Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Differentially private federated deep learning for multi-site medical image segmentation

Alexander Ziller
Technical University of Munich
Dmitrii Usynin
Technical University of Munich
Nicolas W. Remerscheid
Technical University of Munich
Moritz Knolle
Technical University of Munich
Rickmer F. Braren
Technical University of Munich
Daniel Rueckert
Technical University of Munich
Georgios Kaissis (g.kaissis@tum.de)
Technical University of Munich

Article

Keywords:

Posted Date: April 4th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1478332/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Differentially private federated deep learning for multi-site medical image segmentation

Alexander Ziller[1,2,+], **Dmitrii Usynin**[1,2,3,+], **Nicolas W. Remerscheid**[1,+], **Moritz Knolle**[1,2], **Rickmer F. Braren**[2], **Daniel Rueckert**[1,3], **and Georgios Kaissis**[1,2,3,*]

[1]Institute for Artificial Intelligence and Informatics in Medicine, Department of Informatics, Technical University of Munich, 81675 Munich, Germany
[2]Institute of Diagnostic and Interventional Radiology, School of Medicine, Technical University of Munich, 81675 Munich, Germany
[3]Department of Computing, Imperial College London, London, United Kingdom
[*]g.kaissis@tum.de
[+]these authors contributed equally to this work

ABSTRACT

Collaborative machine learning techniques such as federated learning (FL) enable the training of models on effectively larger datasets without data transfer. Recent initiatives have demonstrated that segmentation models trained with FL can achieve performance similar to locally trained models. However, FL is not a fully privacy-preserving technique and privacy-centred attacks can disclose confidential patient data. Thus, supplementing FL with privacy-enhancing technologies (PETs) such as differential privacy (DP) is a requirement for clinical applications in a multi-institutional setting. The application of PETs to FL in medical imaging and the trade-offs between privacy guarantees and model utility, the ramifications on training performance and the susceptibility of the final models to attacks have not yet been conclusively investigated. Here we demonstrate the first application of differentially private gradient descent-based FL on the task of semantic segmentation in computed tomography. We find that high segmentation performance is possible under strong privacy guarantees with an acceptable training time penalty. We furthermore demonstrate the first successful gradient-based model inversion attack on a semantic segmentation model and show that the application of DP prevents it from divulging sensitive image features. Our code is available at https://github.com/TUM-AIMED/PrivateSegmentation.

1 Introduction

Training effective machine learning (ML) models in medical imaging is a data-driven problem, where model utility is typically directly dependent on the quantity and quality of data available during training. In recent works[1,2], federated learning (FL) has been introduced for the utilisation of multi-site clinical datasets. Herein, FL enables and encourages collaboration between sites, and facilitates ML-model building on larger pools of high quality, diverse and representative data while avoiding direct data sharing. Even though FL circumvents centralised data pooling, it should be considered foremost a technique for ascertaining data governance, and not a truly privacy-enhancing technology (PET), as it does not provide any formal notion of privacy in regards to the patient data. Therefore, FL-models are vulnerable to catastrophic privacy breaches through attacks such as model inversion[3–5] or membership inference[6] during model training by malicious parties. The utilisation of PETs such as differential privacy (DP) or encrypted algorithm training have been introduced to this end. PETs need to be tailored in the contexts of collaborative model training as participants in the consortium may have different requirements for the amount of information they are willing to disclose about their data. PETs applied in FL have effects on performance of the jointly trained model. and may result in decreased model utility (*privacy-accuracy-trade-off*), e.g. due to the addition of noise to the training process[7]. Minimising this trade-off is a complex, yet fundamental process, and has so far not been conclusively investigated in the area of medical imaging.

In the present work, we perform federated medical image segmentation under image-level differential privacy guarantees. Through detailed experiments, we demonstrate that the appropriate choice of architecture and training technique can enable excellent model performance despite rigorous privacy guarantees while minimising network overhead and computational demands.

2 Background and Related Work

Federated learning and Differential Privacy We use the term *Federated Learning* (FL,[8]) to denote a collaborative learning protocol in which a number of parties (*nodes/workers*) jointly train a neural network. Training occurs in *rounds* during which the *central server* sends the model to the nodes, local training occurs for a number of iterations, whereupon the updated models are aggregated by Federated (Gradient) Averaging[9]. We assume a *cross-silo topology*, based on a small number of centres with relatively large datasets. Moreover, we assume homogeneous node compute resources and constant node availability.

We use the following definition of *Differential Privacy* (DP) by[10]: For some randomised algorithm (=*mechanism*) $\mathscr{M}$, query function f executed over the data, all subsets S of its image, sensitive dataset D and its neighbouring dataset D', whereby D and D' differ by at most one record (that is, one individual's data), we say that $\mathscr{M}$ is (ε, δ)-differentially private (DP) if, for a (typically small) constant ε and $0 \leq \delta < 1$:

$$\mathbb{P}(\mathscr{M}(f(D) \in S)) \leq e^{\varepsilon}\mathbb{P}(\mathscr{M}(f(D') \in S)) + \delta. \tag{1}$$

The probability $\mathbb{P}$ is taken over the randomness of the mechanism $\mathscr{M}$. DP is, therefore, an attribute of an algorithm that makes its output approximately invariant to exclusion or inclusion of a data point. It quantifies an individuals contribution to the final outcome of the computation. When the difference between the contributions of multiple participants is minimal, it is not possible to reliably determine the presence or the absence of an individual. The parameter ε in this setting can be thought of as a *privacy budget*, that is, the total amount of privacy *spent* during the training of the model.

Prior Work Sheller and colleagues have demonstrated the utilisation of FL in the context of brain tumour segmentation[1,2]. However, neither work utilises PETs to provide privacy guarantees to the included patients. Li et al.[11] also showcase brain tumour segmentation using FL. The *Sparse Vector Technique* utilised in the study however only provides privacy guarantees to the *model's parameters* and not to the dataset records, which is not a meaningful notion of privacy. In comparison, DP-SGD, utilised in our study, provides the guarantees to each individual record (that is, either patient or item of the patient's data), providing the federation with a pragmatic, information-theoretic privacy-preserving solution instead. Fay et al.[12] utilise the *Private Aggregation of Teacher Ensembles* for brain tumour segmentation. This technique was originally developed for classification tasks and imposes strong assumptions on the learning process. Consequently authors could not demonstrate reasonable privacy guarantees in their study while still experiencing a steep utility penalty. Yang et al.[13] utilised FL for COVID-19 lesion segmentation in computed tomography (CT) but did not employ PETs. Sarma et al.[14] demonstrate FL segmentation of prostate volumes in magnetic resonance imaging (MRI), but also did not employ any PETs. Adnan et al.[15] employ both FL and DP in their study on federated pathology image processing, and FL with DP has also been demonstrated in our prior work[16], where we however limited ourselves to a classification use-case. Moreover, in previous work[17] we have demonstrated the utilisation of DP for medical image segmentation. However, this work did not incorporate FL.

A plethora of prior works have demonstrated the vulnerability of FL to adversarial attacks. An overview of these risks is described by Usynin et. al[18]. Concretely, works focused on *gradient-based model inversion attacks*[3–5] have demonstrated that the shared gradients alone are sufficient to reconstruct training data in the setting of federated learning. These prior works however focus solely on classification tasks, whereas no successful attacks against image segmentation architectures have been demonstrated to date.

2.1 Contributions

As witnessed from the survey of previous works above, although several studies have dealt with the topic of FL for medical image segmentation, our work is, to our best knowledge, the first to utilise differentially private stochastic gradient descent (DP-SGD) in addition to FL in order to provide both stringent image-level privacy guarantees and maintain high model utility in the setting of medical image segmentation. We summarise our main contributions below:

- We present an in-depth study on the application of DP to medical image segmentation by successfully training several segmentation model architectures in the federated setting. Contrary to the previous works, our implementation of differentially private training provides strict, provable guarantees with respect to each individual image, which allows the federation to obtain a meaningful measure of privacy.
- Our models trained with FL achieve comparable segmentation performance to centrally trained models while suffering only mild privacy-utility trade-offs.
- We demonstrate the first successful model inversion attack on semantic segmentation architectures, leading to the full reconstruction of input images for certain models. We thus provide evidence that, despite prevailing literature opinion, FL in itself is an insufficient technique for protecting patient privacy. We then empirically show that –consistent with its theoretical privacy guarantees– the addition of DP to model training completely thwarts such privacy-centred attacks.

3 Methods

3.1 Federated Training

For FL experimentation, we simulated a scenario in which three hospitals (=workers/nodes) are collaboratively training the neural network coordinated by a central server. For this, we split the dataset randomly by patient onto the three servers maintaining an equal number of patients per server. We conducted all experimentation using the PriMIA framework[16], a generic open-source software package for privacy-preserving and federated deep learning on medical imaging, which we adapted to semantic medical image segmentation.

3.2 Differentially Private Training

For deep neural network training, we utilised DP-stochastic gradient descent (DP-SGD)[19] which extends DP guarantees to gradient-based optimisation by clipping the L_2-norm of the per-sample gradients of each minibatch to a specific value and adding Gaussian noise of predetermined magnitude to the averaged minibatch gradients before performing an optimisation step. We utilise the *Rényi Differential Privacy Accountant*[20], an extension of the *moments accountant* technique by Abadi et al. for the privacy analysis of this algorithm, i.e. the calculation of privacy loss at each individual site in terms of ε. We note that the reported privacy guarantees are *record-level*, and not patient-level guarantees. Moreover, we regarded all datasets used as public for the purposes of experimentation such as hyperparameter searches, which is common practice in previous works[21]. DP training was performed under three *privacy regimes*, shown in Table 1. In the following, we will refer to these as *low*, *medium* and *high* privacy regimes. The lowest acceptable privacy guarantee of 11.5 was chosen similar to recent work showcasing the training of deep neural networks on ImageNet[21]. Finally, we note that the utilisation of Batch Normalisation layers is incompatible with DP training, as the running statistics of the layers are maintained non-privately. Hence, we converted these layers to Instance Normalisation Layers[22] which are DP-compatible. We note that this is not a limitation of this particular implementation, but rather an inherent incompatibility of Batch Normalisation layers with the formulation of DP itself.

(Hyper-)parameter	Low	Medium	High
Noise Multiplier	0.8	1.0	1.5
L_2 Clipping Norm	1.0	0.5	0.1
δ ($\times 10^{-5}$)	1.0	1.0	1.0
ε local	5.98	3.58	1.82
ε federated	11.5	7.08	3.54

Table 1. Definition of *privacy regimes* used in the study. The *Noise Multiplier* and L_2 *Clipping Norm* are hyperparameters of the DP-SGD algorithm. δ and ε are *record-level* guarantees for the local model and *per-site record-level* guarantees for the federated models. ε also designates the *privacy budget*, i.e. the total privacy loss at which model training is aborted.

3.3 Gradient-based Model Inversion

In order to empirically verify whether reconstruction of training data by an *Honest-but-Curious* (HbC) adversary with white-box model access[23] that participates in the training protocol is possible, we performed a model inversion attack utilising the gradients shared during Federated Averaging. We employ an approach similar to the *Deep Leakage from Gradients* (DLG)[3–5] attack to infer the data that generated the update. The attack relies on approximating the input image through a gradient descent-based perturbation of a randomly initialised noise matrix, based on a differentiable similarity metric to the gradient captured during model training. Unlike the original attack, designed for an image reconstruction in a classification context given a model update and the corresponding label, we provided the gradient update and a segmentation mask, which are used to reconstruct the input. We observed that the use of a segmentation mask of the victim by the adversary greatly improves the results of the reconstruction.

3.4 Model Architectures

All experiments were carried out using U-Net-like architectures[24] with modifications detailed in[25], introducing newer architectures as *backbones* to the encoder portion of the U-Net. Since network input/output overhead is a critical bottleneck for FL, we focused on architectures which provide a good balance between network size and performance on established benchmarks. Hence, we included MobileNet V2[26] and ResNet-18[27] as backbones. Moreover, we considered MoNet[28], a novel lightweight U-Net-like architecture with extremely few parameters specifically optimised for FL. Lastly, for comparability to the original U-Net, we utilised an eleven-layer VGG architecture with Batch Normalisation[29] (VGG11 BN). An overview of the number of parameters (i.e. model size) and Multiply/Accumulate operations (MACs) can be found in Table 2. We note that MoNet

relies on large receptive field dilated (atrous) convolutions which introduce a substantial number of operations despite the small network size. Moreover, the MobileNet-V2 architecture is optimised for CPU performance[30]. We therefore carried out timing experiments both on CPU and GPU (compare Table 4).

	MoNet	U-Net (MobileNet V2)	U-Net (ResNet-18)	U-Net (VGG11 BN)
Parameters	**390k**	6.63M	14.32M	18.26M
MACs	6.11G	**3.37G**	5.32G	14.45G

Table 2. Number of parameters and Multiply/Accumulate operations for a single forward pass (MACs) of the architectures used in the study. Parameter counts in 32-bit floating point units. k: $\times 10^3$, M: $\times 10^6$, G: $\times 10^9$. **Bold** denotes the lowest number of parameters or MACs, correspondingly.

3.5 Hyperparameter Optimisation

We performed hyperparameter optimisation on non-private baseline models in a decentralised manner over the entire federation to obtain suitable values for the learning rate, β parameters for the *Adam* optimiser, as well as translation, rotation and scale values for image augmentation. This corresponds to local hyperparameter optimisation prior to DP training to avoid repeated dataset interaction which would consume the privacy budget. For FL, the synchronisation rate, number of different image augmentations as well as whether or not the model parameters were weighted by the number of samples on the worker during aggregation were additionally optimised. The settings found through optimisation were used to train the corresponding models with DP.

3.6 Training data

All models were trained and evaluated on the MSD Liver segmentation task[31]. To mirror the fact that data is sparse in clinical settings, we limited the overall data to less than 55% (N=9 036 2D slices) of the whole dataset (N=16 768). The dataset was split into a 63% training set (N=5 187), 7% validation set (N=643) and 30% held-out testing set (N=2 563) ensuring that no patients overlap between the splits. All architectures were pre-trained on the MSD pancreatic CT segmentation dataset combined to improve convergence rate and offer equal starting conditions to all models.

4 Results

4.1 Segmentation results

Model evaluation results on the test set are listed in Table 3. We observed that models trained with FL are able to achieve performance on-par with locally trained models. In line with previous results[19], all models showed performance deterioration due to the utilisation of DP. Surprisingly, we found the performance penalty to be especially high on the ResNet-18 architecture, which did not converge at all in the FL setting with DP. Moreover, we consistently found the VGG-11 BN architecture to perform best. This finding suggests that a high parameter count alone is not the sole determinant of performance deterioration due to DP as suggested in previous work[32]. Instead, both the collaborative training setting and the architecture itself seem to contribute to this phenomenon. However, these findings await detailed future analysis. Moreover, the high variance of the per-image Dice score distributions for ResNet-18 and MobileNet V2 backbones suggest a disparate effect of DP on different images (see Figure 1), with the MoNet and VGG-11 BN architectures maintaining a more consistent performance in the privately trained local setting. Exemplary segmentation results are shown in Figure 2.

In Table 4, we compare the time required per epoch of training for the various models. We found the optimised, widely-used architectures to train more efficiently. In comparison, MoNet trades off computational efficiency for a much smaller size to alleviate network input/output constraints.

4.2 Model Inversion Attack

In this study we employ a novel gradient-based model inversion attack based on the works of[5]. In our setting the adversary starts with a randomly initialised (image, segmentation mask) pair and a captured model update. The optimisation task that is executed by the adversary involves perturbing the image to such extent that it (along with a corresponding segmentation mask) produces a gradient update similar to the one that the adversary captured. We utilise a cosine similarity as the cost function in this optimisation process. The procedure is repeated until either the loss starts diverging or the final iteration is reached.

We note that for our method the attacker is assumed to have access to a segmentation mask that corresponds to the sensitive image. We consider this to be a limitation of our approach and note that for smaller networks, utilisation of randomly initialised

Federated	Privacy	MoNet	U-Net (MobileNet V2)	U-Net (ResNet-18)	U-Net (VGG11 BN)
	None	93.41	95.41	94.67	**95.98**
	Low	86.66	87.61	82.20	**92.85**
	Medium	87.72	88.22	84.36	**92.63**
	High	87.22	87.04	83.66	**92.00**
✓	None	94.48	94.13	94.56	**95.08**
✓	Low	82.10	87.81	*DNC*	**90.93**
✓	Medium	83.25	88.76	*DNC*	**90.67**
✓	High	82.75	88.14	*DNC*	**90.27**

Table 3. Dice scores [%] of models trained locally and with FL under different privacy regimes. We note that ResNet-18 did not achieve convergence under FL and DP (*DNC*). **Bold** signifies the highest score in the category.

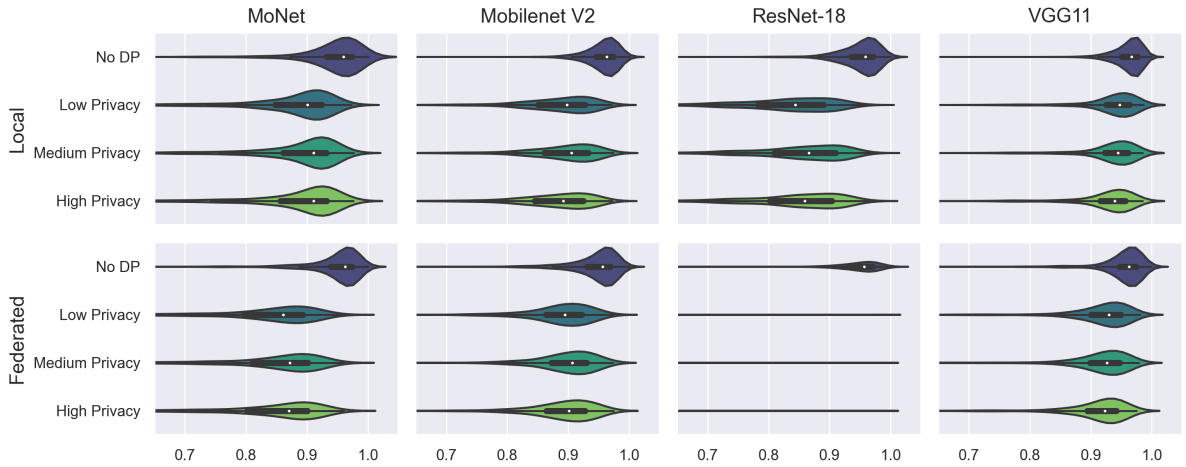

Figure 1. Kernel density estimation plots with inner box plots of Dice scores of the models trained locally (top) and federated (bottom) under different privacy regimes.

segmentation mask can provide the adversary with a suitable reconstruction result. However, for the purposes of medical imaging, larger models, such as VGG-11 or MoNet require a segmentation mask in order to allow the attacker model to converge and to produce an acceptable reconstruction result. We consider this limitation mild, as –in the segmentation setting– the target organ is known, and thus prior knowledge can be employed to improve the result of the attack, as discussed in[33].

Gradient-based model inversion attacks were performed in two settings: initially, gradients captured from FL training without DP were attacked. We then evaluated the attacks on the same architectures with the addition of DP. As presented in Figure 3, non-private models sharing unprotected updates during training risk their data being reconstructed in full. In comparison, models trained with the addition of DP yield no usable information to the inversion attack regardless of architecture. In general, architectural complexity seems to inhibit attack success to a greater degree than shown in previous work on classification[5].

5 Discussion and Conclusion

To the best of our knowledge, this is the first work to demonstrate DP-SGD-based collaborative model training in the context of semantic medical image segmentation. Our main conclusion is that the provision of rigorous privacy guarantees is possible in the FL setting while maintaining high model utility. One area that we highlight as a promising research direction is an investigation into the relationship between privacy-oriented modes of training and the selection of an optimal model architecture. We found that larger model architectures can be more robust to noise addition, whereas lightweight models can be advantageous in non-private FL settings. This highlights a promising research area that considers the application of privacy-preserving mechanisms in task-specific deployments in order to better tailor defence mechanisms to learning tasks with optimal utility preservation. Additionally, we outline a requirement for investigations into the field of *pragmatic* applications of DP, in

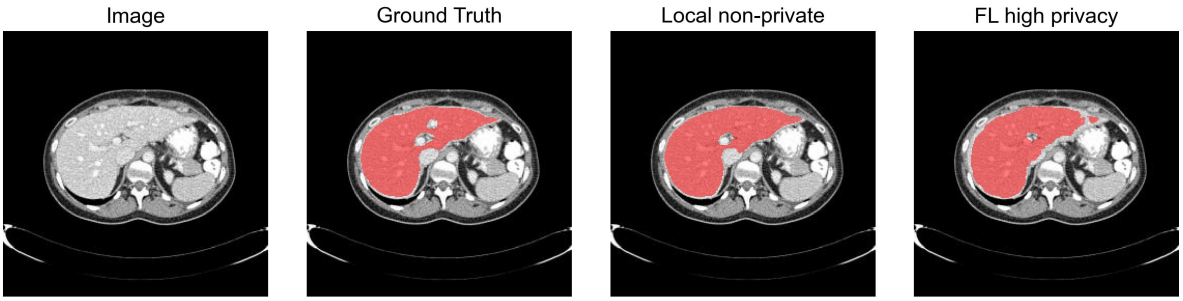

Figure 2. Exemplary segmentation results of the VGG11 BN U-Net architecture, trained in the local non-private setting vs. the FL *high privacy* regime.

Device	Federated	MoNet	U-Net (MobileNet V2)	U-Net (ResNet-18)	U-Net (VGG11 BN)
CPU		4186.50 ± 13.53	769.76 ± 1.68	$\mathbf{708.50 \pm 0.80}$	959.13 ± 0.43
CPU	✓	4281.64 ± 8.85	1200.32 ± 32.96	$\mathbf{892.62 \pm 12.34}$	1114.57 ± 7.56
GPU		92.84 ± 0.62	43.51 ± 0.32	$\mathbf{37.46 \pm 0.36}$	60.73 ± 0.66
GPU	✓	$\mathbf{175.11 \pm 1.61}$	458.30 ± 29.66	190.82 ± 5.22	192.80 ± 7.59

Table 4. Mean and standard deviation [s] for one epoch with equal settings for all models. CPU: Central Processing Unit. GPU: Graphics Processing Unit

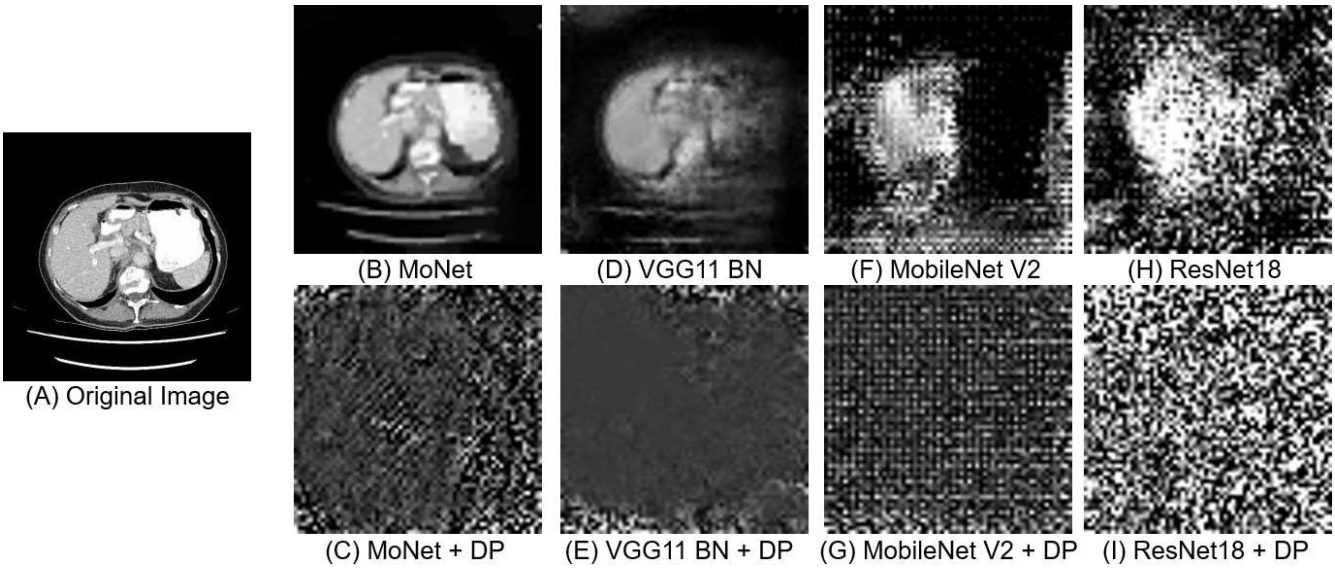

Figure 3. Gradient based reconstructions on the models used in the study. Top row: unprotected models. Bottom row: models with DP (Noise Multiplier 2.0, L_2 clipping norm 0.5).

order to allow privately trained models to be interpretable by machine learning researchers and facilitate the widespread utilisation of private collaborative model training. Our novel model inversion attack in the unprotected FL setting resulted in a catastrophic privacy breach, while only utilising a segmentation mask and a shared model update. We note that there also exist alternative adversarial methods that are applicable to collaborative learning. These include membership inference[34] as well as generator-based model inversion attacks[6], which can be performed simultaneously to the gradient-based methods. This highlights that FL alone is an insufficient privacy preservation mechanism in collaborative learning and should be regarded as a method for preservation of data ownership/governance which facilitates controlled data *access*.

We consider the following limitations of our work: as our method requires possession of a segmentation mask by the adversary, future work will include a natural relaxation of this requirement and the study of robust, scalable model inversion

attacks. We note that similarly to other model inversion implementations, supporting large batches of images is a non-trivial task, therefore we also outline this area as a potential future work direction.

In conclusion, we expect that our work will stimulate further research on privacy-preserving machine learning, essential to large scale, multi-site medical imaging analysis, in order to allow collaborative model training while mitigating associated privacy risks.

References

1. Sheller, M. J., Reina, G. A., Edwards, B., Martin, J. & Bakas, S. Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, 92–104, DOI: 10.1007/978-3-030-11723-8_9 (Springer International Publishing, 2019).

2. Sheller, M. J. *et al.* Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Sci. Reports* **10**, DOI: 10.1038/s41598-020-69250-1 (2020).

3. Zhu, L. & Han, S. Deep leakage from gradients. In *Federated Learning*, 17–31 (Springer, 2020).

4. Zhao, B., Mopuri, K. R. & Bilen, H. idlg: Improved deep leakage from gradients. *arXiv preprint arXiv:2001.02610* (2020).

5. Geiping, J., Bauermeister, H., Dröge, H. & Moeller, M. Inverting gradients–how easy is it to break privacy in federated learning? *arXiv preprint arXiv:2003.14053* (2020).

6. Shokri, R., Stronati, M., Song, C. & Shmatikov, V. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, 3–18 (IEEE, 2017).

7. Bagdasaryan, E. & Shmatikov, V. Differential Privacy Has Disparate Impact on Model Accuracy. *arXiv preprint arXiv:1905.12101* (2019).

8. Konečnỳ, J. *et al.* Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492* (2016).

9. 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*, 1273–1282 (PMLR, 2017).

10. Dwork, C., Roth, A. *et al.* The algorithmic foundations of differential privacy. *Foundations Trends Theor. Comput. Sci.* **9**, 211–407 (2014).

11. Li, W. *et al.* Privacy-preserving federated brain tumour segmentation. In *International Workshop on Machine Learning in Medical Imaging*, 133–141 (Springer, 2019).

12. Fay, D., Sjölund, J. & Oechtering, T. J. Decentralized differentially private segmentation with pate. *arXiv preprint arXiv:2004.06567* (2020).

13. Yang, D. *et al.* Federated semi-supervised learning for COVID region segmentation in chest CT using multi-national data from china, italy, japan. *Med. Image Analysis* **70**, 101992, DOI: 10.1016/j.media.2021.101992 (2021).

14. Sarma, K. V. *et al.* Federated learning improves site performance in multicenter deep learning without data sharing. *J. Am. Med. Informatics Assoc.* DOI: 10.1093/jamia/ocaa341 (2021).

15. Adnan, M., Kalra, S., Cresswell, J. C., Taylor, G. W. & Tizhoosh, H. R. Federated learning and differential privacy for medical image analysis. *Sci. Reports* **12**, DOI: 10.1038/s41598-022-05539-7 (2022).

16. Kaissis, G. *et al.* End-to-end privacy preserving deep learning on multi-institutional medical imaging. *Nat. Mach. Intell.* DOI: 10.1038/s42256-021-00337-8 (2021).

17. Ziller, A. *et al.* Medical imaging deep learning with differential privacy. *Sci. Reports* **11**, 1–8 (2021).

18. Usynin, D. *et al.* Adversarial interference and its mitigations in privacy-preserving collaborative machine learning. *Nat. Mach. Intell.* **3**, 749–758 (2021).

19. Abadi, M. *et al.* Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 308–318 (2016).

20. Mironov, I., Talwar, K. & Zhang, L. Rényi differential privacy of the sampled gaussian mechanism. *arXiv preprint arXiv:1908.10530* (2019).

21. Kurakin, A. *et al.* Toward training at imagenet scale with differential privacy. *arXiv preprint arXiv:2201.12328* (2022).

22. Dai, Z. & Heckel, R. Channel normalization in convolutional neural network avoids vanishing gradients (2019). 1907.09539.

23. Evans, D., Kolesnikov, V., Rosulek, M. *et al.* A pragmatic introduction to secure multi-party computation. *Foundations Trends Priv. Secur.* **2**, 70–246 (2018).

24. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241 (Springer, 2015).

25. Yakubovskiy, P. Segmentation models pytorch. https://github.com/qubvel/segmentation_models.pytorch (2020).

26. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. & Chen, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4510–4520 (2018).

27. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778 (2016).

28. Knolle, M. *et al.* Efficient, high-performance semantic segmentation using multi-scale feature extraction. *Plos one* **16**, e0255397 (2021).

29. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

30. Orsic, M., Kreso, I., Bevandic, P. & Segvic, S. In defense of pre-trained imagenet architectures for real-time semantic segmentation of road-driving images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12607–12616 (2019).

31. Simpson, A. L. *et al.* A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063* (2019).

32. Papernot, N., Chien, S., Song, S., Thakurta, A. & Erlingsson, U. Making the shoe fit: Architectures, initializations, and tuning for learning with privacy. *OpenReview* (2019).

33. Usynin, D., Rueckert, D. & Kaissis, G. Beyond gradients: Exploiting adversarial priors in model inversion attacks. *arXiv preprint arXiv:2203.00481* (2022).

34. He, Z., Zhang, T. & Lee, R. B. Model inversion attacks against collaborative inference. In *Proceedings of the 35th Annual Computer Security Applications Conference*, 148–162 (2019).

Acknowledgements

This research was supported by the UK Research and Innovation London Medical Imaging & Artificial Intelligence Centre for Value Based Healthcare. AZ and RB received funding from the German Research Foundation (DFG) SPP2177 “Radiomics: Next Generation of Biomedical Imaging”. RB received funding from the German Cancer Research Center Joint Funding Programs UPGRADE and SUBPAN. DU received funding from the TUM-Imperial Joint Academy of Doctoral Studies Programme. The funders played no role in the design of the study, the preparation of the manuscript or the decision to publish.

Author contributions statement

AZ implemented the software framework and performed model training. DU designed and analysed the reconstruction attacks. NWR implemented and performed model training. AZ, DU and GK wrote the initial manuscript. GK and RB provided medical expertise in the evaluation of results. All authors revised and edited the manuscript.

Data availability

Our code is available at https://github.com/TUM-AIMED/PrivateSegmentation. All used datasets are available from http://medicaldecathlon.com.

Additional information

The authors declare no competing interests. All methods were carried out in accordance with relevant guidelines and regulations.