Validation of Federated Unlearning on Collaborative Prostate Segmentation

Yann Fraboni, Lucia Innocenti, Michela Antonelli, Richard Vidal, Laetitia Kameni, Sebastien Ourselin, Marco Lorenzi

Validation of Federated Unlearning on Collaborative Prostate Segmentation

Yann Fraboni[1,2], Lucia Innocenti[1,3], Michela Antonelli[3], Richard Vidal[2], Laetitia Kameni[2], Sebastien Ourselin[3], and Marco Lorenzi[1]

[1] Epione Research Group, Inria Sophia Antipolis, Université Côte d'Azur, France
[2] Accenture Labs, Sophia Antipolis, France
[3] King's College London, School of Biomedical Engineering & Imaging Sciences, UK

Abstract. Machine Unlearning (MU) is an emerging discipline studying methods to remove the effect of a data instance on the parameters of a trained model. Federated Unlearning (FU) extends MU to unlearn the contribution of a dataset provided by a client wishing to drop from a federated learning study. Due to the emerging nature of FU, a practical assessment of the effectiveness of the currently available approaches in complex medical imaging tasks has not been studied so far. In this work, we propose the first in-depth study of FU in medical imaging, with a focus on collaborative prostate segmentation from multi-centric MRI dataset. We first verify the unlearning capabilities of a panel of FU methods from the state-of-the-art, including approaches based on model adaptation, differential privacy, and adaptive retraining. For each method, we quantify their unlearning effectiveness and computational cost as compared to the baseline retraining of a model from scratch after client dropout. Our work highlights a new perspective for the practical implementation of data regulations in collaborative medical imaging applications.

Keywords: federated unlearning · segmentation · prostate cancer.

1 Introduction

With the emergence of new data regulations [1, 2], the storage and processing of sensitive personal data is often subject of strict constraints and restrictions. In particular, the "right to be forgotten" states that personal data must be erased upon request, with subsequent potential implications on machine learning models trained by using this data. Machine Unlearning (MU) is an emerging discipline that studies methods to remove the contribution of a given data instance used to train a machine learning model [3].

Motivated by data governance and confidentiality concerns, federated learning (FL) has gained popularity in the last years to allow data owners to collaboratively learn a model without sharing their respective data. FL is particularly suited for Machine Learning applications in domains where data security is critical, such as in healthcare [4, 5]. With the current deployments of FL in the real-world, it is of crucial importance to extend MU approaches to *federated unlearning* (FU), to guarantee the unlearning of data instances from clients wishing

to opt-out from a collaborative training routine. This is not straightforward, since current MU schemes have been proposed essentially for centralized learning, and cannot be seamlessly applied to the federated one without breaking the data governance and privacy setting of FL. Recent FU methods have been proposed in the machine learning literature [6, 7, 8, 9], with their effectiveness being demonstrated on typical machine learning benchmarks [10, 11, 12]. Nevertheless, these benchmarks mostly focus on cross-device scenarios, with partitioning based on heuristics which often do not reflect the complex variability of real-world data analysis problems, such as the cross-site image biases and heterogeneity typical of collaborative medical imaging studies. The translation of FU in medical imaging applications requires the investigation of unlearning through the setup of realistic cross-silo benchmarks.

This work provides the first study of the effectiveness of existing FU approaches in a real-world collaborative medical imaging setup, focusing on federated prostate segmentation. To this end, we develop a benchmark composed by large publicly available prostate segmentation dataset, and define a realistic cross-silo FL scenario with heterogeneity depending on acquisition protocol and scanner. We introduce novel criteria to quantitatively compare the FU methods, assessing the 1) utility of the model after unlearning, 2) unlearning capability, and 3) efficiency of the unlearning procedure. Our results identify critical aspects of current unlearning methods, and show that paradigms based on adaptive retraining are the only effective FU approaches from the state-of-the-art.

This manuscript is structured as follows. In Section 2, we provide formal definitions for FL and the different existing FU schemes. We also introduce the metrics used to measure the effectiveness of an unlearning scheme. In Section 3, we introduce the federated dataset for prostate segmentation used in this work and verify the effectiveness of all the FU schemes.

2 Methodology

After providing in Section 2.1 the formalism of FL, we introduce FU in Section 2.2. We explain the limitations of MU methods for the federated setting in Section 2.3 and detail the existing FU schemes investigated in this paper in Section 2.4.

2.1 Federated Learning

FL consists in optimizing the average of local loss functions $\mathcal{L}_i$ across a set I of clients, weighted by their importance p_i such that $\sum_{i \in I} p_i = 1$, i.e.

$$\mathcal{L}(\boldsymbol{\theta}, I) = \sum_{i \in I} p_i \mathcal{L}_i(\boldsymbol{\theta}), \tag{1}$$

where $\boldsymbol{\theta}$ represents the parameters to be optimized. The weight p_i can be interpreted as the importance given by the server to client i in the federated

optimization problem which, without loss of generality, can be considered identical for every client, i.e. $p_i = 1/n$ where $n = |I|$. We define $\boldsymbol{\theta}^*$ the parameters minimizing the federated problem (1), i.e. $\boldsymbol{\theta}^* \coloneqq \arg\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, I)$.

To estimate the global optimum $\boldsymbol{\theta}^*$, FedAvg [13] is an iterative training strategy based on the aggregation of local model parameters. At each iteration step t, the server sends the current global model parameters $\boldsymbol{\theta}^t$ to the clients. Each client updates the model by minimizing the local cost function $\mathcal{L}_i$ through a fixed amount of SGD initialized with $\boldsymbol{\theta}^t$. Subsequently each client returns the updated local parameters $\boldsymbol{\theta}_i^{t+1}$ to the server. The global model parameters $\boldsymbol{\theta}^{t+1}$ at the iteration step $t+1$ are then estimated as a weighted average, i.e.

$$\boldsymbol{\theta}^{t+1} = \sum_{i \in I} p_i \boldsymbol{\theta}_i^{t+1}. \tag{2}$$

We define $\tilde{\boldsymbol{\theta}}$ the parameters vector obtained after performing FL over T server aggregations, i.e. $\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}^{T+1}$. When the clients' local loss functions $\mathcal{L}_i$ are convex, [14, 15] show that $\tilde{\boldsymbol{\theta}}$ converges to $\boldsymbol{\theta}^*$ as T goes to infinity.

2.2 Federated Unlearning

Removing a client c from the set of clients I modifies the federated problem (1), which becomes $\mathcal{L}(\boldsymbol{\theta}, I \setminus c)$. We define $\boldsymbol{\theta}^*_{-c} = \arg\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, I \setminus c)$ as the optimum of this new optimization problem. An FU scheme can be formalized as a function h taking as input $\tilde{\boldsymbol{\theta}}$, the model trained with every client in I, including c, to return parameters $h(\tilde{\boldsymbol{\theta}}, c)$ ideally equivalent to $\boldsymbol{\theta}^*_{-c}$. In practice, due to the non-convexity and stochasticity characterizing the optimization problems of typical medical imaging tasks, it is challenging to assess the proximity between the model $h(\tilde{\boldsymbol{\theta}}, c)$ and the ideal target $\boldsymbol{\theta}^*_{-c}$ in terms of pre-defined metrics in the parameters space. For this reason, in this work we quantify the quality of FU by introducing a series of criteria motivated by the ideal requirements that an unlearning scheme should satisfy.

To this end, we first notice that the baseline FU approach, here named Scratch, achieves unlearning by performing a new FedAvg training from scratch on the remaining clients $I \setminus c$. We define $\tilde{\boldsymbol{\theta}}_{-c}$ the parameters vectors obtained with Scratch which, by construction, provide perfect unlearning of client c. We note however that this procedure wastes the contribution of the other clients which was already available from the training of $\tilde{\boldsymbol{\theta}}$, i.e. the set of parameters $\{\boldsymbol{\theta}_i^t\}_{i \in I \setminus c, t \in \{1,\dots,T\}}$ gathered during federated optimization. Therefore, an effective FU methods should be more efficient than Scratch in optimizing $h(\tilde{\boldsymbol{\theta}}, c)$. These considerations motivate the following criteria to assess the unlearning quality of FU scheme:

- **Utility.** The predictive capability of the model with parameters $h(\tilde{\boldsymbol{\theta}}, c)$ on the testing sets of the available clients $I \setminus c$ should be equal or superior to the one of Scratch for the FU scheme considered. This criterion shows that the model resulting from FU maintains high predictive performances on the available clients data.

- **Unlearning.** Unlearning of client c implies the loss of predictive capabilities of the model $h(\tilde{\boldsymbol{\theta}}, c)$ on the training set of this client. If the performance of the model after unlearning is superior to the one of SCRATCH, we deduce that FU was ineffective in removing the information from client c.
- **Time.** The amount of server aggregations needed to complete the unlearning of client c should be inferior to the ones achieved by SCRATCH.

2.3 Machine Unlearning vs Federated Unlearning

Several MU methods have been proposed in the centralized learning setting [3]. Most MU approaches consists in defining h as a Newton step based on the Hessian H and gradients G estimated on all the remaining data points from the current model $\tilde{\boldsymbol{\theta}}$, i.e. $h(\tilde{\boldsymbol{\theta}}, c) = \tilde{\boldsymbol{\theta}} - H(\tilde{\boldsymbol{\theta}}, I \setminus c)^{-1} G(\tilde{\boldsymbol{\theta}}, I \setminus c)$ [16, 17, 18, 19, 20, 21]. The main drawback behind the use of this approach in the federated setting is that it requires clients to compute and share gradients and Hessians of the local loss function. This operation should be avoided in FL, as these quantities are known to potentially leak information about the training data [22]. Other approaches to MU consist in applying zero-mean Gaussian perturbations to the model parameters, with magnitude of the standard deviation σ depending on the properties of the data on which unlearning has to be operated [23, 16, 24]. This approach is also not practical in a federated setting, as the estimation of the noise amplitude requires the access to potentially sensitive clients information.

2.4 Federated Unlearning Schemes

To meet the practical requirements of real-world use of FL, we consider FU methods compatible with the following criteria: 1) no additional work has to be performed by the clients withdrawing the study, 2) no additional information beyond model parameters must be exchanged between clients and server, 3) no modification of data is allowed at client side. Following this consideration, we identified 4 state-of-the-art FU approaches for our benchmark [6, 7, 8, 9], and excluded a number of methods not satisfying the criteria [25, 26, 27, 28]. We provide a brief description of the selected approaches, and refer to the related publications for additional details.

FINE-TUNING. Fine-tuning of model parameters after excluding client c is a standard FU baseline [16, 18]. Nevertheless, although fine-tuning can be shown to satisfy the utility criterion, it does not formally guarantee unlearning [9].

FEDACCUM [6]. FEDACCUM implements an heuristic based on the removal of the contribution of the parameters $\{\boldsymbol{\theta}_c^t\}_{t=1}^T$ provided by client during FL optimization. Similarly to SCRATCH, the server performs the training procedure of equation (2), while however integrating in the optimization routine the existing contributions of the remaining clients $\{\boldsymbol{\theta}_i^t\}_{i \in I \setminus c, t \in \{1,\dots,T\}}$.

FEDERASER [6]. This approach performs a retraining from scratch, by however scaling the new contributions by the norm of the ones computed to obtain $\tilde{\boldsymbol{\theta}}$, i.e. $\tilde{p}_i = p_i \left\|\theta_i^t\right\| / \left\|\tilde{\theta}_i^t\right\|$, every Δt aggregation rounds. FEDERASER

is faster than SCRATCH by requiring a smaller amount of local work from the remaining clients, and less aggregations.

Unlearning with Knowledge Distillation (UKD) [7]. UKD consists in subtracting to $\tilde{\boldsymbol{\theta}}$, the model trained with every client in I, all the contributions of client c, i.e. $h(\tilde{\boldsymbol{\theta}}, c) = \tilde{\boldsymbol{\theta}} - p_i \sum_{t=1}^{T} \boldsymbol{\theta}_c^t$. The server subsequently applied fine-tuning to optimize the similarity of the predictions $\tilde{\boldsymbol{\theta}}$ and $h(\tilde{\boldsymbol{\theta}}, c)$ on a control dataset owned by the server. With UKD, no client needs to participate to the unlearning phase, while it is required the use of a dataset owned by the server.

Projected Gradient Ascent (PGA) [8]. This FU scheme unlearns client c by performing a succession of projected gradient ascents (PGA) on $\tilde{\boldsymbol{\theta}}$ to achieve low performance of $h(\tilde{\boldsymbol{\theta}}, c)$ on the dataset of client c. While PGA requires the dataset of client c to unlearn it, we consider this FU scheme to show that minimizing the performances of client c is not sufficient to unlearn it.

Informed Federated Unlearning (IFU) [9]. IFU consists in tracing back the history of global models $\{\boldsymbol{\theta}^t\}_{t=1}^{T}$ and restart FL from a specific round t^*, which is identified by fixing a cutoff on the magnitude of the contributions of client c, measured as $\sum_{t<t^*} \left\| \boldsymbol{\theta}_c^{t+1} - \boldsymbol{\theta}^t \right\|$. Prior to retraining, the global model $\boldsymbol{\theta}^{t^*-1}$ is perturbed with Gaussian noise according to a given unlearning budget (ϵ, δ), with an analogy to randomized mechanisms in differential privacy [29].

3 Experiments

We introduce the federated dataset used for prostate segmentation in Section 3.1, and verify in Section 3.2 the efficiency of the FU schemes introduced in Section 2.4, based on the criteria introduced in Section 2.2. The code for the experiments is publicly available[4].

3.1 Federated Dataset for Prostate segmentation

Our benchmark consists of a FL application on prostate segmentation from a large collection of magnetic resonance images (MRIs) dataset. We consider three publicly available image segmentation benchmarks (Decathlon [30], Promise12 [31], and ProstateX [32]) to create a cross-silo federated partitioning composed by four centers ($C1$ to $C4$), where data split are based on specific image acquisition properties, as summarized in Table 1.

Decathlon [30] is a dataset composed of medical images of ten different organs including prostate. We allocate to C_1 the 32 publicly available Decathlon MRIs of prostate segmentation acquired with different scanners. We merge the masks of the peripheral and transition zone to define the prostate mask.

Promise12 [31] was created for the Prostate MRI Segmentation challenge of 2012. We partition the 50 published training data samples based on the acquisition method: images acquired with and without endorectal coil (respectively allocated to C_2 and C_3).

[4] https://github.com/Accenture/Labs-Federated-Learning/tree/FU_prostate_segmentation

Table 1: Description of the four centers used for FL and the respective training and testing DSC, obtained with the model trained with the four of them.

ID	Samples	Dataset	Description	DSC Train	DSC Test
C_1	32	Decathlon	Full Dataset	91.8(0.37)	87.4(1.4)
C_2	23	Promise12	With Endorectal Coil	96.3(0.15)	81.8(2.7)
C_3	27	Promise12	W\o Endorectal Coil	95.8(0.14)	84.1(5.7)
C_4	184	ProstateX	With Scanner Skyra	96.1(0.22)	84.4(5.8)

ProstateX [32] is a collection of MRIs from different medical studies acquired with two different scanners (Skyra and Triotim, both from Siemens) and segmentation masks provided for 189 of them [33]. We ignore the five data points obtained with Triotim and allocate the remaining 184 images to C_4.

We note that the images in C_2 are the only ones acquired by using the endorectal coil, thus introducing a specific bias for this center. Prostate MRIs were resized to a resolution of $320 \times 320 \times 16$. For each center, we randomly select 80% of its data samples to create a training set and allocate the remaining 20% to a testing set. FedAvg was used to optimized the federated training problem by optimizing a UNET [34] to maximise the dice score (DSC). To ensure generality of the results, we consider 5 random federated splits of the data and 5 different model initialization for the FL process. Hence, mean and standard error of the results reported in this section are estimated across 25 learning and unlearning scenarios. We detail in Appendix A the tuning of the hyperparameters for dropout value, learning rate, and amount of local work. We also detail the implementation of each FU scheme.

The model obtained when training with FL and the four centers has the performances summarized in Table 1. As expected, C_2 is associated with the lowest testing DSC, reflecting the specific heterogeneity of the data in C_2.

3.2 FU Benchmark

The unlearning benchmark here considered consists in unlearning the contributions of center C_2, the only center with MRIs acquired with endorectal coil. We provide in Table 2 a quantification of the impact of the FU schemes on utility and unlearning criteria, when the server performs 500 server aggregations to unlearn center C_2 with each FU scheme. The utility of our FU application is the average testing DSC of the remaining centers (C_1, C_3, and C_4), while the unlearning capability is quantified by the DSC on the training data of center C_2.

As discussed in section 2.2, an optimal FU scheme should lead to a model with utility and unlearning capabilities as close as possible to the ones obtained with Scratch. Based on the results of Table 2, we note that not all the FU schemes provide acceptable unlearning and utility properties. In particular:

1. Fine-Tuning and FedAccum have high utility on the remaining centers but their unlearning criterion is more than 20% higher than Scratch. Figure 1 illustrates this result, where the predictive mask of the model obtained with

Table 2: Unlearning center C_2: FU utility and unlearning criteria described in Section 2.2. We note that only FedEraser and IFU are able to unlearn center C_2, while keeping high utility on the remaining ones.

FU Scheme	Scratch	Fine-Tun.	FedAccum	FedEraser	UKD	PGA	IFU
Utility	84.7(3.4)	85.0(3.6)	84.9(3.6)	84.6(3.6)	34.5(3.6)	85.1(3.7)	84.6(3.3)
Unlearning	62.2(3.5)	84.5(1.2)	84.4(1.4)	60.5(4.4)	24.0(2.9)	83.5(1.4)	58.3(4.2)

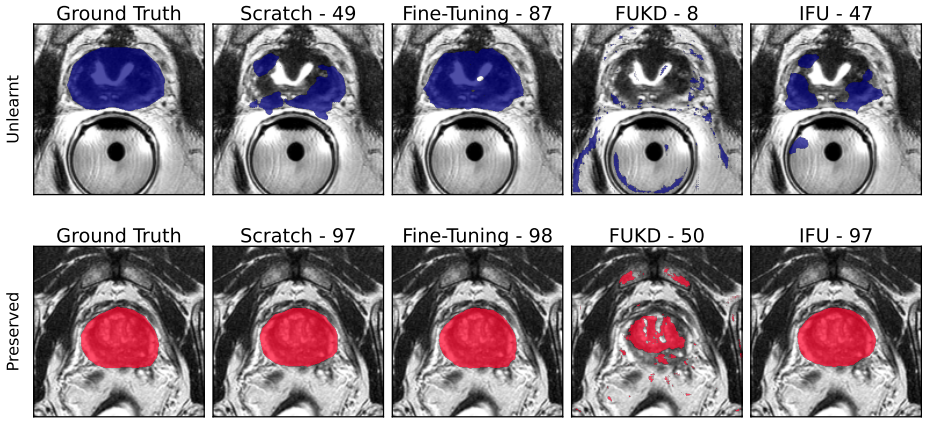

Fig. 1: Prediction Mask on a slice of a sample MRI from center C_2 (in blue) and from center C_3 (red), where FU is applied to the data of C_2. Additional results for all the FU schemes in Figure 2 and 3 are available in the appendix.

Fine-Tuning is almost identical to the ground truth (similar qualitative results are obtained for FedAccum, and are illustrated in Appendix A).

2. UKD has utility and unlearning performances respectively 50% and 30 % different from Scratch, which shows that the predictive accuracy of the model obtained with UKD is degraded on every center. Figure 1 shows that this method provides poor segmentation results for images from both C_2 (to be unlearnt) and C_3 (to be preserved).

3. FedEraser and IFU have identical utility to Scratch, while having only up to a 4% difference in unlearning capability. We see in Figure 1 that while the segmentation performance in C_3 is still of good quality, the correct unlearning of C_2 leads to poor segmentation results, similar to those obtained with Scratch. The slight difference between unlearning performances for IFU, FedEraser, and Scratch is likely due to the variability between model parameters as a result of the associated optimization routine.

The ensemble of results shown in Table 2, Figure 1, and Appendix A, show that only FU schemes based on adaptive retraining (FedEraser and IFU) provide satisfactory unlearning capabilities. On the contrary, the other approaches are either too conservative, thus leading to overly degraded models, or not ef-

Table 3: Optimization rounds when unlearning center C_2 with IFU for varying unlearning budget (3a) or with FedEraser (3b). Scratch requires 500 rounds.

	$\delta = 0.01$	$\delta = 0.025$	$\delta = 0.1$
$\epsilon = 0.1$	276(13)	271(15)	271(14)
$\epsilon = 1$	298(18)	295(16)	298(15)
$\epsilon = 10$	259(18)	251(18)	228(17)

(a) FL iterations (mean, std) required by IFU to unlearn center C_2 for varying unlearning budget parameters (ϵ, δ).

Δt	Utility	Unlearning	FL iter.
1	84.6(3.6)	60.5(4.4)	500(0)
2	84.7(3.3)	61.8(4.1)	250(0)

(b) Utility, unlearning, and FL iterations (mean, std) for FedEraser to unlearn center C_2 for varying frequency Δt.

fective, thus leading to poor unlearning properties. Concerning time efficiency, we report for IFU the amount of server aggregations needed for the resulting model to perform identically to Scratch after 500 server aggregations. For FedEraser, we vary the frequency Δt at which the server requires contributions from the remaining centers. Table 3 shows that both IFU and FedEraser achieve the desired utility and unlearning in a fraction of iterations needed by Scratch (resp. 2× and 1.9× faster). In addition of being able to unlearn center C_2, IFU provides statistical guarantees for the unlearning of the center C_2. We also provide in Table 6 of Appendix A the impact of the unlearning budget (ϵ, δ) associated to IFU on utility and unlearning. These results show that regardless of the unlearning budget, IFU reaches almost identical utility to Scratch, while with the increase in budget ϵ and/or δ, the model is associated with worse unlearning capabilities.

4 Conclusion

We provide in this work an investigation of FU in a practical collaborative segmentation task on prostate imaging data. We first define a benchmark from a collection of large available public dataset, to create a realistic scenario of data heterogeneity in cross-silo applications. We show that FU methods based on adaptive retraining (FedEraser and IFU) lead to optimal results in terms of trade-off between model utility, unlearning, and efficiency.

This study highlights a new perspective for the practical implementation of new data regulations in collaborative medical imaging applications. Future extensions of this work will be devoted to the investigation of FU in general medical applications, and to the assessment of the unlearning properties of the proposed methods, especially related to the definition of unlearning budget and parameters. In particular, since FedEraser does not come with specific guarantees on the effectiveness of the unlearning, we believe that further assessment of the unlearning capabilities of this approaches are needed.

References

[1] Paul Voigt and Axel Von dem Bussche. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 10(3152676):10–5555, 2017.

[2] Elizabeth Liz Harding, Jarno J Vanto, Reece Clark, L Hannah Ji, and Sara C Ainsworth. Understanding the scope and impact of the california consumer privacy act of 2018. *Journal of Data Protection & Privacy*, 2(3):234–253, 2019.

[3] Thanh Tam Nguyen, Thanh Trung Huynh, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. A survey of machine unlearning. *arXiv*, 2022.

[4] Theodora Brisimi, Ruidi Chen, Theofanie Mela, Alex Olshevsky, Ioannis Paschalidis, and Wei Shi. Federated learning of predictive models from federated electronic health records. *International Journal of Medical Informatics*, 2018.

[5] Santiago Silva, Boris A Gutman, Eduardo Romero, Paul M Thompson, Andre Altmann, and Marco Lorenzi. Federated learning in distributed medical databases: Meta-analysis of large-scale subcortical brain data. In *2019 IEEE 16th International Symposium on Biomedical Imaging*. IEEE, 2019.

[6] Gaoyang Liu, Xiaoqiang Ma, Yang Yang, Chen Wang, and Jiangchuan Liu. Federaser: Enabling efficient client-level data removal from federated learning models. In *2021 International Symposium on Quality of Service*, 2021.

[7] Chen Wu, Sencun Zhu, and Prasenjit Mitra. Federated unlearning with knowledge distillation. *arXiv preprint arXiv:2201.09441*, 2022.

[8] Anisa Halimi, Swanand Kadhe, Ambrish Rawat, and Nathalie Baracaldo. Federated unlearning: How to efficiently erase a client in fl? *arXiv*, 2022.

[9] Yann Fraboni, Richard Vidal, Laetitia Kameni, and Marco Lorenzi. Sequential informed federated unlearning: Efficient and provable client unlearning in federated optimization, 2022.

[10] Yann LeCun, Leon Bottou, Yoshua Bengio, and Patrick Ha. LeNet. *Proceedings of the IEEE*, 1998.

[11] Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009.

[12] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of ICCV 2015*, 2015.

[13] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *ICML 2017*, 2017.

[14] Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. In *ICLR 2020*, 2020.

[15] Ahmed Khaled, Konstantin Mishchenko, and Peter Richtarik. Tighter theory for local sgd on identical and heterogeneous data. In *AISTATS 2020*, 2020.

[16] Chuan Guo, Tom Goldstein, Awni Hannun, and Laurens Van Der Maaten. Certified data removal from machine learning models. In *ICML 2020*, 2020.

[17] Zachary Izzo, Mary Anne Smart, Kamalika Chaudhuri, and James Zou. Approximate data deletion from machine learning models. In *AISTATS 2021*, 2021.

[18] Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[19] Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Forgetting outside the box: Scrubbing deep networks of information accessible from input-output observations, 2020.

[20] Aditya Golatkar, Alessandro Achille, Avinash Ravichandran, Marzia Polito, and Stefano Soatto. Mixed-privacy forgetting in deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[21] Ananth Mahadevan and Michael Mathioudakis. Certifiable machine unlearning for linear models. *CoRR*, abs/2106.15093, 2021.

[22] Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. *NeurIPS 2019*, 2019.

[23] Seth Neel, Aaron Roth, and Saeed Sharifi-Malvajerdi. Descent-to-delete: Gradient-based methods for machine unlearning. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*. PMLR, 2021.

[24] Varun Gupta, Christopher Jung, Seth Neel, Aaron Roth, Saeed Sharifi-Malvajerdi, and Chris Waites. Adaptive machine unlearning. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2021.

[25] Yi Liu, Lei Xu, Xingliang Yuan, Cong Wang, and Bo Li. The right to be forgotten in federated learning: An efficient realization with rapid retraining. In *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*, 2022.

[26] Junxiao Wang, Song Guo, Xin Xie, and Heng Qi. Federated unlearning via class-discriminative pruning. In *Proceedings of the ACM Web Conference 2022*, 2021.

[27] Xiangshan Gao, Xingjun Ma, Jingyi Wang, Youcheng Sun, Bo Li, Shouling Ji, Peng Cheng, and Jiming Chen. Verifi: Towards verifiable federated unlearning. *arXiv*, 2022.

[28] Leijie Wu, Song Guo, Junxiao Wang, Zicong Hong, Jie Zhang, and Yaohong Ding. Federated unlearning: Guarantee the right of clients to forget. *IEEE Network*, 36(5):129–135, 2022.

[29] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3–4):211–407, aug 2014.

[30] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature communications*, 13(1):4128, 2022.

[31] Vadim S Koshkin, Vaibhav G Patel, Alicia Ali, Mehmet A Bilen, Deepak Ravindranathan, Joseph J Park, Olesia Kellezi, Marcin Cieslik, Justin Shaya, Angelo Cabal, et al. Promise: a real-world clinical-genomic database to address knowledge gaps in prostate cancer. *Prostate cancer and prostatic diseases*, 2022.

[32] Samuel G Armato III, Henkjan Huisman, Karen Drukker, Lubomir Hadjiiski, Justin S Kirby, Nicholas Petrick, George Redmond, Maryellen L Giger, Kenny Cha, Artem Mamonov, et al. Prostatex challenges for computerized classification of prostate lesions from multiparametric magnetic resonance images. *Journal of Medical Imaging*, 5(4):044501–044501, 2018.

[33] Renato Cuocolo, Arnaldo Stanzione, Anna Castaldo, Davide Raffaele De Lucia, and Massimo Imbriaco. Quality control and whole-gland, zonal and lesion annotations for the prostatex challenge public dataset. *European Journal of Radiology*, 138:109647, 2021.

[34] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI 2015*. Springer, 2015.

A Additional Experiments and Experimental Details

Table 4: Hyperparameters fine-tuned to maximise the testing DSC when training with the four centers on a 5 folds cross-validation scenario, and then used for all our learning and unlearning scenario.

Description	Range	Best Value
Amount of Local Work	1 to 100	5
Amount of Server Aggregations	-	500
Batch Size	-	8
Local learning rate	0.0001 to 0.1	0.001
Dropout value	0 to 0.5	0.2

Table 5: Hyperparameters values for the different unlearning schemes.

Description	FU scheme	Range	Best Value
Unlearning budget parameter ϵ	IFU	$\{0.1, 1, 10\}$	1
Unlearning budget parameter δ	IFU	$\{0.01, 0.025, 0.1\}$	0.025
Amount of unlearning SGDs	PGA	-	100
Upper bound on the training DSC of C_2	PGA	-	0.12
Amount of local work for remaining clients	FedEraser	-	5

Table 6: Impact of the unlearning budget (ϵ, δ) on the difference in utility and unlearning obtained with IFU and Scratch, when unlearning center C_2.

	$\delta = 0.01$	$\delta = 0.025$	$\delta = 0.1$
$\epsilon = 0.1$	.54(.13)	.54(.14)	.54(.13)
$\epsilon = 1$	.29(.19)	.31(.18)	.28(.17)
$\epsilon = 10$	.42(.15)	.46(.15)	.55(.15)

(a) Utility

	$\delta = 0.01$	$\delta = 0.025$	$\delta = 0.1$
$\epsilon = 0.1$	$-4.2(4.6)$	$-3.7(4.7)$	$-3.9(4.5)$
$\epsilon = 1$	$-4.6(3.8)$	$-3.9(3.7)$	$-2.9(3.5)$
$\epsilon = 10$	2.5(5.2)	3.9(4.7)	6.0(4.3)

(b) Unlearning

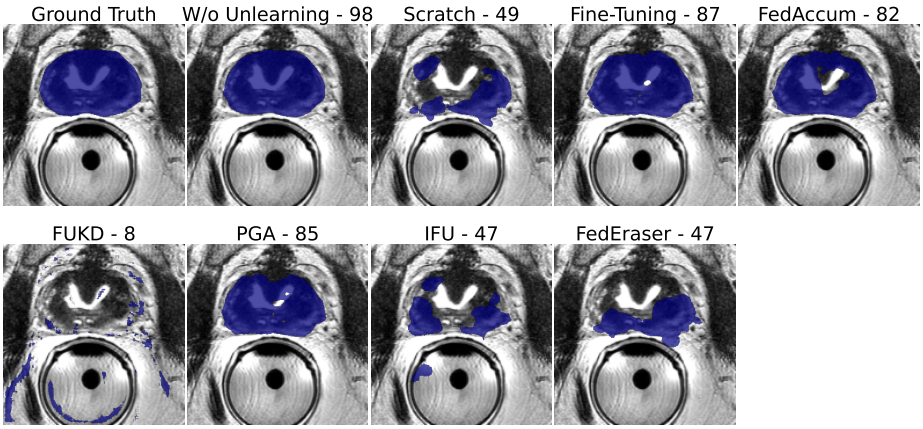

Fig. 2: Prediction Mask on a slice of a sample MRI from center C_2, where FU is applied to the data of C_2.

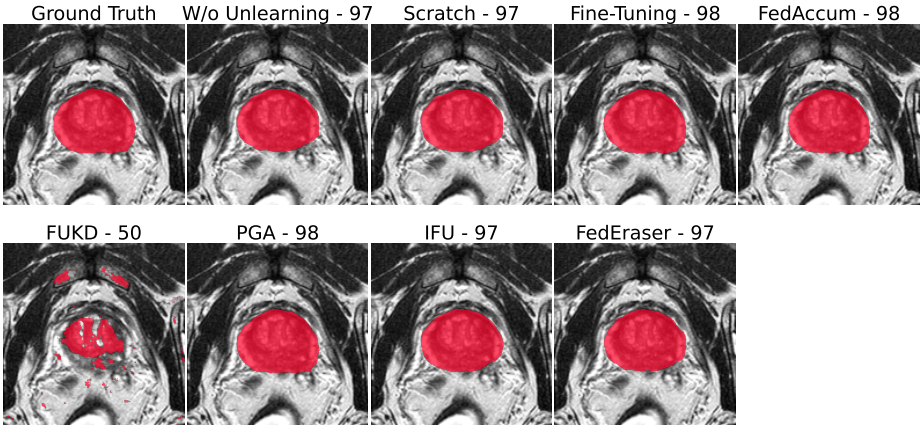

Fig. 3: Prediction Mask on a slice of a sample MRI from center C_3, where FU is applied to the data of C_2.