Privacy-preserving medical image analysis

Alexander Ziller
Institute of Diagnostic
and Interventional Radiology
Institute for Artificial Intelligence
and Informatics in Medicine
Technical University of Munich
Munich, Germany
alex.ziller@tum.de

Jonathan Passerat-Palmbach
Department of Computing
Imperial College London
London, United Kingdom
j.passerat-palmbach@imperial.ac.uk

Théo Ryffel
INRIA, ENS, PSL University
Paris, France
they.ryffel@polytechnique.edu

Dmitrii Usynin
Department of Computing
Imperial College London
London, United Kingdom
dmitrii.usynin16@imperial.ac.uk

Andrew Trask
University of Oxford
Oxford, United Kingdom
andrew@openmined.org

Ionésio Da Lima Costa Junior
Universidade Federal de Campina Grande
Campina Grande, Paraíba, Brazil
ionesiojr@gmail.com

Jason Mancuso
Cape Privacy
New York, United States of America
jason@manc.us

Marcus Makowski
Institute of Diagnostic and
Interventional Radiology
Technical University of Munich
Munich, Germany
marcus.makowski@tum.de

Daniel Rueckert
Institute for Artificial Intelligence
and Informatics in Medicine
Technical University of Munich
Munich, Germany
daniel.rueckert@tum.de

Rickmer Braren
Institute of Diagnostic
and Interventional Radiology
Technical University of Munich
Munich, Germany
rbraren@tum.de

Georgios Kaissis
Institute of Diagnostic
and Interventional Radiology
Institute for Artificial Intelligence
and Informatics in Medicine
Technical University of Munich
Munich, Germany
g.kaissis@tum.de

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

Abstract

The utilisation of artificial intelligence in medicine and healthcare has led to successful clinical applications in several domains. The conflict between data usage and privacy protection requirements in such systems must be resolved for optimal results as well as ethical and legal compliance. This calls for innovative solutions such as privacy-preserving machine learning (PPML).

We present PriMIA (Privacy-preserving Medical Image Analysis), a software framework designed for PPML in medical imaging. In a real-life case study we demonstrate significantly better classification performance of a securely aggregated federated learning model compared to human experts on unseen datasets. Furthermore, we show an inference-as-a-service scenario for end-to-end encrypted diagnosis, where neither the data nor the model are revealed. Lastly, we empirically evaluate the framework's security against a gradient-based model inversion attack and demonstrate that no usable information can be recovered from the model.

1 Introduction

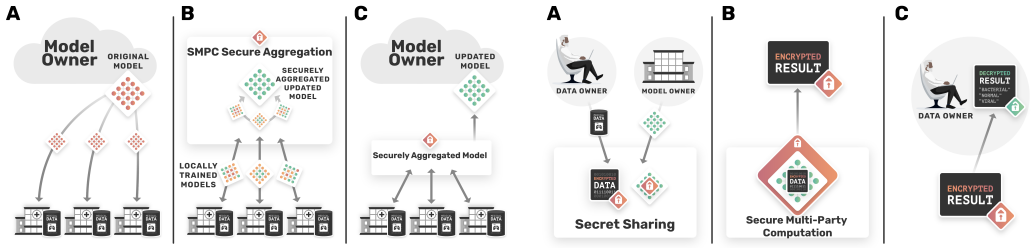
Machine Learning (ML) and Artificial Intelligence (AI) are recent approaches in biomedical data analysis, yielding promising results. These systems can assist clinicians to improve performance in tasks such as the early detection of cancers in medical imaging as shown in several applications [10, 1]. The most prominent challenge regarding AI systems is fulfilling their demands for large scale datasets. The collection of such datasets is often only achievable by multi-institutional or multi-national efforts. Data is typically anonymised or pseudonymised at the source site and stored at the institution performing data analysis and algorithm training [16]. The assembly and transmission of these datasets ethically and legally requires to have measures in place to protect patient privacy [12]. Moreover, information and control about the storage, transmission and usage of health data is a central patient right. Conventional techniques of anonymisation and pseudonymisation have been proven to not sufficiently protect from attacks against privacy [15, 11].

However, such concerns will likely not prevent increased data collection for future data-driven algorithms. Therefore, AI and ML methods will require innovations to sustainably reconcile data utilisiation and privacy protection. One of these approaches is the development of privacy-preserving machine learning (PPML). It aims to protect data security, privacy and confidentiality, while allowing the training and development of algorithms, as well as research in the setting of limited trust and/or data availability.

2 Privacy-preserving medical image analysis

In this work, we present PriMIA (Privacy-preserving Medical Image Analysis), a free-and-open-source software framework, which extends the PySyft/PyGrid ecosystem of open source privacy-preserving machine learning tools [13]. PriMIA adds functionality to allow medical imaging-specific applications, facilitating securely aggregated federated learning and secure inference. A broad variety of applications, including local and distributed model training, can be triggered from an accessible command-line interface. Furthermore, we include functional improvements increasing the framework's flexibility, usability, performance and security. These include gradient descent and federated averaging weighted by the dataset size at each site, local early stopping as a technique to counteract overfitting and *catastrophic forgetting*, hyperparameter optimisation over the whole confederation and the secure aggregation of dataset means/standard deviations.

Our work is clinically validated in a case study of paediatric chest radiography classification into one of three classes: normal (no signs of infection), bacterial pneumonia, viral pneumonia. We utilise a publicly available imaging dataset [8] for model training and two real-life datasets for model evaluation against clinical experts and ground-truth data. We aim to collaboratively train the model on three computation nodes without the disclosure of patient data in an *honest-but-curious* scenario. The topology of our federated learning system is shown in Figure 1a. The classification model is then made available for performing remote inference without revealing either it or the data in plain text (end-to-end-encrypted *inference-as-a-service* (Figure 1b)).



(a) Overview of the federated learning setup. At the beginning of training (**A**), the central server (*Model Owner*) sends the model to the computation nodes for training. Until convergence is achieved, the locally trained models are securely aggregated using secure multi-party computation (*SMPC Secure Aggregation*) and redistributed for a next round of training (**B**). Finally (**C**), the central model is updated and can be used for inference.

(b) Overview of the encrypted inference process. Initially (**A**) the data owner and model owner respectively obfuscate the data and algorithm using *secret sharing*. Inference is then carried out using secure multi-party computation (**B**), upon which the data owner receives an encrypted set of predicted labels, which only they can decrypt (**C**).

Figure 1: Details on the training and inference process in PriMIA.

3 Experiments, Results & Conclusion

In this work we evaluate the performance of a model trained using securely aggregated federated learning (*federated secure model*) against the same model without secure aggregation (*federated non-secure*) and a *locally trained* model. All models are ResNet-18 [6] architectures pre-trained on the ImageNet dataset [3]. We compare securely aggregated federated learning to models trained on locally accumulated data to assess the influence of encryption on model classification performance. Furthermore, we evaluate the performance of expert radiologists to obtain a human baseline as a benchmark for our models. Two separate test sets derived from clinical routine with a total of 497 images (Test Set 1 N=145, Test Set 2 N=352 images) are used. The ground truth is obtained based on clinical patient records. The evaluation measures employed are accuracy, sensitivity/specificity, receiver-operator-characteristic-area-under-the-curve (ROC-AUC) and the Matthews Correlation Coefficient (MCC) [9]. The latter is also the criterion by which we optimise the model hyperparameters during training. Model and expert classification results can be found in Table 1a. The *federated secure* model performs on par with the *federated non-secure* and *locally trained* models (Mc-Nemar test $p>0.05$, Table 1b). All models significantly outperform both human experts (Table 1b). Inter-observer/model agreement, evaluated using Cohen's kappa (κ) is moderate between experts, substantial between models and experts and almost perfect between models (Table 1c).

Furthermore, we empirically evaluate the models' resilience against inversion attacks [4], whereby a reconstruction of input data features is attempted. We apply the *improved deep leakage from gradients* method [5] to re-identify patients from all three models. For the quantification of the reconstruction success we use the mean squared error (MSE) and Frechet Inception Distance (FID) [7]. Both metrics are significantly higher in the setting of securely aggregated federated learning vs. *non-secure* federated learning and *locally trained* models, indicating the highest resistance to model inversion attacks. The average$\pm$standard deviation MSE was 2.26 ± 1.26, 2.62 ± 1.46, 3.22 ± 1.41; FID 2112 ± 1207, 2119 ± 1173, 2651 ± 1293 for *locally trained*, *federated non-secure* and *federated secure*, respectively (both one-way *analysis of variance* (ANOVA) $p<0.001$, MSE *locally trained* vs. *federated secure Student's t-test* $p<0.001$, *federated non-secure* vs. *federated secure Student's t-test* $p=0.03$; FID *locally trained* vs. *federated secure* and *federated non-secure* vs. *federated secure* both *Student's t-test* $p=0.03$). Lastly, we implement secure inference using the *Function Secret Sharing* [2] protocol expanded and adapted for neural networks [14] and observe a significant reduction in inference latency vs. the current state-of-the-art SecureNN [17] especially in the high network latency setting (35% average latency reduction vs. SecureNN at 10ms latency, 47% at 100ms, both Student's t-test $p<0.001$).

In conclusion, we present a free-and-open-source framework for securely aggregated federated learning and end-to-end encrypted inference in a medical context. We showcase a clinically relevant

Table 1: Statistical evaluation of model and expert classification performance

(a) Classification performance comparison of the *federated secure*, *non-secure* and *locally trained* models vs. the experts on the Validation set and on Test Sets 1 and 2. Sensitivity/Specificity metrics refer to normal/bacterial/viral, respectively. *ROC-AUC*: Receiver-operator-characteristic-area-under-the-curve (class-weighted average), *MCC*: Matthews Correlation Coefficient.

	Accuracy			*Sensitivity/Specificity*			*ROC/AUC*			*MCC*		
	Validation	**Test Set 1**	**Test Set 2**	**Validation**	**Test Set 1**	**Test Set 2**	**Validation**	**Test Set 1**	**Test Set 2**	**Validation**	**Test Set 1**	**Test Set 2**
Federated secure	0.88	0.88	0.89	0.98/0.86/0.78	0.88/0.88/0.91	0.89/0.88/0.91	0.90	0.92	0.92	0.83	0.83	0.83
Federated non secure	0.89	0.89	0.90	0.95/0.86/0.86	0.88/0.88/0.94	0.90/0.88/0.93	0.92	0.92	0.93	0.84	0.84	0.85
Locally trained	0.92	0.90	0.91	0.96/0.90/0.87	0.90/0.88/0.94	0.93/0.89/0.92	0.93	0.93	0.94	0.87	0.85	0.87
Expert 1	-	0.79		-	0.96/0.47/0.88	-	-	-	-	-	0.70	-
Expert 2	-	0.79		-	0.96/0.84/0.41	-	-	-	-	-	0.68	-

(b) McNemar test results on Test Set 1. **Bold** text signifies $p<0.05$

	Federated secure	*Federated non-secure*	*Locally trained*	*Expert 1*	*Expert 2*
Federated secure	-	1.0	0.47	**0.034**	**0.025**
Federated non-secure	1.0	-	1.0	**0.021**	**0.014**
Locally trained	0.47	1.0	-	**0.012**	**0.008**
Expert 1	**0.034**	**0.021**	**0.012**	-	0.882
Expert 2	**0.025**	**0.014**	**0.008**	0.882	-

(c) Cohen's κ between models and observers on Test Set 1.

	Federated secure	*Federated non-secure*	*Locally trained*	*Expert 1*	*Expert 2*
Federated secure	-	0.99	0.978	0.574	0.609
Federated non-secure	0.99	-	0.99	0.585	0.621
Locally trained	0.978	0.99	-	0.631	0.594
Expert 1	0.574	0.585	0.631	-	0.512
Expert 2	0.609	0.621	0.594	0.512	-

real-life case study and outperform human experts with a model trained in a federated and secure manner. Further research and development will enable the larger-scale deployment of our framework, the validation of our findings on diverse cross-institutional data and further the widespread utilisation of privacy-preserving machine learning techniques in healthcare and beyond.

Broader Impact

Our work's aim is to develop a technical solution providing objective guarantees of security and privacy to patients and algorithm developers in medical data science approaches. Concretely, we propose a framework for privacy-preserving machine learning on confidential patient data, as well as the provision of algorithmic diagnosis services under a premise of asset protection.

Our work advances the utilisation of privacy-enhancing tools in medically focused artificial intelligence. Such developments are warranted both from an ethical and a legal perspective: The creation of algorithms which assist clinical decision-making hinges on their training on diverse, large datasets to ascertain high performance and fairness and to counteract bias. However, the massive accumulation, transmission or release of private patient health data for this purpose removes sovereignty and control over the data from its owners and thus infringes on patient rights. Concurrently, algorithm owners wish to protect their models from theft or misuse during inference. Our work proposes a solution that will assert both: the safeguarding of data privacy and the protection of algorithms developed for clinical decision support. It promotes the training of models on larger datasets as well as making these models more broadly accessible.

Acknowledgments and Disclosure of Funding

Authors would like to thank Bennett Farkas for creating Figures 1 and 2 as well as the PriMIA logo, Patrick Cason and Héricles Emanuel for helping with PyGrid debugging, Matthias Lau for his input, the PySyft and PyGrid development teams for their foundational work and the OpenMined community for their scientific input, contributions and discussion.

Georgios Kaissis received funding from the Technical University of Munich, School of Medicine Clinician Scientist Programme (KKF), project reference H14. Rickmer Braren received funding from the German Research Foundation, SPP2177/1. Théo Ryffel received funding from the European Community's Seventh Framework Programme (FP7/2007-2013 Grant Agreement no. 339563 – CryptoCloud) and the French project FUI ANBLIC, and is co-founder of ARKHN. The funders played no role in the design of the study, the preparation of the manuscript or the decision to publish.

Authors declare no conflict of interest in relation to the study.

References

[1] Diego Ardila, Atilla P. Kiraly, Sujeeth Bharadwaj, Bokyung Choi, Joshua J. Reicher, Lily Peng, Daniel Tse, Mozziyar Etemadi, Wenxing Ye, Greg Corrado, David P. Naidich, and Shravya Shetty. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature Medicine*, 25(6):954–961, May 2019. doi: 10.1038/s41591-019-0447-x. URL `https://doi.org/10.1038/s41591-019-0447-x`.

[2] Elette Boyle, Niv Gilboa, and Yuval Ishai. Function secret sharing. In *Annual international conference on the theory and applications of cryptographic techniques*, pages 337–367. Springer, 2015.

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[4] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security - CCS 2015*. ACM Press, 2015. doi: 10.1145/2810103.2813677. URL `https://doi.org/10.1145/2810103.2813677`.

[5] Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. Inverting Gradients–How easy is it to break privacy in federated learning? *arXiv preprint arXiv:2003.14053*, 2020.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[7] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium, 2017.

[8] Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131, 2018.

[9] Brian W Matthews. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451, 1975.

[10] Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg C. Corrado, Ara Darzi, Mozziyar Etemadi, Florencia Garcia-Vicente, Fiona J. Gilbert, Mark Halling-Brown, Demis Hassabis, Sunny Jansen, Alan Karthikesalingam, Christopher J. Kelly, Dominic King, Joseph R. Ledsam, David Melnick, Hormuz Mostofi, Lily Peng, Joshua Jay Reicher, Bernardino Romera-Paredes, Richard Sidebottom, Mustafa Suleyman, Daniel Tse, Kenneth C. Young, Jeffrey De Fauw, and Shravya Shetty. International evaluation of an AI system for breast cancer screening. *Nature*, 577(7788):89–94, January 2020. doi: 10.1038/s41586-019-1799-6. URL `https://doi.org/10.1038/s41586-019-1799-6`.

[11] Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pages 111–125. IEEE, 2008.

[12] W Nicholson Price and I Glenn Cohen. Privacy in the age of medical big data. *Nature medicine*, 25(1): 37–43, 2019.

[13] Theo Ryffel, Andrew Trask, Morten Dahl, Bobby Wagner, Jason Mancuso, Daniel Rueckert, and Jonathan Passerat-Palmbach. A generic framework for privacy preserving deep learning. *arXiv preprint arXiv:1811.04017*, 2018.

[14] Théo Ryffel, David Pointcheval, and Francis Bach. Ariann: Low-interaction privacy-preserving deep learning via function secret sharing. *arXiv preprint arXiv:2006.04593*, 2020.

[15] Christopher G Schwarz, Walter K Kremers, Terry M Therneau, Richard R Sharp, Jeffrey L Gunter, Prashanthi Vemuri, Arvin Arani, Anthony J Spychalla, Kejal Kantarci, David S Knopman, et al. Identification of anonymous mri research participants with face-recognition software. *New England Journal of Medicine*, 381(17):1684–1686, 2019.

[16] Micah J. Sheller, Brandon Edwards, G. Anthony Reina, Jason Martin, Sarthak Pati, Aikaterini Kotrotsou, Mikhail Milchenko, Weilin Xu, Daniel Marcus, Rivka R. Colen, and Spyridon Bakas. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Scientific Reports*, 10(1), July 2020. doi: 10.1038/s41598-020-69250-1. URL `https://doi.org/10.1038/s41598-020-69250-1`.

[17] Sameer Wagh, Divya Gupta, and Nishanth Chandran. Securenn: Efficient and private neural network training. *IACR Cryptol. ePrint Arch.*, 2018:442, 2018.