Fortifying Federated Learning Towards Trustworthiness via Auditable Data Valuation and Verifiable Client Contribution

K Naveen Kumar[1], Ranjeet Ranjan Jha[2], C Krishna Mohan[1], and Ravindra Babu Tallamraju[3]
[1]IIT Hyderabad, India [2]IIT Patna, India [3]Sahaj AI Software Pvt Ltd., India

Problem Statement

- Federated Learning (FL) facilitates clients to collaborate on training a shared machine learning model without **exposing individual private data.**
- Requires **auditability** and **verifiability** to ensure **local data integrity** and **trustworthy client updates**.
- Existing FL frameworks rely on a server for **client selection** and **aggregation**, creating a **single point of failure** and **privacy risks**.

Proposed Solution: FAVD (Figure 1)

- **Federated auditable and verifiable data valuation (FAVD)** ensures auditability and verifiability of client contributions without a central authority.
- Utilizes **local data density functions** to construct a global density function for transparent and effective data valuation before training.
- **Gaussian noise** is added to shared density functions to mitigate privacy risks.

Key Contributions

- **Privacy-preserving data valuation** independent of any predefined training algorithm.
- Ensures **benign updates** even in the presence of **malicious data.**
- Theoretical analysis on **convergence**, **auditability**, **verifiability**, and **resilience** against **data poisoning threats.**
- Improves fairness by **distinguishing** high-impact contributions from low-quality data.
- **Comprehensive evaluation on five benchmark datasets with various models:** Covid-chestxray (ResNet50), Camelyon17 (DenseNet121), HAM10000 (FL-FixCaps), CIFAR10 (ResNet18), CIFAR100 (ResNet50).
- **Introduces novel metrics for auditable data valuation**
  - Malicious Sample Detection Rate **(MSDR)**
  - Benign Misclassification Rate **(BMR)**
  - Client Contribution Consistency **(CCC)** Score

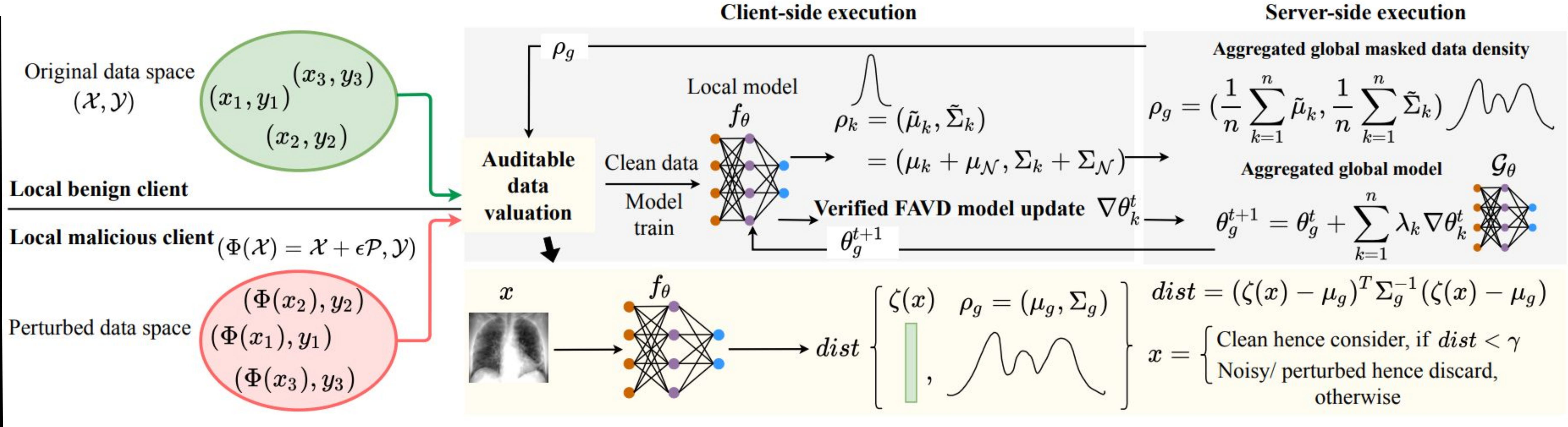

Figure 1. Overview of the FAVD-integrated FL system, with client-side and server-side execution.

Algorithm 1 Standard FL with **our FAVD framework**

Input: Global model $\mathcal{G}_{\theta,t}$, local data $\mathcal{D}_k = (\mathcal{X}_k, \mathcal{Y}_k)$, privacy noise parameters $\rho_\mathcal{N} = (\mu_\mathcal{N}, \Sigma_\mathcal{N})$, auditing parameter γ
Output: Global test accuracy $\mathcal{A}_\mathcal{G}$

```
Client execution (θ_g^t, ρ_g = (μ_g, Σ_g)):
for each client k = 1 to n do
    Initialize the local model θ_k^t ← θ_g^t
    if client k is malicious then
        Φ(X_k) ← X_k + εP
        D_k(X_k) ← D_k(Φ(X_k))
    D_k^F, ρ_k ← FAVD(D_k, γ, ρ_N, f_θ,k)
    for b = 1 to batches ∈ D_k^F = (X_k^F, Y_k) do
        {Q_b,k} ← {σ(f_θ,k(X_k^F[b]))}
        L_CE_k(Q_b,k, Y_b,k) ← Eq. 2, Cross-entropy loss
        θ_k^t ← θ_k^t − η∇_θ_k^t L_CE_k
    ∇θ_k^t ← θ_k^t − θ_g^t                ▷ FAVD verified updates
    return ∇θ_k^t, ρ_k = (μ̃_k, Σ̃_k)
Server execution (∇θ_k^t, ρ_k):
Receive model updates from selected clients ← ∇θ_k^t
Perform model aggregation using FedAvg (Eq. 1)
Update the global model parameter: θ_g^{t+1}
Update ρ_g ← ((1/n) Σ_{k=1}^n μ̃_k, (1/n) Σ_{k=1}^n Σ̃_k)
Share θ_g^t, ρ_g = (μ_g, Σ_g) to all the clients
Compute A_G ← Test (G_{θ_g^{t+1}}, X_test)
return A_G
```

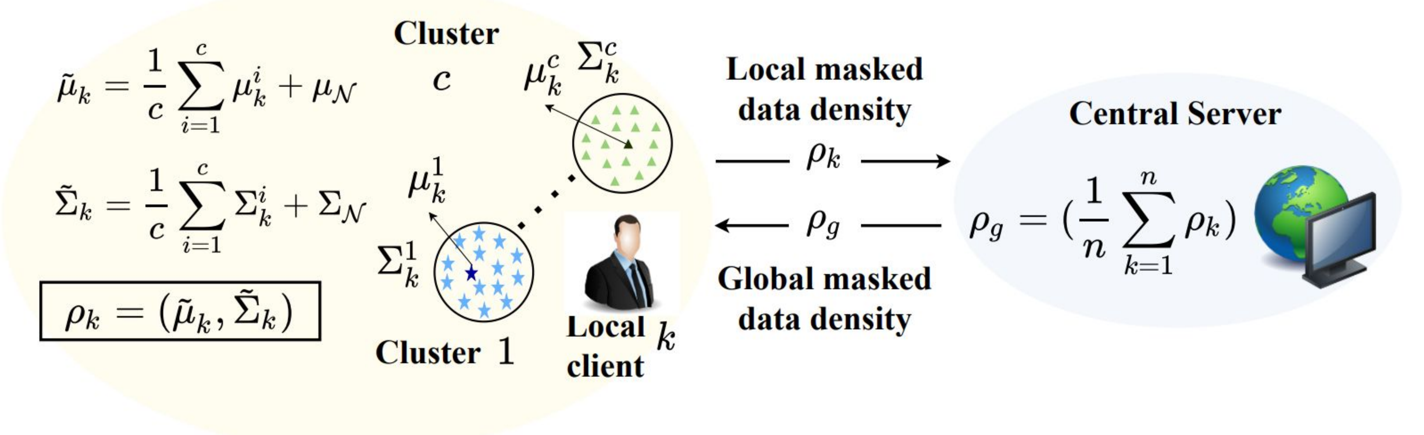

Figure 2. FAVD auditable data valuation process

Definition 3.1 *(ω-bounded poisoned data.) Let $\mathcal{F}$ be the adversary's function class. An adversarial perturbation $\mathcal{U}$ is characterized by the mapping $\mathcal{U} := \mathcal{F} \times \mathcal{X}_k \times \mathbb{R} \to \Phi(\mathcal{X}_k)$. For $\omega > 0$, we define the L_2 norm ball as $\mathcal{B}_2(\mathcal{X}_k, \omega) := \{\Phi(\mathcal{X}_k) \in \mathbb{R}^d : \|\Phi(\mathcal{X}_k) - \mathcal{X}_k\| \leq \omega\} \bigcap \mathcal{X}_k$. We classify the adversarial perturbation $\mathcal{U}$ as ω-bounded if it adheres to the condition $\mathcal{U}(\mathcal{F}, \mathcal{X}_{k,i}, \mathcal{Y}_{k,i})_{i=1}^{\nu} \in \mathcal{B}(\mathcal{X}_{k,i}, \omega)_{i=1}^{\nu}$. Furthermore, for a given $\omega > 0$, we denote the worst-case adversarial perturbation ($\mathcal{U}^+$) as*

$$\mathcal{U}^+ := \underset{\Phi(\mathcal{X}_k)\in\mathcal{B}(\mathcal{X}_k,\omega)}{\arg\max} \mathcal{L}_{CE}(f_\theta(\Phi(\mathcal{X}_{k,i}), \mathcal{Y}_{k,i})_{i=1}^{\nu}),$$

where $\mathcal{L}_{CE}$ and f_θ represent the cross-entropy loss function and parameters of the local black-box model, respectively.

Table 1. Global test accuracy (AG%) ↑ without data poisoning in uniform and non-IID FL settings, with and without FAVD data valuation. "No Val" refers to FL without data valuation.

Dataset →		C-xray [11]	Cam17 [3]	HAM10K [49]	CIFAR10 [24]	CIFAR100 [24]
Shard ↓	FL setting → / Method ↓	$n=4$, $k=4$	$n=5$, $k=5$	$n=64$, $k=64$	$n=100$, $k=70$	$n=10^3$, $k=500$
Uniform	No Val	$91.37_{\pm1.54}$	$91.82_{\pm1.21}$	$73.61_{\pm1.59}$	$85.59_{\pm0.76}$	$78.45_{\pm0.81}$
	FAVD (ours)	$93.75_{\pm1.20}$	$94.62_{\pm1.15}$	$75.48_{\pm0.42}$	$88.42_{\pm0.21}$	$81.19_{\pm0.45}$
non-IID	No Val	$88.62_{\pm1.27}$	$87.81_{\pm1.79}$	$68.79_{\pm0.83}$	$82.93_{\pm1.18}$	$73.56_{\pm1.74}$
	FAVD (ours)	$91.36_{\pm1.31}$	$91.38_{\pm1.86}$	$71.36_{\pm0.86}$	$86.24_{\pm1.22}$	$76.38_{\pm1.83}$

Result Discussion

- **Higher accuracy in no-threat:** FAVD improves accuracy by 2-3% (uniform) and 3-4% (non-IID) by discarding misaligned data.
- **Resilient to poisoning attacks:** Limits accuracy drop to 3-7% (uniform) and 8-10% (in highly non-IID settings).
- **Superior data valuation:** Achieves highest MSDR (0.90) and BMR (0.86), surpassing other methods.
- **Better client contribution consistency (CCC):** Maintains high CCC (~1), ensuring verifiable contributions even under threats.

Limitations and Future Work

- **Benign data exploitation:** Enhance feature alignment, explore encryption & secure multi party computation protocols.
- **Adaptive attack evasion:** Add behavioral analysis & temporal consistency checks.

Summary and Conclusion

FAVD enhances auditability, verifiability, and privacy in FL by leveraging local and global density functions for precise data valuation. Extensive evaluations show superior performance over state-of-the-art methods across diverse datasets.

References

K. Naveen Kumar, *et al.*, "The Impact of Adversarial Attacks on Federated Learning: A Survey," *IEEE TPAMI,* **2024.**

[1]{cs19m20p000001@, ckm@cse.}iith.ac.in, [2]rrjha@iitp.ac.in, [3]ravindrababu.t@gmail.com