



HAL
open science

Benchmarking Collaborative Learning Methods Cost-Effectiveness for Prostate Segmentation

Lucia Innocenti, Michela Antonelli, Francesco Cremonesi, Kenaan Sarhan,
Alejandro Granados, Vicky Goh, Sebastien Ourselin, Marco Lorenzi

► **To cite this version:**

Lucia Innocenti, Michela Antonelli, Francesco Cremonesi, Kenaan Sarhan, Alejandro Granados, et al..
Benchmarking Collaborative Learning Methods Cost-Effectiveness for Prostate Segmentation. ECML
- PharML - Applications of Machine Learning in Pharma and Healthcare (Workshop at ECML PKDD
2023), Sep 2023, Turin (IT), Italy. 10.48550/arXiv.2309.17097 . hal-04357349

HAL Id: hal-04357349

<https://inria.hal.science/hal-04357349>

Submitted on 21 Dec 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

Benchmarking Collaborative Learning Methods Cost-Effectiveness for Prostate Segmentation

Lucia Innocenti^{1,2}, Michela Antonelli², Francesco Cremonesi¹, Kenaan Sarhan²,
Alejandro Granados², Vicky Goh², Sebastien Ourselin², and Marco Lorenzi¹

¹ Epione Research Group, Inria, Sophia Antipolis, France

² King's College London, London, UK

Abstract. Healthcare data is often split into medium/small-sized collections across multiple hospitals and access to it is encumbered by privacy regulations. This brings difficulties to use them for the development of machine learning and deep learning models, which are known to be data-hungry. One way to overcome this limitation is to use collaborative learning (CL) methods, which allow hospitals to work collaboratively to solve a task, without the need to explicitly share local data.

In this paper, we address a prostate segmentation problem from MRI in a collaborative scenario by comparing two different approaches: federated learning (FL) and consensus-based methods (CBM).

To the best of our knowledge, this is the first work in which CBM, such as label fusion techniques, are used to solve a problem of collaborative learning. In this setting, CBM combine predictions from locally trained models to obtain a federated strong learner with ideally improved robustness and predictive variance properties.

Our experiments show that, in the considered practical scenario, CBMs provide equal or better results than FL, while being highly cost-effective. Our results demonstrate that the consensus paradigm may represent a valid alternative to FL for typical training tasks in medical imaging.

Keywords: Collaborative Learning · Cost-Effectiveness · Prostate Segmentation.

1 Introduction

Prostate cancer is the most frequently diagnosed cancer in men in more than half of the countries worldwide [1]. While accurate prostate segmentation is crucial for effective radiotherapy planning [2], traditional manual segmentation is expensive, time-consuming, and dependent on the observer [3]. Automated or semi-automated methods are needed for efficient and reliable prostate segmentation [4], and deep learning is nowadays the main tool for solving the segmentation task [5]. Hospital data are highly sensitive and are difficult to collect in data silos for centralized training. This makes their use in notoriously data-hungry deep learning systems problematic. For this reason, collaborative learning (CL) is emerging as a powerful approach: it allows different decentralized entities to

collaborate in solving a task, and researchers are exploring ways to do this by keeping the local data private [6].

Federated learning (FL) [7] has gained great attention since the first apparition. FL solves a collaborative training problem in which a model is collectively optimized by different clients, each of them owning a local private dataset [8]. Through different training rounds, a server orchestrates local optimization and aggregation of trained parameters across clients. Since training data is kept on the client’s side, FL addresses the problems of data privacy and governance. Nevertheless, FL still poses several challenges in real-world applications [9, 10], consisting of 1) the sensitivity of the optimization result to the heterogeneity of system and data distribution across clients, and 2) the need for a large number of communication rounds, making communication cost a critical aspect. Moreover, from a practical perspective, FL systems are costly, since they are based on the setup and maintenance of complex computational infrastructures in hospitals, and thus require the availability of local resources and personnel [11–13].

Consensus-based methods (CBM) are a class of algorithms widely explored in machine learning, where the outputs from an ensemble of weak-predictors are aggregated to define a strong-predictor, outperforming the weak experts in terms of predictive robustness [14]. In medical imaging, CBM are often at the core of state-of-the-art approaches for image segmentation tasks [15, 16].

In this paper, we propose a comparison of these two different collaborative methods. Our specific focus is on collaborative prostate segmentation applied to magnetic resonance images (MRI). Differently from FL, in CBM independent models are locally trained by each client only once and, at testing time, a strong predictor is obtained by aggregating the output of the local models. Contrarily to FL, the setup of a CBM system in a hospital is straightforward, since no coordination in training is needed. Moreover, CBM provides data privacy and governance guarantees akin to FL, because no private information is shared during training, and model parameters are shared only once after training. Note that CBM has been coupled to FL training in previous works [17–23]. Nevertheless, most of these approaches are still based on distributed optimization, and thus they require setting up the whole FL infrastructure in hospitals, while the CBM we are analyzing here overtaken this limitation.

We present in this work a thorough benchmark of these models based on a cross-silo collaborative prostate segmentation task. The contributions of this paper are the following:

- We generate a distributed scenario based on natural data splits from a large collection of prostate MRI datasets currently available to the community, thus defining a realistic federated simulation.
- We define novel metrics to compare FL and CBM in terms of accuracy, robustness, cost-effectiveness, and utility.
- We apply the two CL approaches to this federated scenario and evaluate them in terms of accuracy and new-proposed metrics.

The paper is structured as follows. In Section 2 we present the data and the learning models used for the benchmark, i.e. federated learning and consensus-

based methods, and present the experiments and evaluation methods adopted in this work. Section 3 presents the experiment setting and results. Finally, Section 4 discusses our findings and future perspectives.

2 Benchmark definition

Starting from a large publicly available collection of data for prostate segmentation, we first define the federated setting by partitioning the data based on image acquisition characteristics and protocols. This allows us to obtain splits with controlled inter-center heterogeneity, thus simulating a realistic collaborative training scenario. We further define experiments to evaluate segmentation accuracy, cost-effectiveness, robustness to data heterogeneity, and utility for clients. Finally, we apply the differential privacy (DP) paradigm to different methods and we analyze how they respond to it.

2.1 Distributed Scenario

We gathered data provided by 3 major publicly available datasets on prostate cancer imaging analysis, and by 1 private dataset:

- **Medical Segmentation Decathlon - Prostate** [24] provides 32 prostate MRIs for training.
- **Promise12** [25] consists of 50 training cases obtained with different scanners. Of those, 27 cases were acquired by using an endorectal coil.
- **ProstateX** [26] contains prostate MRIs acquired by using two different scanners (Skyra and Triotim, both from Siemens). Segmentations of 194 cases are available [27].
- **Private Hospital Dataset** (PrivateDS) is composed of 36 MRIs collected by using a Siemens Aera scanner during a project on active surveillance for prostate cancer detection. An expert radiologist produced prostate masks. This dataset is used as an independent test set.

Datasets were split as in Table 1, to define centers characterized by specific image acquisition properties, thus allowing to obtain heterogeneous image distributions among centers. The common preprocessing pipeline applied to all the data comprised of flipping, cropping/padding to the same dimension, and intensity normalization. N4-bias-correction has also been applied to the data from Promise12 in N03 in order to compensate for the intensity artifacts introduced by the endorectal coil.

2.2 Collaborative Learning Frameworks

In our scenario we consider M hospitals, each having a local dataset $\mathcal{D}_i = \{z_{k,i}\}_{k=1}^{N_i}$. Given z , a volumetric MRI, and a vector of parameters θ , we define a segmentation problem in which a model g produces binary masks $h_z = g(z, \theta)$. Each hospital is a client indexed by $i \in [0, M]$, and the local training consists in solving the loss minimization problem, considering a loss function $f(\cdot)$.

Table 1: Description of the different centers here considered for the distributed learning scenario, derived by partitioning the four dataset Decathlon, ProstateX, Promise12, and PrivateDS.

ID	#Samples	Dataset	Subset Selection	Training	Test
N01	32	Decathlon	Full Dataset	Y	Y
N02	23	Promise12	No Endorectal Coil	Y	Y
N03	27	Promise12	Only Endorectal Coil	Y	Y
N04	184	ProstateX	Only Scanner Skyra	Y	Y
N05	5	ProstateX	Only Scanner Triotim	N	Y
N06	36	PrivateDS	Full Dataset	N	Y

Federated learning. FL is a collaborative optimization problem defined by:

$$\theta_g = \arg \min_{\theta} (\mathcal{L}(\theta)) \text{ s.t. } \mathcal{L}(\theta) := \sum_{i=1}^n p_i \mathcal{L}_i(\theta_i). \quad (1)$$

In FL, local losses are weighted by p_i , such that $\sum_{i=1}^n p_i = 1$, where the weights p_i are arbitrarily set, for example, based on the local dataset size. Different strategies on how to optimize the weights have been proposed in the literature, with the aim of mitigating the impact of data heterogeneity or client drift. In this paper, we consider the following FL strategies from the state-of-the-art:

- **FEDAVG**[28] is the backbone of FL optimization where, at round r , each client locally executes a number of stochastic gradient descent steps, and sends the partially optimized model θ_i^r to the server. The received models are weighted and averaged by the server into a global one, θ_g^{r+1} , which is then sent back to the clients to initialize the next optimization round. This process is repeated for R rounds until convergence.
- **FEDPROX**[29] tackles the problem of federated optimization with data heterogeneity across clients. This approach extends FEDAVG by introducing a proximal term to the local objective function to penalize model drift from the global optimization during local training. The proximal term is controlled by a trade-off hyperparameter, μ , through the following optimization problem:

$$\mathcal{L}_i(\theta)^r := \frac{1}{N_i} \sum_{k=1}^{N_i} \mathcal{L}(z_{k,i}, \theta_i^r) + \frac{\mu}{2} \|\theta_i^r - \theta_g^r\|^2. \quad (2)$$

Consensus-based methods. With CBM, a global federated ensemble of weak predictors is composed by aggregating the outputs from the different local models. During *training*, each client fully optimizes the segmentation model $g(z, \theta_i)$ on its local dataset D_i , by independently minimizing the local objective function \mathcal{L}_i . Trained local models are subsequently centralized and, for a given test image z' at *inference* time, the segmentation masks from all the local models are computed and aggregated by applying an ensembling strategy:

$$h_{z'} = \text{ensembling}(\{h_i(z')\}_{i=1}^M) \text{ s.t. } h_i(z') = g(z', \theta_i). \quad (3)$$

Among the different approaches to ensembling proposed in the literature [30], in this work we consider:

- **Majority Voting** [31] (MV) is a simple merging method that assigns to each voxel the label predicted by the majority of the local models.
- **STAPLE** [32] optimizes a consensus based on Expectation-Maximization (E-M) defined by the following iterative process:
 - the E-step computes a probabilistic estimate of the true segmentation, that is a weighted average of each local prediction;
 - the M-step assigns a performance level to each individual segmentation, which will be used as weights for the next E-step.
- **Uncertainty-Based Ensembling** (UBE) is based on weighted averaging of local decisions, in which the weights represent the uncertainty of each local model on the prediction task. As uncertainty can be quantified in different ways, in this work we adopt dropout [33] to compute a measure of the global uncertainty of each local model for the segmentation of a testing image z . In particular, here the uncertainty is computed as the total voxel-wise variance at inference time, defined as: $p_i = \sum_{x \in \Omega} \text{Var}(g(z, \theta_i))[x]$, where Ω is the set of voxels in z , $\text{Var}(\cdot)[x]$ is the sampling variance estimated from S stochastic forward passes of the model, computed at voxel x .

2.3 Experiments details

The benchmark is based on four experiments, quantifying a different aspect for comparison between different strategies. The experiments are characterized by the same baseline model used for segmentation, which is presented below.

Segmentation accuracy was quantified through 5-fold cross-validation across all nodes, by testing all training strategies for each unique combination of training/testing split. The final result was obtained by averaging across all splits.

Additionally, N05 and N06 from Table 1 were not used for training, and exclusively reserved for use as independent test set. The performance of the trained model was evaluated using the Dice Score (DSC) and a Normalised Surface Distance (NSD), following the guidelines from the Decathlon Segmentation Challenge [34].

We benchmarked the following strategies. *Local*: model trained only on the data from a single node, without aggregation; *Centralized*: model trained on the aggregated data from all the centers; *Federated*: federated training using both FEDAVG and FEDPROX as FL strategies; *Consensus*: ensembling of prediction using the CBMs strategies presented above.

Cost-effectiveness was investigated in terms of training and inference time and communication bandwidth [35–37]. For estimating the bandwidth we consider the amount of data exchanged through the network during the training phase; this value depends on the model size, that in our setting is constant among all the experiments, and the number of exchanges, which is strategy-dependent.

Model robustness was assessed to compare FL and CBM with respect to varying data heterogeneity across clients. To this end, we evaluated the change in

performance of the methods when removing N03 from the experiment. We expect a large variation in performance depending on the presence of N03, being this client the only one with images acquired with an endorectal coil, thus introducing large heterogeneity in the collaborative segmentation task.

Clients Utility refers to the evaluation of how beneficial it is for an individual client to participate in a collaborative method, and which specific method would bring the most value to that client. To determine this, we consider the accuracy of different models on various test sets.

Let's consider a client labeled as l . We have two models: a local model denoted as \mathcal{M}_l , and a collaborative model denoted as \mathcal{M}_c . We evaluate the performance of these models on two different test sets: \mathcal{T}_l , which is the local test set specific to client l , and \mathcal{T}_e , which is the union of all test sets excluding \mathcal{T}_l .

To compare the utility of the two models, we examine two metrics:

- variation in accuracy on the local test set: This is computed as the difference between the accuracy of the collaborative model on the local test set ($\text{DSC}_{\mathcal{M}_c, \mathcal{T}_l}$) and the accuracy of the local model on the same test set ($\text{DSC}_{\mathcal{M}_l, \mathcal{T}_l}$).
- variation in accuracy on the external test sets: This is calculated as the difference between the accuracy of the collaborative model on the combined external test sets ($\text{DSC}_{\mathcal{M}_c, \mathcal{T}_e}$) and the accuracy of the local model on the same combined external test sets ($\text{DSC}_{\mathcal{M}_l, \mathcal{T}_e}$).

By analyzing these two metrics, we can quantify the impact of using either the local or collaborative methods on both internal and external datasets. Ideally, a positive value for both metrics indicates that collaboration is beneficial for the client in all scenarios. However, it is more common to observe that collaboration improves model generalization but may affect local performance. Therefore, striking a balance between these two values is crucial.

In summary, the client's utility aims to determine the most advantageous approach for a client by comparing the accuracy variations of local and collaborative models on local and external test sets, respectively.

Privacy mechanisms such as differential privacy (DP) [38] have been proposed in the literature to quantify the privacy that a protocol provides and to train a model in a privacy-preserving manner. In the context of DP, the term "budget" refers to the amount of privacy protection available for the entire federated learning process, and represents the cumulative privacy loss allowed during the training phase. This budget is typically defined as a function of ϵ , where ϵ is used to control the strength of privacy guarantees for each round of federated learning updates. Here we compared the accuracy we can obtain by spending a fixed privacy budget ϵ while protecting different collaborative methods.

A common **baseline model** was defined to obtain comparable results across strategies. We employed a 3D UNet architecture with residual connections [5]. The training was based on the optimization of the DICE Loss, by using the ADAMW optimizer for all experiments [39]. The UNet implementation is avail-

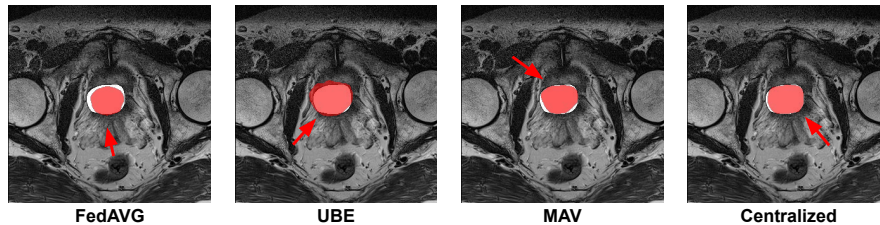


Fig. 1: A representation of the segmentation task on a sample image using different strategies. In white, the ground truth; in red, the segmentation provided by each training approach.

able in the MONAI library³. We fixed model hyper-parameters and maintained consistency in the amount of training, loss type, and optimizer used across all configurations. Hyperparameter search was performed by varying training parameters for all experiments (see Appendix Table 5) and selecting those performing averagely better on the local models, obtaining a learning rate of 0.001, a batch size of $B = 8$, and a dropout value of 0.3. All the experiments were executed using Fed-BioMed [40], an open-source platform that simulates the FL infrastructure. The code for running the experiments is available on the GitHub page of the author.

The number of epochs and rounds were defined using a standard strategy [41], which ensured comparable numbers of training steps among local and federated training for each node. Specifically, the number of rounds R for FL methods was defined as follows: $R = E \cdot N_T / M / B / s$, where E is the number of epochs required to train the model locally, $s = 20$ is the fixed number of local SGD steps, and N_T is the total number of samples in the training set.

3 Results

Segmentation Accuracy. Table 2 presents the average DSC among the 5-Fold evaluations obtained with the different collaborative learning strategies, while an illustrative example of the results on a sample image is available in Figure 1. The best results are indicated in **bold**. Similar results are obtained with the NSD metric and can be found in Appendix Table 1 and Table 2. Details about standard deviation among the K runs can be found in Appendix Table 3.

Overall, CBM obtain better or at least comparable results than FL: the last row in Table 2 shows that UBE is on average the best-performing method, but all the CBM provide very similar results. In general, distributed methods highly outperform local methods, which fail to generalize.

Cost-Effectiveness. We consider the total training time for FL and the longest time for local training across clients for CBM. Federated training is roughly three times longer than CBM training (~ 2 hours vs ~ 30 minutes).

³ <https://monai.io/index.html>

Table 2: Comparison of the 5-fold DSC obtained in the segmentation task by different training strategies.

	Local				Centralized	Federated		Consensus		
	N01	N02	N03	N04		FEDAVG	FEDPROX	UBE	STAPLE	MV
N01-test	0.86	0.64	0.49	0.44	0.92	0.85	0.70	0.89	0.83	0.84
N02-test	0.80	0.69	0.66	0.73	0.90	0.82	0.75	0.85	0.87	0.87
N03-test	0.64	0.72	0.75	0.44	0.83	0.70	0.75	0.73	0.75	0.76
N04-test	0.79	0.66	0.62	0.88	0.91	0.88	0.84	0.87	0.86	0.86
N05	0.57	0.68	0.71	0.73	0.77	0.71	0.67	0.72	0.68	0.68
N06	0.75	0.63	0.61	0.75	0.83	0.82	0.80	0.80	0.82	0.82
Average	0.73	0.67	0.64	0.66	0.86	0.80	0.75	0.81	0.80	0.80

Table 3: Comparison of costs of different training strategies in terms of training and inference time and training bandwidth.

	Local				Centralized	Federated		Consensus		
	N01	N02	N03	N04		FEDAVG	FEDPROX	UBE	STAPLE	MV
Train. time (min)	22	35	38	36	421	116	116	38	38	38
Inf. time (sec)	0.4	0.4	0.4	0.4	0.4	0.3	0.3	16.3	3.7	0.9
Train. Bandwidth (MB)	30	30	30	30	0	9600	9600	120	120	120

Among the CBM methods, UME is associated with the largest testing time, having to perform many inferences to estimate the uncertainty map. MAV is the most efficient and takes two times longer than the average FL (though still in the order of seconds). However, we note that testing time is a magnitude lower than training time, making its impact irrelevant in a real case application. The amount of exchanged data for FL is equal to $2 \cdot M \cdot m_s \cdot R$, where R is the number of rounds and m_s is the model size. For CBM, is only $M \cdot m_s$, resulting in a difference of $C \cdot m_s \cdot (2 \cdot R - 1)$. Considering the UNet used in the experiment, $m_s = 30MB$, the difference between FL and CBM is roughly of 9.25 GBytes.

Model Robustness. The performance of local models reported in Table 2 (panel "Local") allows to appreciate the heterogeneity across clients. As expected, N03 emerges as the client with the highest heterogeneity from this analysis, given the drop in testing performance of the models locally trained on the other clients. As shown in Appendix Table 4, CBM leads to an average absolute DSC variation of 1.7%, 2.4%, and 2.7%, for respectively UBE, MV, and Staple, as compared to the 3.1% and 5.7% DSC change respectively associated with FEDAVG and FEDPROX. A graphical representation of this property is available in Appendix Figure 1. This result denotes the improved robustness of CBM to clients' heterogeneity. The overall results obtained after removing N03 are compatible with those shown in Table 2, and confirm the positive performances of CBM as compared to FL.

Clients Utility. Figure 2 presents a comparison of the utility of different collaborative methods for the four clients in the experiment.

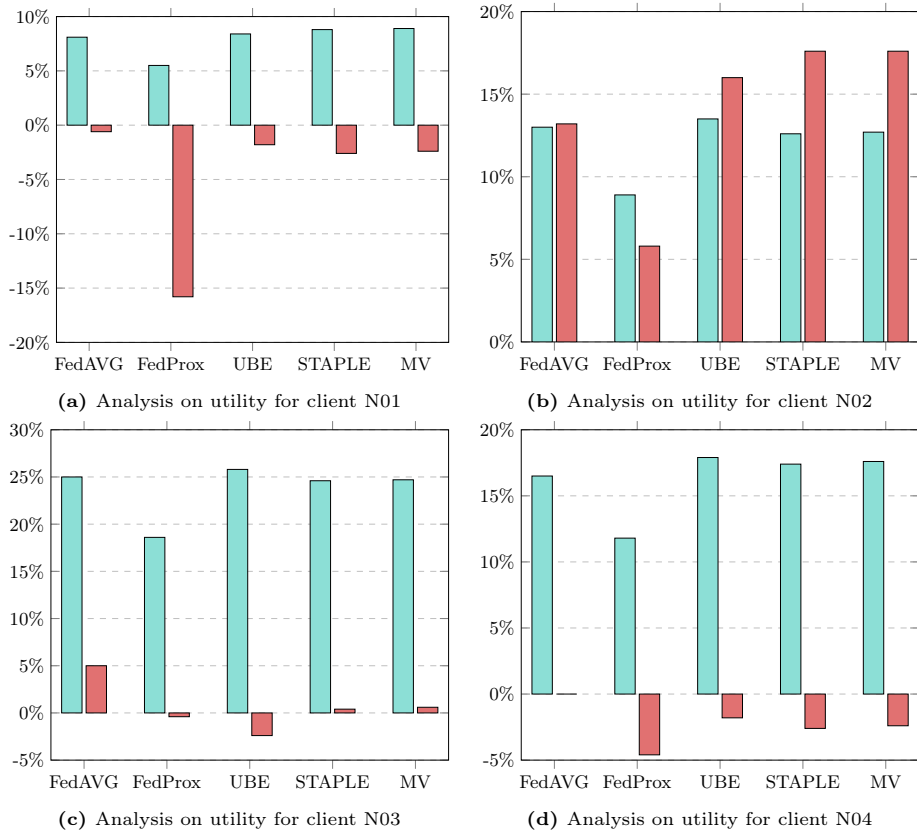


Fig. 2: The chart shows the utility of collaborative methods with respect to local models when used on the local test sets (red bar) or external test sets (blue bar) for each client indicated in the sub-captions. Each histogram corresponds to a different client. For all clients, collaborative methods improved generalization by a difference of up to 25%, while decreasing local performance by at most 15% and in some cases even improving it. A significant degree of heterogeneity can be observed in the impact on generalization and local performance among different test sets as well as different methods.

For all clients and all methods, collaborative methods lead to improvements in model generalization when evaluated on external test sets. This implies that collaborating with other clients helps to enhance the overall performance of the models on unseen data. Additionally, it is worth noting that even for small clients like N02, collaborative methods also result in improved local performance. This suggests that even clients with limited local data can benefit from participating in the collaboration. Surprisingly, even the largest client, N04, still experiences advantages by joining the collaboration. This indicates that size alone does not diminish the benefits of collaborative methods and that even clients

with substantial local datasets can gain value from collaboration. Overall, in this particular experiment, the performance of CBM is comparable to that of FL. However, UBE method consistently demonstrates the most substantial improvements across various metrics, making it the preferred choice among the collaborative methods evaluated.

Privacy mechanisms. The privacy analysis is performed in the framework of Rényi Differential Privacy (RDP) [42], a relaxation of the classical definition [43] allowing a convenient way to keep track of the cumulative privacy loss. This allows us to quantify the privacy budget ϵ corresponding to SGD optimization with parameters defined as for the baseline model of Section 2.3.

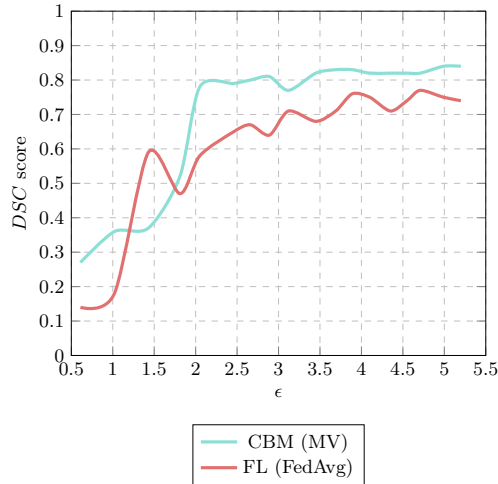


Fig. 3: The chart compares the accuracy reached by different methods when spending a privacy budget ϵ for differential privacy. The two compared methods are majority voting (MV) for CBM and federated averaging (FedAvg) for FL. CBM obtain on average better performances when ϵ is fixed, and it already reaches the plateau with $\epsilon \approx 3$.

Following [44], the DP Gaussian mechanism was defined with noise $\sigma = 4$. One can show that the privacy budget for obtaining the results presented in Table 2 is $\epsilon_{CBM} = 5.2$ in the CBM scenario, and $\epsilon_{FL} = 7.9$ in the federated one, denoting the lower privacy cost of CBM. CBM is also characterized by a lower privacy cost in relation to our chosen performance metric: DSC. We compared how DSC on unseen data evolved for the ensembling method MV and the FedAVG aggregation strategy when the privacy budget ϵ varied between 0.5 and 5.5. Figure 3 shows that the CMB method achieved a higher DSC than FL with a lower privacy budget: CBM reached a plateau at roughly $\epsilon = 3$ while FL reached a plateau only after $\epsilon > 4$, and at a lower DSC value.

4 Conclusions

In this paper, we proposed a realistic benchmark for collaborative learning methods for prostate segmentation. To this end, we used a collection of large public and private prostate MRI datasets to simulate a realistic distributed scenario across hospitals and we defined experiments and metrics to compare local training with different collaborative learning methods, namely FL and CBM, in terms of performances, cost-effectiveness, robustness and privacy of the models. For the considered scenario of cross-silo federated prostate segmentation, our results show that CBM represent a reliable alternative to FL in terms of performances, while being highly competitive in terms of robustness, and superior in cost-effectiveness when considering the practical implementation and required resources. Indeed, CBM avoid synchronization of training across hospitals, while the setup of an FL infrastructure is costly and time-consuming, and often prohibitive for typical hospital applications.

By simply sharing locally trained models and applying CBM to local predictions, we can rely on established theory from the state-of-the-art of multi-atlas segmentation to obtain competitive results at much less cost, as CBM avoid synchronization of training across hospitals.

Our preliminary results on privacy-preserving methods based on differential privacy show that CBM can guarantee a stronger level of privacy protection.

Moreover, secure aggregation techniques could be used at inference time for CBM in order to avoid sharing the whole model, adding another privacy layer to the framework. Other FL schemes could be included in our benchmark, such as SCAFFOLD [45] or FEDOPT [46], to better account for heterogeneity. Nevertheless, given previous benchmark results on similar medical imaging tasks [41], we do not expect a substantial change in the overall message of this study, especially concerning the comparison of cost-effectiveness between FL and CBM paradigms. Different consensus strategies could be implemented in the future, for example, to account for voxel-wise uncertainty across local models. The benchmark here proposed focuses on a cross-silo setup, typical of FL applications in hospitals proposed so far. Future investigations could extend our study to include a larger number of clients, thus allowing to better exploit the robustness guarantees associated with consensus strategies.

References

1. Winnie WY Sung et al. A cost-effectiveness analysis of systemic therapy for metastatic hormone-sensitive prostate cancer. *Frontiers in Oncology*, 11:627083, 2021.
2. Zia Khan et al. Recent automatic segmentation algorithms of mri prostate regions: a review. *IEEE Access*, 9:97878–97905, 2021.
3. Anne Sofie Korsager et al. The use of atlas registration and graph cuts for prostate segmentation in magnetic resonance images. *Medical physics*, 42(4):1614–1624, 2015.

4. Ronald A. Castellino. Computer aided detection (cad): an overview. *Cancer Imaging*, 5(1):17, 2005.
5. Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI*, 2015.
6. Brendan McMahan et al. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, 2017.
7. Nicola Rieke et al. The future of digital health with federated learning. *NPJ digital medicine*, 3(1):1–7, 2020.
8. Chen Zhang et al. A survey on federated learning. *Knowledge-Based Systems*, 216:106775, 2021.
9. Tian Li et al. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020.
10. Peter Kairouz et al. Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(1-2):1–210, 2021.
11. Wenqi Li et al. Privacy-preserving federated brain tumour segmentation. In *Machine Learning in Medical Imaging: 10th International Workshop, MLMI 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019*, 2019.
12. Prateek Chhikara et al. Federated learning meets human emotions: A decentralized framework for human–computer interaction for iot applications. *IEEE Internet of Things Journal*, 8(8):6949–6962, 2020.
13. Li Huang et al. Patient clustering improves efficiency of federated machine learning to predict mortality and hospital stay time using distributed electronic medical records. *Journal of biomedical informatics*, 99:103291, 2019.
14. Lars Kai Hansen and Peter Salamon. Neural network ensembles. *IEEE transactions on pattern analysis and machine intelligence*, 12(10):993–1001, 1990.
15. Truong Dang et al. Weighted ensemble of deep learning models based on comprehensive learning particle swarm optimization for medical image segmentation. In *2021 IEEE Congress on Evolutionary Computation (CEC)*, 2021.
16. Truong Dang et al. Ensemble of deep learning models with surrogate-based optimization for medical image segmentation. In *2022 IEEE Congress on Evolutionary Computation (CEC)*, 2022.
17. An Xu et al. Closing the generalization gap of cross-silo federated medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.
18. Naichen Shi et al. Fed-ensemble: Improving generalization through model ensembling in federated learning. *arXiv preprint arXiv:2107.10663*, 2021.
19. Jenny Hamer, Mehryar Mohri, and Ananda Theertha Suresh. Fedboost: A communication-efficient algorithm for federated learning. In *International Conference on Machine Learning*, 2020.
20. Jenny Hamer, Mehryar Mohri, and Ananda Theertha Suresh. Fedboost: A communication-efficient algorithm for federated learning. In *International Conference on Machine Learning*, 2020.
21. Neel Guha, Ameet Talwalkar, and Virginia Smith. One-shot federated learning. In *Workshop on Machine Learning on the Phone and other Consumer Devices (NeurIPs)*, 2018.
22. Tao Lin et al. Ensemble distillation for robust model fusion in federated learning. In *Advances in Neural Information Processing Systems 33*, pages 2351–2363, 2020.
23. Fernando E. Casado et al. Ensemble and continual federated learning for classification tasks. *Machine Learning*, pages 1–41, 2023.

24. Michela Antonelli et al. The medical segmentation decathlon. *Nature communications*, 13(1):4128, 2022. Data: <http://medicaldecathlon.com/>.
25. Geert Litjens et al. Evaluation of prostate segmentation algorithms for mri: the promise12 challenge. *Medical image analysis*, 18(2), 2014. Data: <https://promise12.grand-challenge.org/>.
26. Samuel G. Armato III et al. Prostatex challenges for computerized classification of prostate lesions from multiparametric magnetic resonance images. *Journal of Medical Imaging*, 5(4):044501–044501, 2018. Data: <https://prostatex.grand-challenge.org/>.
27. Renato Cuocolo et al. Quality control and whole-gland, zonal and lesion annotations for the prostatex challenge public dataset. *European Journal of Radiology*, 138:109647, 2021.
28. SK Warfield, KH Zou, and WM Wells. Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation. *IEEE Trans Med Imaging*, 2004.
29. Tian Li et al. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020.
30. Xibin Dong et al. A survey on ensemble learning. *Frontiers of Computer Science*, 14:241–258, 2020.
31. Josef Kittler et al. On combining classifiers. *IEEE transactions on pattern analysis and machine intelligence*, 20(3):226–239, 1998.
32. Simon K. Warfield, Kelly H. Zou, and William M. Wells. Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation. *IEEE transactions on medical imaging*, 23(7):903–921, 2004.
33. Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International conference on machine learning*, pages 1050–1059, 2016.
34. Michela Antonelli et al. The medical segmentation decathlon. *Nature communications*, 13(1):4128, 2022.
35. Bing Luo et al. Cost-effective federated learning design. *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*, 2021.
36. Hung T. Nguyen et al. Fast-convergent federated learning. *IEEE Journal on Selected Areas in Communications*, 39(1):201–218, 2020.
37. Mingzhe Chen et al. Convergence time optimization for federated learning over wireless networks. *IEEE Transactions on Wireless Communications*, 20(4):2457–2471, 2020.
38. Alexander Ziller et al. Medical imaging deep learning with differential privacy. *Scientific Reports*, 11(1):1–8, 2021.
39. Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *arXiv preprint arXiv:1711.05101*, 2017.
40. Santiago Silva et al. Fed-biomed: A general open-source frontend framework for federated learning in healthcare. *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning: Second MICCAI Workshop, DART 2020, and First MICCAI Workshop, DCL 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4–8, 2020, Proceedings 2*, 2020.
41. Jean Ogier du Terrail et al. Flamby: Datasets and benchmarks for cross-silo federated learning in realistic healthcare settings. *arXiv:2210.04620*, 2022.
42. Ilya Mironov. Rényi differential privacy. *2017 IEEE 30th computer security foundations symposium (CSF)*, 2017.
43. Cynthia Dwork et al. Calibrating noise to sensitivity in private data analysis. *Third Theory of Cryptography Conference*, 2006.

44. Martin Abadi et al. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.
45. Sai Praneeth Karimireddy et al. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, 2020.
46. Muhammad Asad, Ahmed Moustafa, and Takayuki Ito. Fedopt: Towards communication efficiency and privacy preservation in federated learning. *Applied Sciences*, 10(8):2864, 2020.

5 Supplementary Material

Table 4: Comparison of the NSD obtained in the segmentation task by different training strategies on the different test sets. Training strategies are grouped by approach.

	Local				Centralized	Federated		Consensus		
	N01	N02	N03	N04		FedAvg	FedProx	UBE	Staple	MV
N01-test	0.82	0.62	0.22	0.45	0.91	0.86	0.66	0.89	0.83	0.83
N02-test	0.64	0.63	0.37	0.74	0.91	0.84	0.71	0.85	0.80	0.80
N03-test	0.41	0.54	0.73	0.36	0.78	0.5	0.64	0.64	0.58	0.54
N04-test	0.67	0.65	0.27	0.92	0.94	0.89	0.76	0.87	0.83	0.83
N05	0.37	0.46	0.18	0.63	0.62	0.64	0.57	0.74	0.61	0.60
N06	0.39	0.52	0.26	0.68	0.75	0.66	0.55	0.71	0.60	0.58
Average	0.55	0.57	0.34	0.63	0.82	0.73	0.65	0.78	0.71	0.70

Table 5: Comparison of the deviation standard in the NSD obtained in the segmentation task by different training strategies on the different test sets. Training strategies are grouped by approach.

	Local				Centralized	Federated		Consensus		
	N01	N02	N03	N04		FedAvg	FedProx	UBE	Staple	MV
N01-test	0.09	0.32	0.31	0.14	0.06	0.30	0.07	0.04	0.05	0.05
N02-test	0.15	0.32	0.38	0.10	0.09	0.38	0.07	0.06	0.09	0.09
N03-test	0.28	0.33	0.07	0.19	0.34	0.23	0.24	0.23	0.33	0.28
N04-test	0.06	0.32	0.38	0.02	0.13	0.39	0.04	0.03	0.03	0.03
N05	0.21	0.25	0.38	0.16	0.09	0.33	0.15	0.05	0.21	0.21
N06	0.17	0.20	0.28	0.03	0.14	0.28	0.08	0.01	0.03	0.04
Average	0.16	0.29	0.30	0.11	0.14	0.32	0.11	0.07	0.12	0.12

Table 6: Comparison of the deviation standard in the DSC obtained in the segmentation task by different training strategies on the different test sets. Training strategies are grouped by approach.

	Local				Centralized	Federated		Consensus		
	N01	N02	N03	N04		FedAvg	FedProx	UBE	Staple	MV
N01-test	1.0E-01	3.3E-01	2.6E-01	1.9E-01	1.9E-02	6.4E-02	1.8E-01	3.4E-02	9.0E-02	8.3E-02
N02-test	6.1E-02	3.4E-01	1.5E-01	1.1E-01	2.4E-02	7.0E-02	1.3E-01	3.7E-02	2.9E-02	2.9E-02
N03-test	1.9E-01	1.7E-01	5.1E-02	1.4E-01	1.1E-01	1.6E-01	1.2E-01	1.3E-01	1.5E-01	1.5E-01
N04-test	7.0E-04	1.0E-01	3.1E-02	1.3E-04	3.8E-04	4.4E-03	2.3E-02	2.3E-02	1.8E-02	2.5E-04
N05	3.6E-02	3.3E-01	1.9E-01	6.7E-02	3.6E-02	3.9E-02	1.1E-01	3.3E-02	1.3E-02	1.1E-02
N06	1.8E-02	2.7E-02	2.6E-02	2.1E-02	6.0E-02	1.5E-02	1.6E-02	1.3E-02	2.1E-02	1.8E-02
Average	5.1E-02	2.8E-01	1.6E-01	9.3E-02	2.0E-02	4.4E-02	1.1E-01	4.4E-02	3.7E-02	3.1E-02

Table 7: Absolute differences in DSC score, among all the test sets, between the training with and without N03. The lower the difference, the higher the robustness to data heterogeneity for the strategy.

	FEDAVG	FEDPROX	UBE	STAPLE	MV
N01-test	6.0 %	4.2%	0.6 %	7.8 %	6.8 %
N02-test	4.6 %	7.2%	5.0 %	6.4 %	6.2 %
N03-test	0.6 %	1.8%	1.4 %	1.0 %	0.8 %
N04-test	2.2 %	5.4%	0.8 %	0.2 %	0.4 %
N05	2.0 %	8.8%	2.0 %	0.4 %	0.0 %
N06	3.2 %	4.6%	0.4 %	0.6 %	0.2 %
Average	3.1 %	4.3%	1.7 %	2.7 %	2.4 %

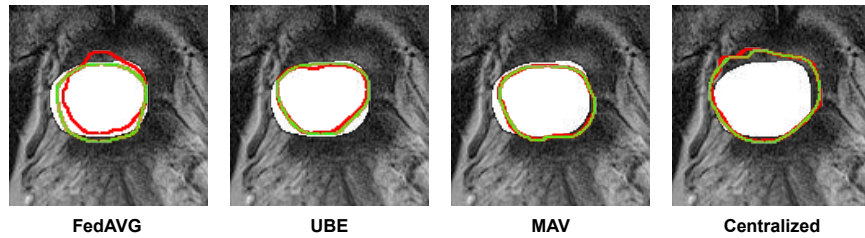


Fig. 4: A representation of the segmentation task on the same sample image using different strategies, comparing the model trained with, or without, client N03. In white, the ground truth; the red the contour obtained by the model trained with N03; in blue, the opposite.

Table 8: Hyperparameters and respective values explored during the tuning phase. Selected value in **bold**. The selection of dropout value was driven by the need to use it for the UBE method. Finally, we set the number of epochs for local training $E = \frac{1}{M} \sum_{i=1}^M E_i$, where E_i is the number of epochs computed by each local node while taking K gradient steps

Parameter	Values
Learning Rate	0.0001; 0.001 ; 0.01; 0.1; 1
Batch Size	4, 8 , 16
Dropout	0.1, 0.3 , 0.5
Local Steps	10, 15, 20 , 25
K	300, 400, 450 , 500