Maverick: *Collaboration-free* Federated Unlearning for Medical Privacy

Win Kent Ong and Chee Seng Chan

Faculty of Comp. Sci. and Info. Tech., Universiti Malaya, Malaysia
{winkent.ong, cs.chan}@um.edu.my

Abstract. Federated Learning (FL) enables decentralized model training while preserving patient privacy, making it essential for medical AI applications. However, regulatory frameworks such as GDPR, CCPA, and LGPD mandate *"the right to be forgotten"*, requiring patient data removal from trained models upon request. This has driven growing interest in Federated Unlearning (FU), but existing methods require the collaborative participation of all clients, which is often impractical and raises privacy concerns. This paper proposes Maverick, a novel *Collaboration-free* FU framework that enables localized unlearning at the target client by minimizing model sensitivity, without requiring global collaboration from all clients to unlearn a target client. Theoretical analysis and extensive experiments on three medical imaging datasets, Colorectal Cancer Histology, Pigmented Skin Lesions, and Blood Cells, demonstrate Maverick's effectiveness in sample, class, and client unlearning scenarios. Maverick ensures trustworthy FL in healthcare while complying with regulations. The code is publicly available at https://github.com/OngWinKent/Maverick

Keywords: Trustworthy AI · Federated Unlearning.

1 Introduction

Federated Learning (FL) [8,24] enables decentralized model training across multiple parties without sharing raw data, a critical feature for preserving privacy in the medical domain. However, with the advent of stringent privacy regulations such as the GDPR [26], CCPA [17], and LGPD [7], the collaborative learning landscape has fundamentally changed. These regulations enforce the *"right to be forgotten"*, requiring that individuals can request the removal of their personal data from trained models. Federated Unlearning (FU) [21] addresses this by enabling selective data removal without retraining [3], thereby reshaping how FL systems must handle data deletion in compliance with modern legal requirements.

Despite advancements in FU [23,35,28], most existing methods require the coordinated participation of all clients (*i.e.,* global collaboration) to remove a target client's data. For example, if a medical institution $C_u \in \mathcal{C}$ decides to withdraw its data due to an institution-specific privacy policy change, the process becomes challenging because it requires the involvement of all other institutions $\mathcal{C} \setminus C_u$. Not all institutions may be willing to participate because this approach increases the risk of privacy leakage [27] and imposes higher computational costs.

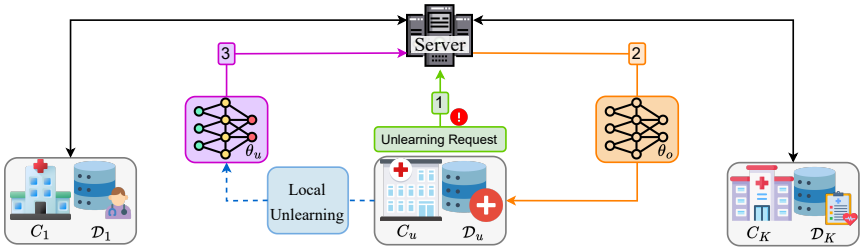


Fig. 1: Overview of Maverick. Upon receiving an unlearning request from client C_u, the server initializes the global model θ_o for *local unlearning* at $\mathcal{C}_u$. After unlearning optimization, $\mathcal{C}_u$ uploads the unlearned model θ_u back to the server.

In this paper, we propose Maverick, the first *Collaboration-free* FU framework, enabling unlearning based on a single client's request as shown in Fig. 1 for medical applications. As soon as a client requests data removal, our method can act independently, eliminating the need for global collaboration from other clients. This key innovation not only simplifies the unlearning process, but also minimizes privacy risks by keeping the sensitive operation localized to the requesting client.

Our key contributions are:

1. **Collaboration-free Unlearning Framework:** We propose Maverick, that unlearns a target client without requiring global collaboration of other clients, preventing privacy leakage and representing the first significant contribution.
2. **Model Sensitivity Reduction:** We introduce model sensitivity to quantify a model's output changes under input perturbations. To achieve unlearning, we propose locally minimizing model sensitivity by the target client.
3. **Comprehensive Validation:** Through theoretical analysis and extensive experiments on three medical imaging datasets, Colorectal Cancer Histology Slides, Pigmented Skin Lesions, and Blood Cells. We demonstrate our method's robustness across sample, class, and client unlearning scenarios.

2 Related Works

Federated Unlearning (FU) addresses sample, class, and client unlearning, as centralized unlearning methods [3,13] are ineffective due to incremental learning and restricted dataset access. *Sample unlearning*, initiated with FedRR [22] to remove individual samples with methods like QuickDrop [6], FedFilter [29] and FedAU [14]. *Class unlearning*, initiated with FedCDP [28] to eliminate data classes using techniques like Momentum Degradation (MoDE) [36]. *Client unlearning*, introduced with FedEraser [21], employs methods such as FRU [33], FedRecover [2], FCU [5], FedRecovery [35] and VeriFI [10] to remove the influence of clients.

Among existing approaches, the most related works [20,31,34] rely on server-side Fisher information sharing, introducing side-channel risks and communications overhead. In contrast, Maverick applies a single, noise-hardened update

on the unlearning client's data, without external regularization or cross-party communication, improving both efficiency and privacy.

Nevertheless, despite these advancements, current FU methods still require global client collaboration even when unlearning is requested by only a single client. This raises critical concerns around privacy, scalability and practicality, particularly in sensitive domains such as medical imaging. There remains a **clear gap**: the absence of a framework that enables unlearning in a fully client-local manner *without* involving other participants.

3 Problem Definition

A FL system consists of multiple clients, $\mathcal{C} = \{C_1, \cdots, C_K\}$ where K denotes the number of clients (*e.g.,* hospitals) and a central server collaboratively trains a global model, $\theta_o = \mathcal{A}(\mathcal{D})$, using a learning algorithm $\mathcal{A}$, on a distributed medical dataset, $\mathcal{D} = \{\mathcal{D}_1, \cdots, \mathcal{D}_K\}$, where $\mathcal{D}$ is the aggregate dataset between all clients.

Federated Unlearning. An unlearn client C_u (*i.e.,* target client C_k^u) requests the removal of their local medical dataset $\mathcal{D}_k^u \subseteq \mathcal{D}$ (*i.e.,* unlearn dataset $\mathcal{D}_u$) from the originally trained θ_o, ensuring it no longer contains $\mathcal{D}_u$. Hence, the unlearning algorithm $\mathcal{U}$ is to produce an unlearned model $\theta_u = \mathcal{U}(\theta_o)$ with $\mathcal{D}_u$'s influence is no longer present. So technically, the goal is for θ_u to perform similarly to a model that retrains from scratch on retain dataset $\mathcal{D}_r = \mathcal{D} \setminus \mathcal{D}_u$, $\theta_r = \mathcal{A}(\mathcal{D}_r)$.

Definition 1 (Exact Unlearning). *An unlearning algorithm $\mathcal{U}$ is considered (ϵ, δ)-unlearned if the distributions $u = \mathbb{P}(\theta_u)$ and $r = \mathbb{P}(\theta_r)$ are (ϵ, δ)-close.*

Specifically, u and r are (ϵ, δ)-close if $u(\mathcal{H}) \leq e^\epsilon r(\mathcal{H}) + \delta$ and $r(\mathcal{H}) \leq e^\epsilon u(\mathcal{H}) + \delta$ for all measurable events $\mathcal{H}$.

According to Def. 1, an unlearning algorithm $\mathcal{U}$ will achieve *exact unlearning* [11] if $\epsilon = \delta = 0$, yielding a distribution u identical to r. However, this strict compliance is often impractical due to high computational costs and utility loss.

Approximate Unlearning. To address this, *approximate unlearning* algorithms [12] have been introduced by relaxing (ϵ, δ)-bounds, offering a balance between efficiency and performance comparable to *exact unlearning*. For this, θ_u must meet two primary requirements with respect to $\mathcal{D}_u$ and $\mathcal{D}_r$:

1. *Fidelity*: $\mathcal{U}$ should not compromise the accuracy of θ_u on $\mathcal{D}_r$. Specifically, the logits output of θ_u should be consistent with θ_o for inputs from $\mathcal{D}_r$:
$$\underset{i}{\text{argmax}}\ f^i_{\theta_u}(x) = \underset{i}{\text{argmax}}\ f^i_{\theta_o}(x),\ \forall x \in \mathcal{D}_r, \tag{1}$$
where $f^i_\theta(x)$ denotes the i_{th} logit for input x.
2. *Effectiveness*: θ_u should avoid memorization [9] of $\mathcal{D}_u$ by exhibiting incorrect predictions [12] on $\mathcal{D}_u$. Particularly, the logits output of θ_u should not correspond to the ground-truth label y for inputs from $\mathcal{D}_u$:
$$\underset{i}{\text{argmax}}\ f^i_{\theta_u}(x) \neq y,\ \forall (x, y) \in \mathcal{D}_u. \tag{2}$$

Unlearning Scenarios. In this paper, we consider three different unlearning scenarios. First, *sample unlearning*: This occurs when an individual data owner (*i.e.,* patients) $c_j \in C_u$ withdraws consent for the use of his/her specific medical records, thereby eliminating its influence on θ_o (*i.e.,* $\mathcal{D}_u \subseteq \mathcal{D}^u_{k,j}$ for a particular sample j). Second, *class unlearning*: This arises when C_u decides to remove an entire imaging class m (*e.g.,* a hospital no longer wants to share all CT scans of a particular disease type) from its original dataset, thereby excluding that class from θ_o's generalization boundary (*i.e.,* $\mathcal{D}_u \subseteq \mathcal{D}^u_{k,m}$). Finally, *client unlearning*: This takes place when a C_u opts to exit the FL medical ecosystem (*e.g.,* a participating hospital decides to withdraw from the consortium), thereby necessitating the complete removal of the client's dataset $\mathcal{D}^u_k$ from θ_o (*i.e.,* $\mathcal{D}_u = \mathcal{D}^u_k$). Overall, these three unlearning scenarios: *sample*, *class*, and *client* are designed to ensure that the global FL model θ_o can be selectively "forgotten" with respect to different levels of data granularity, thereby satisfying personal (*e.g.,* patient) privacy rights and institutional (*e.g.,* clinic, hospital) data governance requirements.

4 Methodology

4.1 Model Sensitivity

Inspired by Lipschitz continuity [30], which analyzes model behavior through input perturbations, we introduce *model sensitivity* s in Def. 2 to quantify memorization, focusing on local input variations rather than a global perspective [19].

Definition 2 (Model Sensitivity). *The model sensitivity s of the model f_θ with respect to the sample x is defined as $s = \mathbb{E}_\delta \frac{\|f_\theta(x) - f_\theta(x+\delta)\|_2}{\|\delta\|_2}$, where δ represents a perturbation*[1] *applied to the sample x.*

Def. 2 quantifies the rate of change in the model's output relative to input perturbations. A smaller value of s indicates that f_θ exhibits minimal memorization of sample x. This formulation averages output variation over perturbations δ, eliminating dependence on the entire dataset.

4.2 Maverick

The proposed framework achieves unlearning by minimizing model sensitivity s, reducing the model's response to variations in samples $x \in \mathcal{D}_u$ to "forgets" x through local unlearning, as shown in Fig. 2. When C_u requests data removal for $\mathcal{D}_u$, the global model θ_o is updated to an unlearned model θ_u in three steps.

Firstly, *perturbation sampling* introduces controlled noise δ drawn from Gaussian distribution to evaluate model sensitivity on samples $x \in \mathcal{D}_u$:

$$\tilde{x} = x + \delta, \text{ where } \delta \sim \mathcal{N}(0, \sigma^2), \tag{3}$$

where σ is the standard deviation of the injected Gaussian noise $\mathcal{N}$.

[1] δ can be sampled from various distributions, such as Gaussian, uniform, etc.

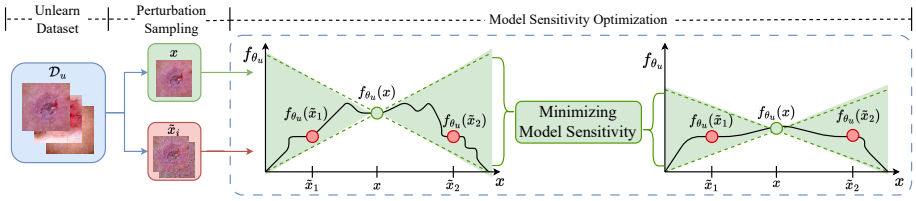


Fig. 2: Local unlearning reduces the bounded Lipschitz constant to minimize model sensitivity s on unlearn dataset $\mathcal{D}_u$, while maintaining overall performance.

Next, *Monte Carlo approximation* estimates model sensitivity using a finite sample method over perturbation sample size N, as defined in Def. 2:

$$\mathbb{E}_\delta \frac{\|f_{\theta_o}(x) - f_{\theta_o}(\tilde{x})\|_2}{\|x - \tilde{x}\|_2} \sim \frac{1}{N} \sum_{i=1}^{N} \frac{\|f_{\theta_o}(x) - f_{\theta_o}(\tilde{x}_i)\|_2}{\|\delta_i\|_2}, \tag{4}$$

where δ_i is the i_{th} perturbation sampled according to Eq. 3.

Finally, the *local unlearning* step derives θ_u through local optimization of model sensitivity as shown in Fig. 2. The optimization is defined as follows:

$$\theta_u = \underset{\theta_o}{\operatorname{argmin}}\, \mathbb{E}_{(x,y)\in\mathcal{D}_u} \frac{1}{N} \sum_{i=1}^{N} \frac{\|f_{\theta_o}(x) - f_{\theta_o}(\tilde{x}_i)\|_2}{\|\delta_i\|_2}, \tag{5}$$

ensuring that θ_u no longer retains information about $\mathcal{D}_u$. Maverick leverages Def.2 to operate locally at C_u, **enabling the unlearning of a single target client without requiring global collaboration from other clients**.

4.3 Theoretical Analysis

Theorem 1 demonstrates that Maverick satisfies the design requirements in Sec. 3.

Theorem 1 (Theoretical Bounds). *For C_u removing $\mathcal{D}_u \subseteq \mathcal{D}$ from θ_o, with $\mathcal{D}_r = \mathcal{D} \backslash \mathcal{D}_u$, the ℓ_p-perturbation $\Delta_p = \|\delta\|_p$ is bounded by $\beta_L \leq \Delta_p \leq \beta_U$. Within these bounds, Maverick ensures both fidelity and effectiveness requirements:*

$$\begin{cases} \underset{i}{\operatorname{argmax}}\ f^i_{\theta_u}(x) = \underset{i}{\operatorname{argmax}}\ f^i_{\theta_o}(x), & \forall x \in \mathcal{D}_r, \\ \underset{i}{\operatorname{argmax}}\ f^i_{\theta_u}(x) \neq y, & \forall (x,y) \in \mathcal{D}_u. \end{cases} \tag{6}$$

5 Experimental Results

5.1 Experimental Setup

Implementations. We simulate a horizontal FL setup with $K = 10$ clients in an IID setting, where each client receives 10% of the dataset. For *sample* and

client unlearning, we employ backdoor techniques [1] based on prior works [15,16], unlearning 40% or the entire $\mathcal{D}_k^u$ for client C_u. *Class unlearning* removes class 1 from $\mathcal{D}_k^u$. Our hyperparameters settings are: learning rate $\eta = 0.0001$, perturbation sample size $N = 10$, and Gaussian noise with σ ranging from 0.05 to 0.5 (see Sec.5.7). Each experiment is repeated over five trials on a single NVIDIA A100 GPU, with results reported as mean and standard deviation.

Model, Datasets & Evaluation Metrics. We use ResNet18 [18], following prior studies [28,35] on three publicly available medical imaging datasets[2] [32]: i) Colorectal Cancer Histology Slides (Path), ii) Pigmented Skin Lesions (Derma) and iii) Blood Cells (Blood).

For evaluation metrics, four different metrics are used: i) *Fidelity*, the accuracy on the retain dataset $\mathcal{D}_r$, where higher $\mathcal{D}_r$ accuracy indicates a greater fidelity. ii) *Effectiveness*, the accuracy on the unlearn dataset $\mathcal{D}_u$, where lower $\mathcal{D}_u$ accuracy indicates a greater effectiveness. iii) *Privacy*, is assessed via *Membership Inference Attack (MIA)* [4]. The attack success rate (ASR) is employed to determine if specific data were used in training. Lower ASR indicates a strong privacy guarantee. iv) *Efficiency*, is measured by runtime in seconds.

Baselines. We compare Maverick against the following methods: i) *Baseline*: The original model before unlearning. ii) *Retrain*: Retraining from scratch on $\mathcal{D}_r$ until convergence as the gold standard. iii) *Fine-tune(FT)*: Fine-tuning the baseline model on $\mathcal{D}_r$ for five epochs. iv) *FedCDP* [28]: A class unlearning approach using TF-IDF-guided channel pruning. v) *FedRecovery* [35]: A sample and client unlearning approach using client gradient submissions without retraining.

5.2 Fidelity Guarantee

We evaluate fidelity by measuring $\mathcal{D}_r$ accuracy, as shown in Tab. 1. While FT achieves high $\mathcal{D}_r$ accuracy, it is unsuitable for unlearning due to its ineffectiveness (see Sec. 5.3), privacy risks (see Sec. 5.4), and computationally expensive (see Sec. 5.6). FedCDP and FedRecovery outperform Maverick on $\mathcal{D}_r$ accuracy by 1-2% but lack consistency across scenarios, as they target specific unlearning scenario. In contrast, **Maverick maintains high fidelity** with consistent $\mathcal{D}_r$ accuracy across all scenarios with minimal deterioration.

5.3 Effectiveness Guarantee

We assess effectiveness by measuring accuracy on $\mathcal{D}_u$, as shown in Tab. 1. While all baselines remove $\mathcal{D}_u$ information to some extent, the FT method reduces $\mathcal{D}_u$ accuracy less effectively than others. FedCDP and FedRecovery show higher $\mathcal{D}_u$ accuracy than Maverick and lack scenario consistency. In contrast, **Maverick achieves the highest effectiveness**, with the lowest $\mathcal{D}_u$ accuracy across all scenarios, indicating a successful unlearning.

[2] `https://github.com/MedMNIST/MedMNIST`

Scenarios	Datasets	Metrics	Accuracy(%)					
			Baseline	Retrain	FT	FedCDP[28]	FedRecovery[35]	Maverick
Sample	Path	$\mathcal{D}_r$ ↑	91.37±1.16	92.50±1.10	**93.04±1.05**	70.19±1.52	90.14±0.96	89.43±1.49
		$\mathcal{D}_u$ ↓	90.48±0.92	0.00±0.00	46.13±1.72	22.62±1.16	2.35±1.16	**0.71±0.02**
		ASR ↓	92.51±1.13	8.69±0.65	55.49±1.28	38.05±1.33	13.43±0.62	**10.04±0.74**
	Derma	$\mathcal{D}_r$ ↑	81.63±1.27	80.97±1.71	**81.83±1.20**	67.44±0.79	80.30±1.74	79.35±1.62
		$\mathcal{D}_u$ ↓	93.18±1.65	0.00±0.00	54.05±1.04	17.36±1.63	1.35±0.11	**0.54±0.07**
		ASR ↓	93.05±0.93	6.49±0.73	62.84±0.78	30.15±0.66	9.32±0.72	**7.31±0.58**
	Blood	$\mathcal{D}_r$ ↑	93.17±1.25	93.32±1.07	**94.47±1.55**	80.34±1.66	92.85±1.15	92.54±1.38
		$\mathcal{D}_u$ ↓	91.53±1.49	0.00±0.00	37.59±1.15	20.79±1.48	1.56±0.26	**0.43±0.05**
		ASR ↓	95.53±0.87	5.59±0.91	43.35±0.79	31.61±0.64	9.76±0.69	**6.15±0.31**
Class	Path	$\mathcal{D}_r$ ↑	92.04±0.38	94.91±0.20	**95.39±0.82**	91.48±0.74	79.38±1.47	90.37±0.75
		$\mathcal{D}_u$ ↓	98.03±0.65	0.00±0.00	52.84±2.74	0.92±0.03	20.41±1.63	**0.37±0.05**
		ASR ↓	97.55±1.41	6.07±0.62	45.35±0.59	8.13±0.29	27.81±1.39	**7.74±0.47**
	Derma	$\mathcal{D}_r$ ↑	82.52±0.55	80.39±0.85	**81.38±0.37**	79.31±0.73	55.51±0.59	79.18±0.63
		$\mathcal{D}_u$ ↓	80.88±0.30	0.00±0.00	53.69±1.75	0.51±0.05	31.40±0.73	**0.18±0.02**
		ASR ↓	90.62±0.64	2.60±0.18	40.44±1.62	5.17±0.46	34.16±0.94	**0.49±0.31**
	Blood	$\mathcal{D}_r$ ↑	95.41±0.79	96.92±0.51	**96.05±0.48**	95.03±0.20	69.27±1.74	94.24±0.53
		$\mathcal{D}_u$ ↓	97.73±0.94	0.00±0.00	58.21±0.71	0.72±0.02	27.61±0.63	**0.49±0.01**
		ASR ↓	95.17±1.03	3.07±0.25	57.45±1.30	4.01±0.37	30.48±0.90	**3.26±0.43**
Client	Path	$\mathcal{D}_r$ ↑	89.13±1.51	91.67±1.23	**92.95±1.33**	73.19±2.36	87.94±1.05	87.08±1.26
		$\mathcal{D}_u$ ↓	91.98±1.39	0.00±0.00	48.83±1.57	27.52±1.94	2.85±1.94	**0.80±0.03**
		ASR ↓	93.49±1.04	8.59±0.32	57.49±0.53	40.37±1.82	14.63±0.45	**10.96±0.22**
	Derma	$\mathcal{D}_r$ ↑	78.36±0.92	79.34±1.49	**80.98±1.65**	65.95±1.57	77.59±1.49	76.17±0.93
		$\mathcal{D}_u$ ↓	95.33±2.98	0.00±0.00	59.60±1.94	19.45±2.11	1.63±0.19	**0.67±0.05**
		ASR ↓	95.27±1.63	6.05±0.59	65.38±0.96	35.59±0.94	11.32±0.71	**7.92±0.49**
	Blood	$\mathcal{D}_r$ ↑	91.21±1.16	91.90±2.41	**93.38±1.53**	79.58±1.07	89.54±1.09	88.33±1.64
		$\mathcal{D}_u$ ↓	92.83±0.62	0.00±0.00	42.38±0.82	25.29±1.44	1.95±0.27	**0.53±0.09**
		ASR ↓	96.71±1.28	5.78±0.51	52.57±1.20	39.85±1.52	10.95±0.33	**6.73±0.52**

Table 1: Comparison of accuracy on $\mathcal{D}_r$ and $\mathcal{D}_u$, along with the ASR of MIA across different unlearning methods and scenarios. **Bold** indicates the best.

5.4 Privacy Guarantee

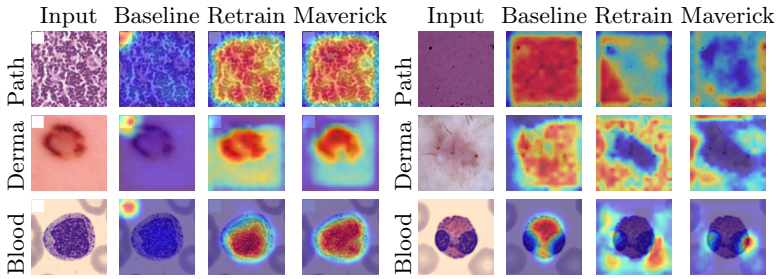
We evaluate privacy by measuring the ASR of MIA, ensuring that the unlearned model does not leak information about $\mathcal{D}_u$ as shown in Tab. 1. The FT method shows a high ASR, indicating minimal removal of $\mathcal{D}_u$ information. FedCDP and FedRecovery have higher ASR than Maverick and lack consistency across scenarios. In contrast, **Maverick provides the strongest privacy guarantee**, with the lowest ASR across all scenarios.

5.5 Attention Map

We analyze the attention maps [25] on $\mathcal{D}_u$ as shown in Fig. 3. The attention map highlights the key input features that influence model predictions. Maverick exhibits an attention distribution similar to the retrained model, avoiding the focus on unlearned regions. Specifically, the unlearned model ignores the top-left backdoor trigger in sample unlearning. In class unlearning, attention is shifted to the background, rather to the main object. This suggests $\mathcal{D}_u$ has minimal impact on the unlearned model's output, demonstrating the effectiveness of Maverick.

5.6 Efficiency Guarantee

Fig. 4 compares the runtime performance of different unlearning methods. Retrain is the slowest, while FT has a better speed but remains slower than other methods



(a) Sample unlearning. (b) Class unlearning.

Fig. 3: Attention maps for unlearning methods in *sample* and *class* scenarios.

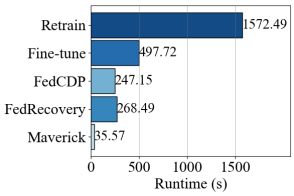
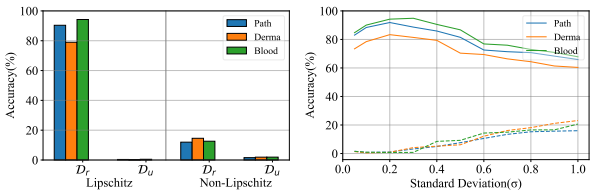


Fig. 4: Runtime comparison of unlearning methods, measured in seconds.



(a) Non-Lipschitz (b) Gaussian noise(σ)

Fig. 5: Ablation studies of Maverick for class unlearning. Solid line: $\mathcal{D}_r$; Dashed line: $\mathcal{D}_u$.

due to fine-tuning on large $\mathcal{D}_r$. FedCDP and FedRecovery perform better than FT but are hindered by reliance on global training datasets and costly gradient residual calculations. In contrast, **Maverick achieves the highest efficiency**, running 8x to 45x faster than all other baselines using only the local unlearning of target client's dataset and completing the unlearning in a single epoch.

5.7 Ablation Studies

Non-Lipschitz. We assess unlearning performance by removing the denominator in Eq. 5, termed the Non-Lipschitz method as shown in Fig. 5a. The results reveal catastrophic forgetting, with $\mathcal{D}_r$ accuracy falling below 10% due to misclassification into random classes. The failure stems from unbounded optimization, unlike the bounded Lipschitz constant provides theoretical guarantees in Theorem 1.

Gaussian Noise. We assess the impact of Gaussian noise on unlearning by varying the standard deviation, as shown in Fig. 5b. For $0.05 \leq \sigma \leq 0.5$, $\mathcal{D}_r$ accuracy remains high, and $\mathcal{D}_u$ accuracy remains low, fulfilling the unlearning requirements. Therefore, we adopted this range of σ in our study.

6 Conclusion

This paper introduces Maverick, a novel *Collaboration-free* federated unlearning framework. This framework represents a first step towards enabling local unlearning at the target client without the involvement of other clients in medical domain. It is achieved through model sensitivity optimization based on Lipschitz continuity. Our theoretical analysis and experimental work suggest that Maverick can improve fidelity, effectiveness, privacy, and efficiency across various unlearning scenarios. To further support community wide benefit, Maverick is designed to impose minimal disruption and computational burden on retained clients. This is an important consideration for privacy-sensitive and resource-constrained settings such as healthcare. These findings indicate its potential as a practical tool for advancing trustworthy federated learning in sensitive areas, with notable implications for societal and clinical practices.

Acknowledgments. This work is supported by the ASEAN-China Cooperation Fund (ACCF) under the project titled *"Deep Ensemble under Non-Ideal Conditions and Its Typical Applications in Computer Vision"*.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bagdasaryan, E., Veit, A., Hua, Y., Estrin, D., Shmatikov, V.: How to backdoor federated learning. In: AISTATS. pp. 2938–2948. PMLR (2020)
2. Cao, X., Jia, J., Zhang, Z., Gong, N.Z.: Fedrecover: Recovering from poisoning attacks in federated learning using historical information. In: 2023 IEEE Symposium on Security and Privacy (SP). pp. 1366–1383. IEEE (2023)
3. Cao, Y., Yang, J.: Towards making systems forget with machine unlearning. In: 2015 IEEE symposium on security and privacy. pp. 463–480. IEEE (2015)
4. Carlini, N., Chien, S., Nasr, M., Song, S., Terzis, A., Tramer, F.: Membership inference attacks from first principles. In: 2022 IEEE Symposium on Security and Privacy (SP). pp. 1897–1914. IEEE (2022)
5. Deng, Z., Luo, L., Chen, H.: Enable the right to be forgotten with federated client unlearning in medical imaging. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 240–250. Springer (2024)
6. Dhasade, A., Ding, Y., Guo, S., Kermarrec, A.m., De Vos, M., Wu, L.: Quickdrop: Efficient federated unlearning by integrated dataset distillation. arXiv preprint arXiv:2311.15603 (2023)
7. Dourado, D.d.A., Aith, F.M.A.: The regulation of artificial intelligence for health in brazil begins with the general personal data protection law. Revista de Saúde Pública **56**, 80 (2022)
8. Fan, T., Gu, H., et al.: Ten challenging problems in federated foundation models. IEEE TKDE **37**(07), 4314–4337 (2025)
9. Feldman, V.: Does learning require memorization? a short tale about a long tail. In: Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing. pp. 954–959 (2020)

10. Gao, X., Ma, X., Wang, J., Sun, Y., Li, B., Ji, S., Cheng, P., Chen, J.: Verifi: Towards verifiable federated unlearning. IEEE TDSC (2024)
11. Ginart, A., Guan, M., Valiant, G., Zou, J.Y.: Making ai forget you: Data deletion in machine learning. Advances in neural information processing systems **32** (2019)
12. Graves, L., Nagisetty, V., Ganesh, V.: Amnesiac machine learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 11516–11524 (2021)
13. Gu, H., Ong, W., Chan, C.S., Fan, L.: Ferrari: federated feature unlearning via optimizing feature sensitivity. Advances in Neural Information Processing Systems **37**, 24150–24180 (2024)
14. Gu, H., Zhu, G., Zhang, J., Zhao, X., Han, Y., Fan, L., Yang, Q.: Unlearning during learning: An efficient federated machine unlearning method. In: Larson, K. (ed.) IJCAI-24. pp. 4035–4043 (8 2024), main Track
15. Guo, Y., Zhao, Y., Hou, S., Wang, C., Jia, X.: Verifying in the dark: Verifiable machine unlearning by using invisible backdoor triggers. IEEE TIFS (2023)
16. Halimi, A., Kadhe, S., Rawat, A., Baracaldo, N.: Federated unlearning: How to efficiently erase a client in fl? arXiv preprint arXiv:2207.05521 (2022)
17. Harding, E.L., Vanto, J.J., Clark, R., Hannah Ji, L., Ainsworth, S.C.: Understanding the scope and impact of the california consumer privacy act of 2018. Journal of Data Protection & Privacy **2**(3), 234–253 (2019)
18. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on CVPR. pp. 770–778 (2016)
19. Latafat, P., Themelis, A., Stella, L., Patrinos, P.: Adaptive proximal algorithms for convex optimization under local lipschitz continuity of the gradient. Mathematical Programming pp. 1–39 (2024)
20. Li, Y., Lyu, X., Koren, N., Lyu, L., Li, B., Ma, X.: Anti-backdoor learning: Training clean models on poisoned data. Advances in Neural Information Processing Systems **34**, 14900–14912 (2021)
21. Liu, G., Ma, X., Yang, Y., Wang, C., Liu, J.: Federaser: Enabling efficient client-level data removal from federated learning models. In: 2021 IEEE/ACM 29th international symposium on quality of service (IWQOS). pp. 1–10. IEEE (2021)
22. Liu, Y., Xu, L., Yuan, X., Wang, C., Li, B.: The right to be forgotten in federated learning: An efficient realization with rapid retraining. In: IEEE INFOCOM 2022-IEEE Conference on Computer Communications. pp. 1749–1758. IEEE (2022)
23. Liu, Z., Jiang, Y., Shen, J., Peng, M., Lam, K.Y., Yuan, X., Liu, X.: A survey on federated unlearning: Challenges, methods, and future directions. ACM Computing Surveys **57**(1), 1–38 (2024)
24. McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: Artificial intelligence and statistics. pp. 1273–1282. PMLR (2017)
25. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE ICCV. pp. 618–626 (2017)
26. Voigt, P., Von dem Bussche, A.: The eu general data protection regulation (gdpr). A Practical Guide, 1st Ed., Cham: Springer International Publishing **10**(3152676), 10–5555 (2017)
27. Wang, F., Li, B., Li, B.: Federated unlearning and its privacy threats. IEEE Network **38**(2), 294–300 (2023)
28. Wang, J., Guo, S., Xie, X., Qi, H.: Federated unlearning via class-discriminative pruning. In: Proceedings of the ACM Web Conference 2022. pp. 622–632 (2022)

29. Wang, P., Yan, Z., Obaidat, M.S., Yuan, Z., Yang, L., Zhang, J., Wei, Z., Zhang, Q.: Edge caching with federated unlearning for low-latency v2x communications. IEEE Communications Magazine (2023)
30. Weng, T.W., Zhang, H., Chen, P.Y., Yi, J., Su, D., Gao, Y., Hsieh, C.J., Daniel, L.: Evaluating the robustness of neural networks: An extreme value theory approach. In: ICLR (2018)
31. Wu, L., Guo, S., Wang, J., Hong, Z., Zhang, J., Ding, Y.: Federated unlearning: Guarantee the right of clients to forget. IEEE Network **36**(5), 129–135 (2022)
32. Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., Ni, B.: Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. Scientific Data **10**(1), 41 (2023)
33. Yuan, W., Yin, H., Wu, F., Zhang, S., He, T., Wang, H.: Federated unlearning for on-device recommendation. In: Proceedings of the sixteenth ACM international conference on web search and data mining. pp. 393–401 (2023)
34. Zhang, K., Tao, G., Xu, Q., Cheng, S., An, S., Liu, Y., Feng, S., Shen, G., Chen, P.Y., Ma, S., et al.: Flip: A provable defense framework for backdoor mitigation in federated learning. arXiv preprint arXiv:2210.12873 (2022)
35. Zhang, L., Zhu, T., Zhang, H., Xiong, P., Zhou, W.: Fedrecovery: Differentially private machine unlearning for federated learning frameworks. IEEE Transactions on Information Forensics and Security (2023)
36. Zhao, Y., Wang, P., Qi, H., Huang, J., Wei, Z., Zhang, Q.: Federated unlearning with momentum degradation. IEEE Internet of Things Journal (2023)