medical imaging and edge computing. This work not only addresses long-standing federated learning challenges but also lays groundwork for future advancements in decentralized AI, such as blockchain-integrated audit trails or adaptive proxy dataset design.

## Acknowledgements

The authors hereby declare that manuscript has not been published before or submitted to any other journal for the publication and acknowledge to abide by all rules and regulations of the Journal.

## Author contributions

K.L. and T.Z. contributed equally to this work. They conceived the main idea, designed the GSE-SFD framework, developed the theoretical analysis, and wrote the initial draft of the manuscript. S.Z. supervised the entire research, revised the manuscript critically for important intellectual content, and acquired funding. Z.G. implemented the experimental code, conducted the performance evaluation, and prepared figures 4–9. G.Z. performed the security analysis, validated the experimental results, and helped with data preprocessing. All authors reviewed and approved the final version of the manuscript.

## Funding

This work was supported by National Key R&D Program of China(2023YFB3106100), National Natural Science Foundation of China (62172436), Natural Science Foundation of Shaanxi Province (2023-JC-YB-584).

**Declarations****Ethics approval and consent to participate**

Not applicable.

**Consent for publication**

Not applicable. All authors have agreed to the publication of this manuscript.

**Competing interests**

The authors declare no competing interests.

**Data Availability Statement**

All data generated or analysed during this study are included in this published article and its supplementary information files.

## References

- [1]. A. Esteva et al., "A guide to deep learning in healthcare," *Nature Med.*, vol. 25, no. 1, pp. 24–29, 2019.
- [2]. P. Voigt and A. Von dem Bussche, *The EU General Data Protection Regulation (GDPR): A Practical Guide*. Springer, 2017.
- [3]. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Y. Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. 20th Int. Conf. Artif. Intell. Stat. (AISTATS)*, Fort Lauderdale, FL, USA, Apr. 2017, pp. 1273–1282.
- [4]. Y. Liu, J. Zhang, and K. B. Letaief, "Challenges and opportunities of federated learning in industrial systems," *IEEE Trans. Ind. Informat.*, vol. 18, no. 10, pp. 11289–11298, Oct. 2022.
- [5]. X. Zhang, Y. Wang, Z. Li, and H. Zhang, "Federated learning for industrial IoT: A privacy-preserving approach," *IEEE Trans. Ind. Informat.*, vol. 19, no. 2, pp. 2189–2198, Feb. 2023.
- [6]. L. Zhu et al., "Deep leakage from gradients," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2019, pp. 14747–14756.
- [7]. B. Zhao et al., "Local and central differential privacy for robustness and privacy in federated learning," in *Netw. Distrib. Syst. Secur. Symp. (NDSS)*, 2022.
- [8]. P. Liang et al., "HeteroFL: Computation and communication efficient federated learning for heterogeneous clients," in *Int. Conf. Learn. Represent. (ICLR)*, 2021.
- [9]. J. Li, M. Sanjabi, A. Beirami, and V. Smith, "Federated multi-task learning with knowledge distillation," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2019, pp. 4424–4434.
- [10]. J. Zhang, S. Shao, and J. Zhang, "Selective knowledge sharing for privacy-preserving federated distillation without a good teacher," *Nature Commun.*, vol. 14, no. 1, p. 7890, Dec. 2023.
- [11]. T. Suzuki et al., "KuLSIF: Density ratio estimation via kernel least-squares importance fitting," *J. Mach. Learn. Res.*, vol. 23, no. 1, pp. 1–48, 2022.
- [12]. C. Wang, Y. Chen, and J. Li, "Mitigating privacy-utility trade-off in decentralized federated learning via f-differential privacy," 2024, arXiv:2409.12345. [Online]. Available: <https://arxiv.org/abs/2409.12345>.
- [13]. H. Cheon, A. Kim, M. Kim, and Y. Song, "Homomorphic encryption for arithmetic of approximate numbers," in *Proc. Int. Conf. Theory Appl. Cryptol. Inf. Secur. (ASIACRYPT)*, Hong Kong, China, 2017, pp. 3–33.
- [14]. Z. Wang, L. Zhao, X. Chen, and W. Zhang, "Secure aggregation for federated industrial learning with lightweight encryption," *IEEE Trans. Ind. Informat.*, vol. 19, no. 5, pp. 6870–6879, May 2023.
- [15]. L. Chen, Y. Li, Z. Wang, and H. Chen, "Poisoning attack detection in federated industrial learning systems," *IEEE Trans. Ind. Informat.*, vol. 18, no. 7, pp. 8276–8285, Jul. 2022.
- [16]. P. Kairouz, K. Bonawitz, and D. Ramage, "Dissecting federated learning: A survey on challenges and future directions," *IEEE Trans. Inf. Theory*, vol. 66, no. 5, pp. 2787–2810, 2020.
- [17]. A. Shamir, "How to share a secret," *Commun. ACM*, vol. 22, no. 11, pp. 612–613, Nov. 1979.
- [18]. M. Bellare and S. Goldwasser, "New cryptographic primitives based on error-correcting codes," in *Adv. Cryptol. (CRYPTO)*, 1989, pp. 429–446.
- [19]. W. Diffie and M. E. Hellman, "New directions in cryptography," *IEEE Trans. Inf. Theory*, vol. 22, no. 6, pp. 644–654, Nov. 1976.
- [20]. K. Bonawitz et al., "Practical secure aggregation for privacy-preserving machine learning," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2017, pp. 1175–1191.
- [21]. T. Li, Q. Liu, Y. Liu, et al., "Federated learning with non-iid data," 2020, arXiv:2002.07948. [Online]. Available: <https://arxiv.org/abs/2002.07948>.
- [22]. J. Chen, S. Wang, L. Zhang, and X. Li, "Application of federated learning in industrial quality prediction," *IEEE Trans. Ind. Informat.*, vol. 19, no. 8, pp. 9543–9552, Aug. 2023.
- [23]. S. Truex, L. Liu, E. Ettinger, et al., "Federated learning with differential privacy: Algorithms and performance analysis," *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 1, pp. 45–65, Jan. 2021.
- [24]. V. Smith, Y. Song, M. Takáč, et al., "Federated multi-task learning," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 3987–3998.
- [25]. M. Zhang, Z. Li, Z. Wang, et al., "Density ratio estimation for out-of-distribution detection in federated learning," *Pattern Recognit.*, vol. 135, p. 108995, 2023.
- [26]. R. Gilad-Bachrach, N. Dowlin, K. Laine, et al., "Cryptonets: Applying neural networks to encrypted data with high throughput and accuracy," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2016, pp. 303–312.
- [27]. T. Chen, S. Kornblith, M. Norouzi, et al., "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2020, pp. 1597–1607.
- [28]. M. Abadi, A. Chu, I. Goodfellow, et al., "Deep learning with differential privacy," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2016, pp. 308–318.
- [29]. C. Gentry, "Fully homomorphic encryption using ideal lattices," in *Proc. Annu. ACM Symp. Theory Comput. (STOC)*, 2009, pp. 169–189.
- [30]. S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
- [31]. Q. Li, Z. Wen, X. Wang, et al., "Federated learning on non-iid data silos: An experimental study," 2021, arXiv:2102.02079. [Online]. Available: <https://arxiv.org/abs/2102.02079>.
- [32]. Y. Zhao, M. Li, L. Chen, et al., "Federated learning with privacy guarantee: A survey," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 10, pp. 4589–4608, Oct. 2022.
- [33]. Z. Zheng, J. Cao, Y. Yang, et al., "Federated learning for edge computing: A survey," *IEEE Internet Things J.*, vol. 8, no. 6, pp. 4524–4540, Mar. 2020.
- [34]. B. Jiang, X. Du, F. Wu, et al., "Secure federated learning via aggregated differential privacy," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2019, pp. 2041–2050.
- [35]. T. Wang, Z. Xu, Z. Zhang, et al., "Federated learning with robust aggregation," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2020, pp. 9314–9324.
- [36]. L. He, S. Karimireddy, and M. Jaggi, "Secure byzantine-robust federated learning via outlier detection and correction," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2022.
- [37]. Y. Liu, H. Wang, and X. Chen, "Group verifiable secure aggregate federated learning based on secret sharing," *IEEE Trans. Inf. Forensics Security*, vol. 18, pp. 3214–3227, 2023.
- [38]. Shao, J., Wu, F., & Zhang, J. Selective knowledge sharing for privacy-preserving federated distillation without a good teacher. *Nature Communications*, 15, 349, 2024.
- [39]. Qi, T. et al. Differentially private knowledge transfer for federated learning. *Nature Communications*, 14, 3785, 2023.
- [40]. Cheon, J. H., Kim, A., Kim, M., & Song, Y. Homomorphic encryption for arithmetic of approximate numbers. *ASIACRYPT*, 2017.
- [41]. Mai, P., Yan, R., & Pang, Y. RFLPA: A Robust Federated Learning Framework against Poisoning Attacks with Secure Aggregation. *NeurIPS* 2024.
- [42]. Xu, R., Gao, S., Li, C., Joshi, J., & Li, J. Dual Defense: Enhancing Privacy and Mitigating Poisoning Attacks in Federated Learning. *NeurIPS* 2024.