nature machine intelligence

ARTICLES

https://doi.org/10.1038/s42256-021-00337-8

Check for updates

End-to-end privacy preserving deep learning on multi-institutional medical imaging

Georgios Kaissis[1,2,3,4,13], Alexander Ziller[1,2,4,13], Jonathan Passerat-Palmbach[3,4,5], Théo Ryffel[4,6,7], Dmitrii Usynin[1,2,3,4], Andrew Trask[4,8], Ionésio Lima Jr[4,9], Jason Mancuso[4,10], Friederike Jungmann[1], Marc-Matthias Steinborn[11], Andreas Saleh[11], Marcus Makowski[1], Daniel Rueckert[2,3] and Rickmer Braren[1,12] ✉

Using large, multi-national datasets for high-performance medical imaging AI systems requires innovation in privacy-preserving machine learning so models can train on sensitive data without requiring data transfer. Here we present PriMIA (Privacy-preserving Medical Image Analysis), a free, open-source software framework for differentially private, securely aggregated federated learning and encrypted inference on medical imaging data. We test PriMIA using a real-life case study in which an expert-level deep convolutional neural network classifies paediatric chest X-rays; the resulting model's classification performance is on par with locally, non-securely trained models. We theoretically and empirically evaluate our framework's performance and privacy guarantees, and demonstrate that the protections provided prevent the reconstruction of usable data by a gradient-based model inversion attack. Finally, we successfully employ the trained model in an end-to-end encrypted remote inference scenario using secure multi-party computation to prevent the disclosure of the data and the model.

The rapid evolution of artificial intelligence (AI) and machine learning (ML) in biomedical data analysis has recently yielded encouraging results, showcasing AI systems able to assist clinicians in a variety of scenarios, such as the early detection of cancers in medical imaging[1,2]. Such systems are maturing past the proof-of-concept stage and are expected to reach widespread application in the coming years as witnessed by rising numbers of patent applications[3] and regulatory approvals[4]. The common denominator of high-performance AI systems is the requirement for large and diverse datasets for training the ML models, often achieved by voluntary data sharing on behalf of the data owners and multi-institutional or multi-national dataset accumulation. It's common for patient data to be anonymized or pseudonymized at the originating institution, then transmitted to and stored at the site of analysis and model training (known as centralized data sharing)[5]. However, anonymization has proven to provide insufficient protection against re-identification attacks[6,7]. Therefore, large-scale collection, aggregation and transmission of patient data is critical from a legal and an ethical viewpoint[8]. Furthermore, it is a fundamental patient right to be in control of the storage, transmission and usage of personal health data. Centralized data sharing practically eliminates this control, leading to a loss of sovereignty. Moreover, anonymized data, once transmitted, cannot easily be retrospectively corrected or augmented, for example by introducing additional clinical information that becomes available.

Despite these concerns, the increasing demand for data-driven solutions is likely to increase health-related data collection, not only from medical imaging datasets, clinical records and hospital patient data, but also for example via wearable health sensors and mobile devices[9]. Hence, innovative solutions are required reconcile data and protect privacy. Secure and privacy-preserving machine learning (PPML) aims to protect data security, privacy and confidentiality, while still permitting useful conclusions from the data or its use for model development. In practice, PPML enables state-of-the-art model development in low-trust environments despite limited local data availability. Such environments are common in medicine, where data owners cannot rely on other parties' privacy and confidentiality compliance. PPML can also provide guarantees to model owners that their model will not be modified, stolen or misused, for example by its encryption during use. This lays the groundwork for sustainable collaborative model development and commercial deployment by alleviating concerns of asset protection.

Evidence from prior work

Recent work has shown the utility of PPML in biomedical science and medical imaging in particular. For instance, federated learning (FL) is a decentralized computation technique based on distributing machine learning models to the data owners (also referred to as computation nodes) for decentralized training instead of centrally aggregating datasets. It has been proposed as a method to facilitate multi-national collaboration while obviating data transfer. In the setting of the COVID-19 pandemic[10,11] FL was used to allow the retention of data sovereignty and the enforcement of local governance policies over data repositories. In medical imaging, recent studies[5,12] demonstrated that federated training of deep learning models on brain tumour segmentation or breast density classification performs on-par with local training and that it fosters the inclusion of data from more diverse sources, leading to improved generalization.

[1]Institute of Diagnostic and Interventional Radiology, Technical University of Munich, Munich, Germany. [2]Artificial Intelligence in Medicine and Healthcare, Technical University of Munich, Munich, Germany. [3]Department of Computing, Imperial College London, London, UK. [4]OpenMined. [5]ConsenSys Health, New York, NY, USA. [6]INRIA, ENS, PSL University, Paris, France. [7]Arkhn, Paris, France. [8]Centre for the Governance of AI, University of Oxford, Oxford, UK. [9]Universidade Federal de Campina Grande, Campina Grande, Paraíba, Brazil. [10]Cape Privacy, New York, NY, USA. [11]München-Klinik Schwabing, Munich, Germany. [12]German Cancer Consortium (DKTK), Partner Site Munich, Munich, Germany. [13]These authors contributed equally: Georgios Kaissis and Alexander Ziller. ✉e-mail: rbraren@tum.de

However, FL in itself is not a fully privacy-preserving technology. Previous studies[13,14] demonstrate that inversion attacks can reconstruct images from model weights or gradient updates with impressive visual detail. Moreover, in the setting of inference-as-a-service[15], exposure of the model to a non-trusted third party can enable model misuse or outright theft. Therefore, FL must be augmented by additional privacy-enhancing techniques to truly preserve privacy. For example, FL with secure aggregation (SecAgg) of weights or gradient updates or differential privacy (DP) can prevent dataset reconstruction attacks, and the utilization of secure multi-party computation (SMPC) protocols during model inference can protect the models in use. We provide an overview of these techniques in our previous work[16].

Aim and contributions

The clinical application of PPML in medical imaging requires the development of frameworks for security and privacy, and their validation on non-trivial clinical tasks. Here we present PriMIA, a free, open-source framework for end-to-end privacy-preserving decentralized deep learning on medical images. Our framework incorporates differentially private federated model training with encrypted aggregation of model updates as well as encrypted remote inference. Our contribution provides the following innovations:

- We demonstrate the training of a deep convolutional neural network (CNN) on the clinically challenging task of paediatric chest radiography classification using FL augmented with PriMIA's privacy-enhancing techniques over the public Internet.
- Our framework is compatible with a wide range of medical imaging data formats, easily user-configurable and introduces functional improvements to FL training (weighted gradient descent/federated averaging, diverse data augmentation, local early stopping, federation-wide hyperparameter optimization, DP dataset statistics exchange), increasing flexibility, usability, security and performance.
- We examine the computational and classification performance of models trained with and without privacy-enhancing techniques against models trained centrally on the accumulated dataset, personalized models trained on subsets of the data and against expert radiologists on unseen real-life datasets to evaluate various scenarios typical in medical imaging research.
- We assess the theoretical and empirical privacy and security guarantees of our framework and provide examples of applying a state-of-the-art gradient-based model inversion attack against the models under a number of training scenarios.
- Finally, we showcase the utilization of the trained model in a secure inference-as-a-service scenario without the disclosure of either the data or the model in plain text and demonstrate the improvements in inference latency of our SMPC protocol.

Library functionality

PriMIA was developed as an extension to the PySyft/PyGrid ecosystem of open-source PPML tools. PySyft (https://github.com/OpenMined/PySyft) is a Python framework allowing the remote execution of machine learning tasks (for example, tensor manipulation) and for encrypted deep learning by interfacing with common machine frameworks such as PyTorch. PyGrid provides server/client functionality for the deployment of such workflows on servers and edge computing devices. A detailed description of the generic functionality provided by these frameworks can be found in our previous work[17]. PriMIA builds upon this functionality towards medical-imaging-specific applications by being natively compatible with medical imaging data formats such as DICOM and able to operate on medical datasets of arbitrary modality and dimensionality (for example, computed tomography, radiography, ultrasound and magnetic resonance imaging). Outside of the above-mentioned PPML techniques, it offers solutions to common challenges in medical imaging analysis workflows, such as dataset imbalance, advanced image augmentation, federation-wide hyperparameter tuning functionality. Furthermore, it provides an accessible user interface for applications ranging from local experimentation on the user's machine to distributed training on remote compute nodes to facilitate the application of PPML best practices in medical consortia. The source code and documentation for the library and the publicly available data are provided at https://doi.org/10.5281/zenodo.4545599[18].

Case study, system design and threat model

We present a case study for the application of PriMIA on clinical data by training an 11.1 million parameter ResNet18 CNN[19] on the paediatric pneumonia dataset originally proposed by Kermany et al.[20] on cloud compute nodes over the public Internet with the aim of classifying paediatric chest radiographs into one of three categories: normal (no signs of infection), viral pneumonia or bacterial pneumonia. Pneumonia is a leading cause of paediatric mortality[21]. Chest radiography is routinely performed for differential diagnosis and therapy selection, but classifying paediatric chest radiographs is challenging. The case study is set up according to the following real-life scenario:

FL training phase. A confederation of three hospitals wishes to train a deep learning model for chest radiography classification. As they neither possess enough data on their own nor the expertise to train the model on this data, they enlist the support of a model developer to orchestrate the training on a central server. In the training phase, we refer to the hospitals holding patient data as the data owners. We utilize the term 'model' throughout the manuscript to refer to the structure and parameters of a deep neural network. We assumed an honest-but-curious threat model as defined previously[22] for the training phase. Here, participants trust each other to not actively undermine the learning protocol with utility degradation in mind, for example by actively supplying adversarial inputs or low-quality data (honest). However, individual participants and colluding groups of participants are assumed to actively attempt to extract private information from other participants' data (curious). Our framework's privacy-enhancing techniques are designed to protect from this behaviour, which we describe in detail in later sections. In brief, DP gradient descent[23] extends the guaranteed properties of DP to deep neural network training. Specifically, it bounds the worst-case privacy loss of individual patients in the datasets and provides privacy guarantees against model inversion/reconstruction attacks carried out against federation participants or against model owners at inference time. PriMIA implements DP for each FL node (local DP) to provide patient-level guarantees. Per-node privacy budgeting is performed using the Rényi Differential Privacy Accountant[24]. SMPC allows parties to jointly compute a function over a set of inputs without disclosing their individual contributions. During training, it is utilized to securely average the network weight updates (SecAgg). Additive secret sharing based on the SPDZ protocol[25] is used for SecAgg. The training phase is shown in Fig. 1. It concludes with all participants holding a copy of the fully trained final model.

Remote inference phase. Once fully trained, the model can be used for remote inference. In our case study, we assume that a different data owner, in this case a physician at a remote location holds some patient data and wants to receive an inference result for diagnostic assistance from the model. The inference service is provided over the internet by the model owner. The data and model owners do not trust each other and wish their data and model to remain private. PriMIA's SMPC protocol guarantees the cryptographic security of

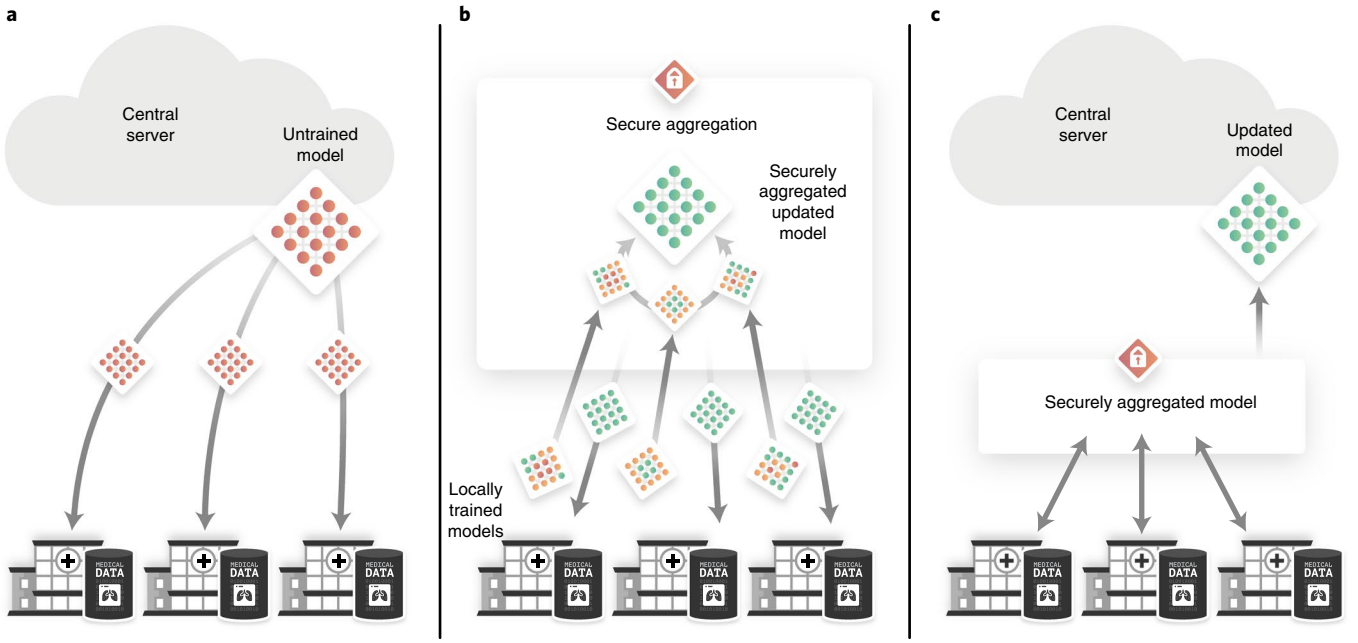

Fig. 1 | Overview of the FL training phase in the PriMIA case study. Three data owners (hospitals) wish to cooperate to train a model; a central server orchestrates the training. **a**, At the beginning of training, the central server sends the untrained model (red) to the computation nodes (hospitals/data owners) for training. **b**, Until convergence is achieved, the models are trained locally at each hospital. Intermittently, the models (coloured) are securely averaged (SecAgg). The SecAgg procedure occurs only between the three data owners. The SMPC protocol guarantees that the individual models cannot be exposed by other participants. After SecAgg, the updated model (green) is redistributed for another round of training. **c**, After the final iteration, the central model is updated with the (now fully trained) securely aggregated model (green) and can be used for inference.

both the model and the data in the inference phase. The AriaNN framework described in our previous work[26] is used, which we have adapted to end-to-end encrypted inference.

A common SMPC technique[25] is the utilization of cryptographically secure random numbers (cryptographic primitives) generated ahead of time (so-called offline phase) to accelerate certain computations. The trusted system (for example, a hardware device) providing these primitives is referred to as a cryptographic provider and is not involved in the actual inference procedure (online phase), nor does it ever come in contact with any party's data. In fact, a 'stockpile' of cryptographic primitives can be provided to the protocol participants ahead of time to be used up over multiple inference procedures. The encrypted inference process is summarized in Fig. 2.

Classification performance

We trained FL models without SecAgg or DP (DP-/SecAgg-), with SecAgg only (DP-/SecAgg+) and with both techniques (DP+/SecAgg+). Furthermore, we trained a model on the entire dataset pooled on a single machine (centrally trained) and separate models on the individual data owners' subsets of the dataset (personalized). The centrally trained model represents the centralized data sharing scenario described in the introduction. The personalized models each represent a single institution training exclusively on their own data, a typical case in current medical imaging research workflows. FL aims to enable the training of models that are better than personalized training and—ideally—as good as the centrally trained model.

We tested the classification performance of the models on the validation set and against the classification performance of two expert radiologists on test set 1 (145 images) and against clinical ground truth data on test set 2 (345 images). We used accuracy, sensitivity/specificity (recall), receiver-operator-characteristic-area-under-the-curve (ROC-AUC) and the Matthews correlation coefficient (MCC)[27] for assessment. Details can be found in the Methods section. Model and expert classification performance on the datasets can be found in Table 1.

The FL model trained with neither SecAgg nor DP performed best with no statistically significant difference to the centrally trained model. The addition of SecAgg to the model slightly, but non-significantly reduced performance. Both FL models and the centrally trained model significantly outperformed the human observers. The DP training procedure ($\epsilon = 6.0$, $\delta = 1.9 \times 10^{-4}$ at an α-value (divergence order) of 4.4) significantly reduced model performance, however the model still performed statistically on par with human observers and retained stable performance on the out-of-sample data of test sets 1 and 2. We note that the ϵ-value represents the total privacy budget spent at the end of training. The personalized models trained only on the data owners' individual data subsets performed approximately on par only on the validation data, but significantly worse on the out-of-sample data of test sets 1 and 2, indicating poor generalization. The statistical evaluation of these results alongside inter-rater/model agreement metrics can be found in Supplementary Section 2 and Supplementary Tables 1 and 2.

Training and inference performance benchmarking

To assess the performance ramifications of PriMIA's privacy-enhancing techniques, we benchmarked the training and inference performance in a variety of scenarios, shown in Fig. 3. Training timings were measured as average time per batch at a constant batch size to decouple them from dataset size. Compared to training locally, FL incurs a performance penalty due to network communications, which is further increased by the addition of SecAgg and DP, yielding a threefold increase in training time when both SecAgg and DP are used. Large neural network architectures require proportionally longer to train due to network transfer

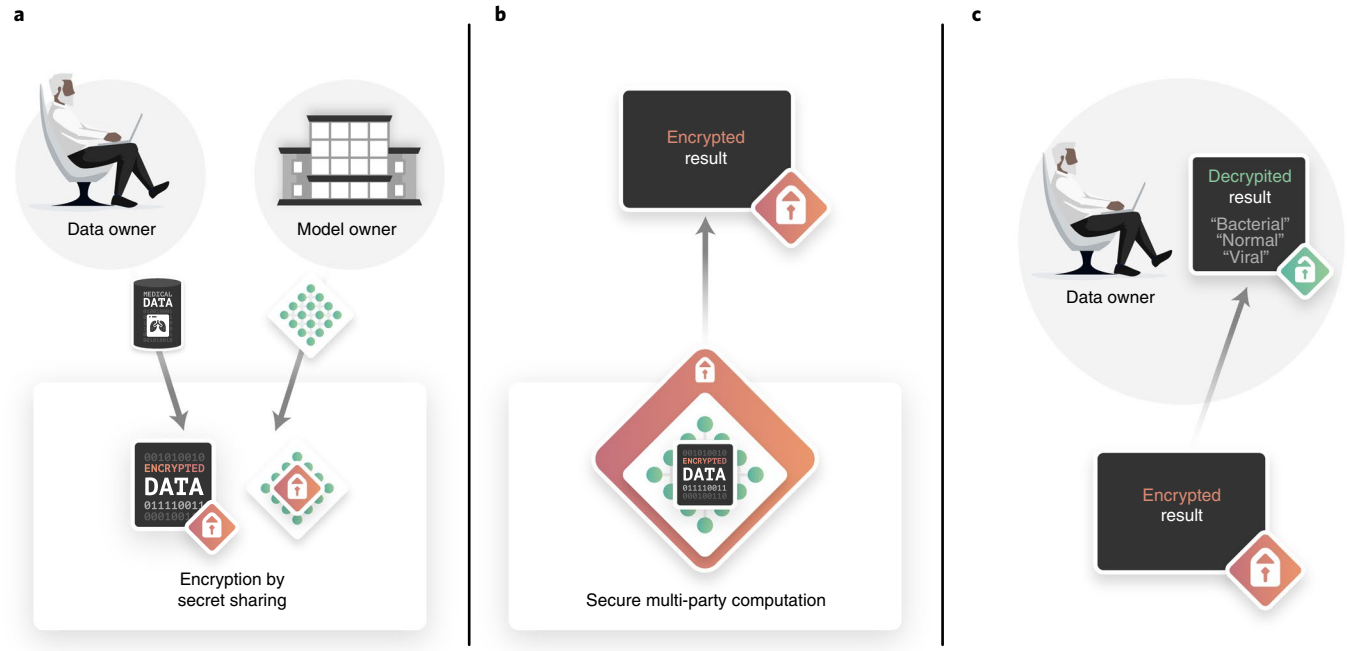

Fig. 2 | Overview of the encrypted inference process. The data owner (in this case, a physician located at a remote location) requests an inference result from the model over the Internet but wants the confidential patient data they hold to remain secret. Similarly, the model owner provides inference as a service but wants to keep their model confidential. The use of SMPC enables the following scenario. **a**, Initially the data owner and model owner respectively encrypt the data and model using secret sharing. This process relies on splitting the data/model into shares, which in themselves do not contain any usable information and can therefore be exchanged (shared) with the other party. **b**, Inference is then carried out by jointly computing a function (in this case the neural network inference procedure) using SMPC. **c**, The data owner receives an encrypted result, which only they can decrypt.

requirements, providing justification for the use of the ResNet18 architecture in our study compared with larger ResNets. The addition of more worker nodes led to a linear increase in times when utilizing SecAgg due to the communication overhead of the protocol. However, due to the small number of operations per round, the protocol scales well to multiple parties: linear regression analysis of the scaling yielded $t(w) = 0.57w + 2.61$ with t expressing time in seconds and w the number of workers ($R^2 = 0.98$, $p < 0.001$, $N = 100$ samples per number of workers tested). Training time was nearly constant without SecAgg. Training times per batch were constant for larger dataset sizes, signifying that training duration is dependent only on dataset size all other things being equal. Lastly, we benchmarked our encrypted inference implementation[26] based on the function secret sharing (FSS) protocol[28], which offers increased efficiency in the evaluation of comparison operations, max-pooling and batch normalization layers compared to the widely used SecureNN[29]. The utilization of FSS for encrypted inference significantly reduced inference times. In particular, in the high-latency setting, FSS yielded a proportionally better performance in comparison to SecureNN. Implementation details can be found in the Methods section and the statistical evaluation can be found in Supplementary Section 3.

Model inversion attack

Prior work[13,30] has demonstrated that model inversion attacks are able to reconstruct features or entire dataset records (in our case, chest radiographs), rendering them a threat to patient privacy in FL settings. To exemplify the susceptibility of models trained with and without the privacy-enhancing techniques offered by PriMIA, we utilized the improved deep leakage from gradients attack[31,32] with small modifications detailed in the Methods section. We chose this method because it was the first technique shown to be highly effective against the ResNet18 architecture used in our case study. Figure 4 shows exemplary results from the chest radiography case study. We utilized the pixelwise mean squared error (MSE), signal-to-noise ratio (SNR) and Fréchet inception distance (FID) metrics for quantifying attack success. Empirical evaluation yielded that the attack's success depends highly on the L2-norm of the gradient updates and the batch size used. To thus generate a best-case baseline of a highly successful attack, we attacked the centrally trained model with a batch size of one at the start of training, when the loss magnitude (and thus gradient norm) is highest. The attacks on the FL model with SecAgg used for our case study were not successful, most likely due to the high effective batch size of 600. Consistent with DP's privacy guarantees, the attacks were ineffective when DP training was used. Results showing that DP negates the attack even when the model is attacked locally or when SecAgg is not used are shown in Supplementary Section 5 and Supplementary Fig. 2.

To further underline the high risk of privacy-centred attacks in the healthcare imaging setting and thus the importance of privacy-enhancing techniques for collaborative model training, we performed additional experiments on the publicly available MedNIST dataset and were able to recover images disclosing sensitive patient attributes when DP was not utilized. No images could be recovered with DP in place (Fig. 5). Further details on the attack and the statistical evaluation can be found in the Methods and Supplementary Sections 4 and 6.

Discussion

We've presented PriMIA, an open-source framework for privacy-preserving FL and encrypted inference on medical images. We've demonstrated the decentralized collaborative training of an expert-level deep convolutional neural network in the challenging clinical task of paediatric chest radiography classification. Further, we've showcased end-to-end encrypted inference, which can be leveraged for secure diagnostic services without the disclosure of confidential data or exposure of the model. Our work serves

Table 1 | Classification performance comparison of models on the validation set and test sets 1 and 2

	Accuracy			Sensitivity/specificity			ROC-AUC			MCC		
	Val	Test 1	Test 2	Val	Test 1	Test 2	Val	Test 1	Test 2	Val	Test 1	Test 2
Federated DP−/SecAgg−	0.89	0.89	0.90	0.95	0.88	0.90	0.92	0.92	0.93	0.84	0.84	0.85
				0.86	0.88	0.88						
				0.86	0.94	0.93						
Federated DP−/SecAgg+	0.88	0.88	0.89	0.98	0.88	0.89	0.90	0.92	0.92	0.83	0.83	0.83
				0.86	0.88	0.88						
				0.78	0.91	0.91						
Federated DP+/SecAgg+	0.85	0.85	0.84	0.97	0.87	0.86	0.89	0.88	0.87	0.78	0.76	0.77
				0.76	0.81	0.83						
				0.82	0.85	0.86						
Centrally trained	0.92	0.90	0.91	0.96	0.90	0.93	0.93	0.93	0.94	0.87	0.85	0.87
				0.90	0.88	0.89						
				0.87	0.94	0.92						
Personalized 1	0.89	0.67	0.63	0.90	0.96	1.00	0.92	0.72	0.71	0.83	0.48	0.47
				0.88	0.19	0.25						
				0.88	0.71	0.65						
Personalized 2	0.87	0.69	0.58	0.88	0.85	0.91	0.90	0.74	0.67	0.80	0.51	0.37
				0.85	0.65	0.29						
				0.87	0.41	0.50						
Personalized 3	0.87	0.68	0.66	0.86	0.68	1.00	0.90	0.75	0.79	0.80	0.50	0.48
				0.90	0.79	0.72						
				0.84	0.53	0.00						
Expert 1	-	0.79	-	-	0.96	-	-	-	-	-	0.70	-
					0.47							
					0.88							
Expert 2	-	0.79	-	-	0.96	-	-	-	-	-	0.68	-
					0.84							
					0.41							

Federated, model trained with federated learning; DP+/−, model trained with (+) or without (−) DP gradient descent; SecAgg+/−, model trained with (+) or without (−) SecAgg; Centrally trained, model trained on the entire dataset on a single machine. Personalized 1–3, models trained only on the data owner's local data set. Expert 1/2, human experts. Sensitivity/specificity metrics refer to normal/bacterial/viral, respectively.

as the first step towards the implementation of next-generation privacy-preserving methods in medical imaging workflows. It applies to both multi-institutional research and to enterprise model development settings, allowing the preservation of data governance and sovereignty over confidential patient health data. Our framework can be used in inference-as-a-service scenarios in which diagnosrsquo support can be provided remotely with theoretical and empirical guarantees of privacy, confidentiality and asset protection. PriMIA represents a targeted evolution of our previous work[17] towards healthcare-sector-focused deployment. Although we focused on a classification task for the presented case study, PriMIA is highly adaptable to a variety of medical imaging analysis workflows employing different network architectures, datasets and more. We present an additional case study focused on semantic segmentation in computed tomography scans of the abdomen in Supplementary Section 7 and Supplementary Fig. 3, to demonstrate this flexibility.

Model classification performance. Recent work has evaluated the ramifications of data quality (overly homogeneous/independent and identically distributed data versus overly heterogeneous data) and distributed system topology on federated model performance, for example generalization to out-of-sample data. In our case study, models trained with FL performed on par with the centrally trained model similar to ref. [5] and outperformed human observers. Models trained only on subsets of the data (personalized models) showed drastically diminished performance on out-of-sample data. Since personalized model training is the standard in most mono-centric medical imaging studies, this finding serves as a reminder that the inclusion of larger quantities of more diverse data from multiple sources enabled through FL can allow the training of models with better generalization performance, as is demanded by current best practices[33]. DP model training is able to offer objective privacy guarantees and resilience against model inversion attacks[30,32]. The utilization of DP diminished model performance, which was, however, still on par with human observers. At the same time, the DP guarantees achieved ($\epsilon = 6$) by the selected model are only moderate. This phenomenon (privacy–utility trade-off) is a well-known observation in the still nascent area of deep learning with DP. For instance, previous work[23] reached an ϵ-value of approximately 8 on the CIFAR-10 dataset and another study reported[34] ϵ-values between 6.9 and 8.48. Both studies also report a diminished performance by the final model. We regard methods to improve the training of DP models as a promising direction for future research.

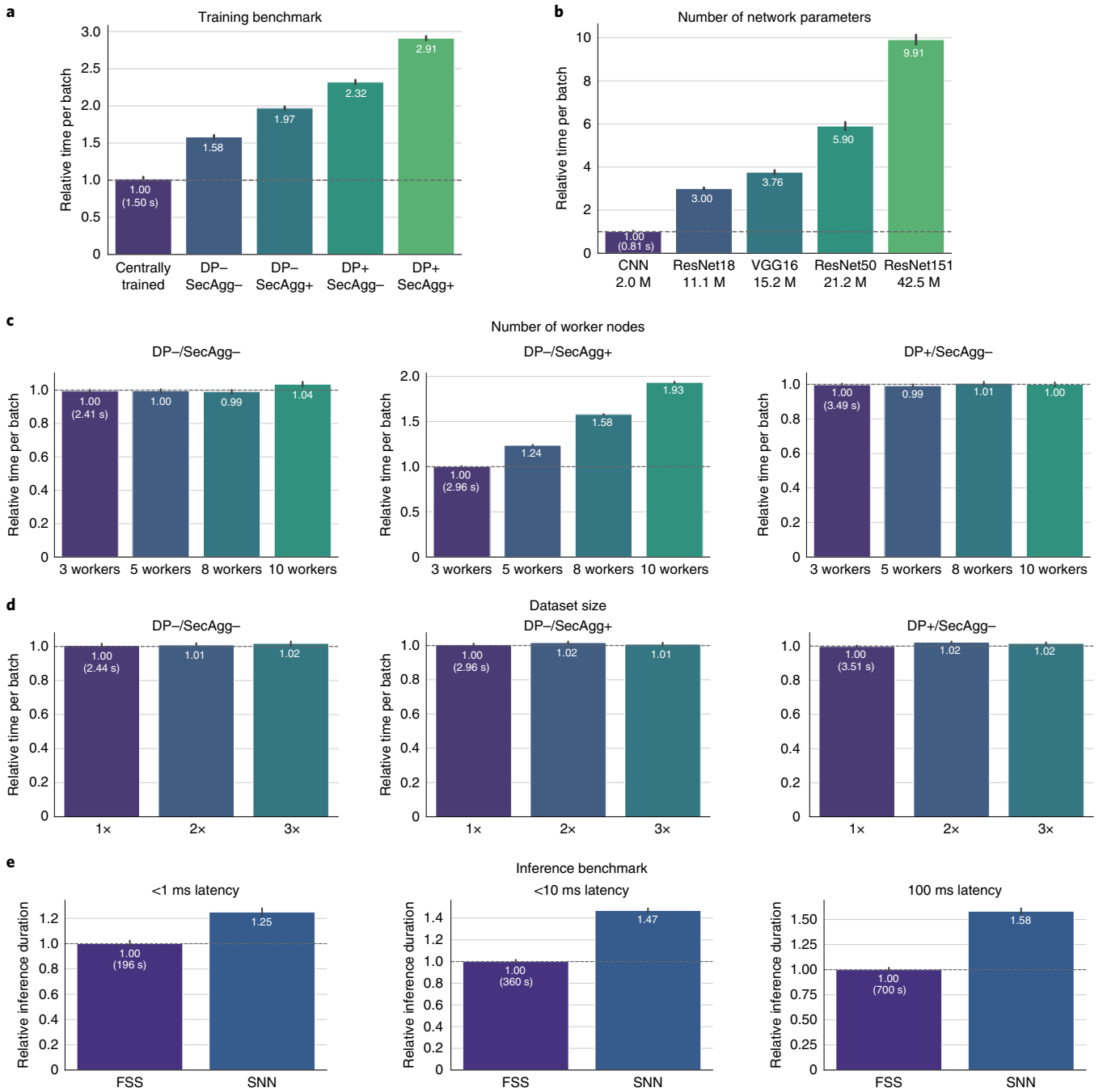

Fig. 3 | Results of training and inference benchmarks. a–**d**, Timing benchmarks in the training phase. All times shown in white are relative to the baseline for a batch size of 8 at a constant synchronization rate of 1 averaged over 100 runs. For DP, a microbatch size of 1 was used. The baseline is provided in parentheses. Bars denote standard deviation. Centrally trained: local training. DP+/– and SecAgg+/–: with/without DP gradient descent/SecAgg. **a**, Training latency for local training in various scenarios. **b**, The influence of neural network model parameters. Models shown: CNN architecture included with PriMIA (2.0 million parameters), ResNet18 (11.1 million parameters), VGG16 (15.2 million parameters), ResNet50 (21.2 million parameters) and ResNet151 (42.5 million parameters). **c**, The influence of the number of workers (data owners) in the federation. **d**, The influence of the dataset size per worker between one (1×) and three (3×) times the amount of data. As times shown are per batch, timings are independent of dataset size. **e**, Timing benchmark in the inference phase. FSS, function secret sharing-based inference (ours). SNN, SecureNN protocol[29]. 100 repetitions each. Latency, average 10-round-trip ping latency.

Functional improvements to FL. To increase framework usability and flexibility as well as FL model performance, our framework includes the following functional improvements. (1) Besides incorporating adaptive client optimization in the form of the Adam optimizer recently shown to yield improved convergence results[35], we include a wide range of advanced image augmentation techniques including MixUp, which has been shown to encompass privacy-enhancing attributes[36]. (2) We implement techniques to address imbalances in data volume between nodes (local early stopping), as well as between dataset classes (class-weighted gradient descent and federated averaging[37]). (3) We include facilities to carry out centrally coordinated hyperparameter optimization

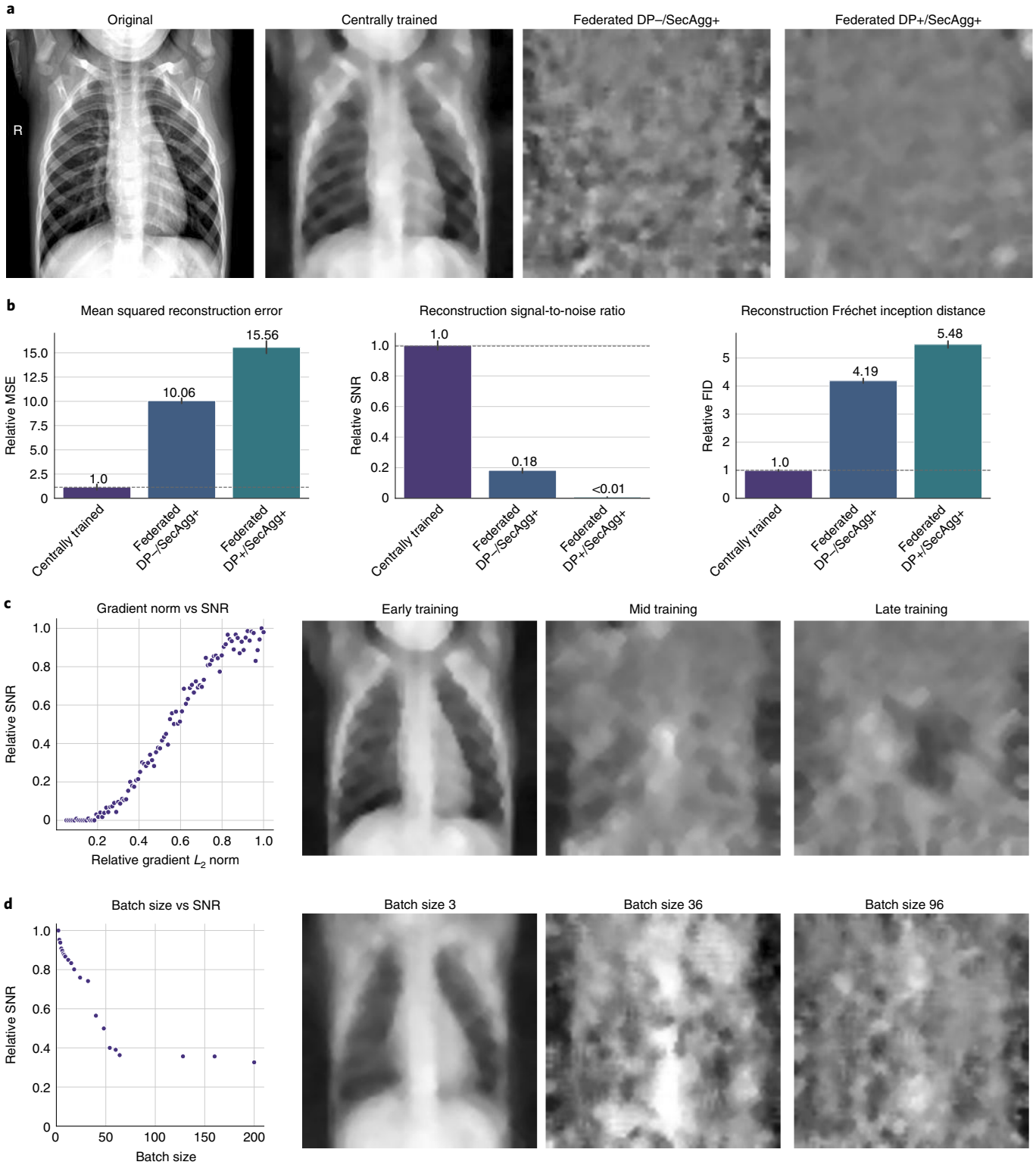

Fig. 4 | Overview of the gradient-based privacy attacks against PriMIA using the paediatric pneumonia dataset. a, Left to right: the target image (original); best-case reconstruction derived from attacking the centrally trained model early during training with a batch size of 1; typical case of an attack against the FL model trained with SecAgg (effective batch size 600, epoch 5 of 20); worst-case attack performed against a model trained with DP. **b**, Normalized metrics of attack success. Lower values for pixel-wise MSE and FID (mirroring human perception of similarity) and higher values for signal-to-noise ratio indicate increased success, respectively. **c**, Attack success, measured as relative signal-to-noise ratio dependent on the model's global L_2-norm. As training progresses, loss decreases and thus the gradient norm diminishes, reducing attack success. **d**, The influence of effective batch size on attack success measured as relative signal-to-noise ratio. High batch sizes substantially impede attack success. Chest radiographs from Mendeley Data[67].

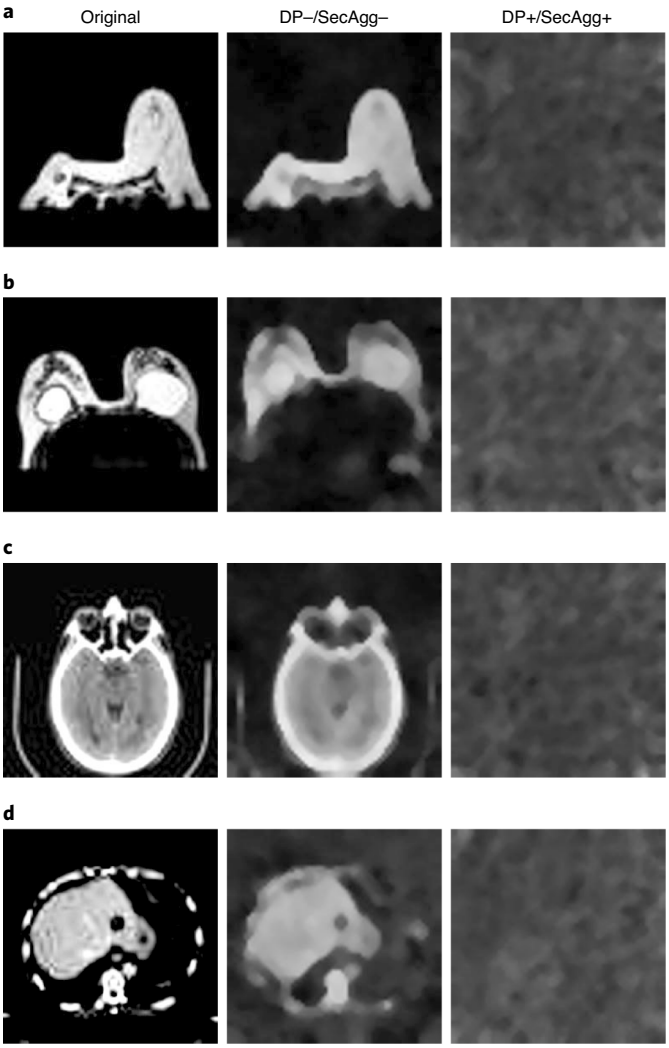

Fig. 5 | Overview of the gradient-based privacy attacks against PriMIA using the MedNIST dataset in a variety of scenarios. The original image is shown (original) alongside the reconstruction results from a model trained without secure aggregation or DP (DP–/SecAgg–) as well as a model trained with DP and SecAgg (DP+/SecAgg+). In every case, the attack reveals confidential information about the patient when the model is trained without privacy-enhancing techniques. **a**, Breast MRI revealing absence of the right breast, likely due to operative removal due to breast cancer. **b**, Breast MRI revealing breast implants. Both **a** and **b** also allow assumptions about the patient's sex. **c**, Cranial computed tomography image at the level of the nose. Facial contours reconstructed from such images can lead to personal identification[39]. **d**, Abdominal CT at the level of the liver, allowing visualization of a hypodense lesion in the left liver lobe in the reconstructed image. In every case, using DP thwarts the attack, disallowing any usable image features from being visualized. CT images licensed under the Creative Commons CC BY-SA 4.0.

over the entire confederation using the Tree-Structured Parzen Estimator algorithm[38]. Experimental data showcasing the utilization of our hyperparameter selection framework to search for the optimal FL model can be found in Supplementary Section 1 and Supplementary Fig. 1. All above-mentioned training optimizations are implemented locally on the nodes and do not negatively impact privacy guarantees. Hyperparameter tuning, however, must be considered when DP is utilized, as it relies on multiple training repetitions.

Discussion on privacy-enhancing techniques. The inclusion of methods offering provable privacy and security guarantees in the FL process is a crucial step towards the widespread implementation of privacy-preserving AI technologies[8]. The successful reconstruction of images from unprotected models in our attack experiments underline the risks of such attacks to patient privacy, which has also been discussed in previous work[6,39]. DP training provides objective privacy guarantees in case of attacks against the model both by confederation members and during inference and is not limited to the gradient-based inversion attack we use in our example. SecAgg utilizing SMPC only discloses the aggregate model update to the parties, even in case up to $n-1$ out of n parties collude to reveal data. The DP secure aggregation of dataset statistics (means and standard deviations) we propose can protect FL participants from data leakage, especially when non-imaging data is included in model building (for example clinical records, in which the means of features such as age represent sensitive information). Finally, encrypted inference reveals no information about the data or the model to either party.

Compared with fully homomorphic encryption protocols[40] relying on key-based cryptography, whose implementation for neural network training and inference is impeded by the computational complexity of the encryption process and the performance decrease due to function approximation for for example activation functions, communication overhead has traditionally been the limiting factor for SMPC. In our recent work, we introduced AriaNN[26], an SMPC protocol leveraging function secret sharing (FSS)[28] and building upon SPDZ[25]. It represents an alternative to protocols like SecureNN[29] or Falcon[41], and computes private comparisons with a single round of communication. This renders FSS substantially more communication-efficient than other SMPC protocols, especially when parties are geographically distant and communicate with high latency, for example when performing inference over the public web as showcased in our study. Through the present use-case, we confirm the results obtained in our previous work on other datasets: secure inference gains proportionally greater benefits from the FSS protocol in the high-latency setting. Thus, we propose its utilization over SecureNN in cases a reduction in latency is desired in an honest-but-curious setting.

Comparison to prior work. Several current works aim to introduce PPML techniques to biomedical imaging: Silva et al.[42] present a front-end FL framework for biomedicine, but do not consider DP, SecAgg or encrypted inference. Xu and colleagues (https://bit.ly/3pl5dD1) provide a framework for FL using homomorphic encryption for SecAgg, but do not utilize DP or provide encrypted inference capabilities. Sheller et al.[43] showcase an FL use-case based on segmentation. They do not assess either DP, SecAgg or the option for encrypted inference. Li et al.[44] also demonstrate an FL segmentation task. Their DP implementation relies on an alternative technique (sparse vector) and the framework does not provide secure aggregation or encrypted inference. The work by Lu and colleagues[45] demonstrates FL with DP, however their use-case is focused around pathology slides and does not employ SecAgg or provide encrypted inference capabilities. Li et al.[46] utilize DP, however assume a fixed sensitivity and do not conduct privacy analysis. Their framework does not offer SecAgg or encrypted inference.

Limitations. We consider the following limitations of our work. The computational requirements for deploying our system are substantial, and the latency resulting from encrypted inference is still very high compared to unencrypted inference, despite the proposed protocol improvements. The underlying remote execution environment currently offers experimental graphics processing unit (GPU) support, with full support planned for an upcoming version. The success of FL models is largely dependent on high data quality on

the nodes. The auditing and curation of the data and its quality, methods to quantify the contribution of individual datasets to the model or to detect local overfitting are still under investigation[47]. Our library is designed to be used in an honest-but-curious regime, which we believe to represent the standard in healthcare consortia. Thus, although we provide comprehensive privacy protection measures, we included no specific countermeasures against malicious contributions of low-quality or adversarial data to the FL process or to verify/guarantee to the data owner that the model used in the inference setting is the one promised. Furthermore, we point out that discussions of the theoretical threat model are a level of abstraction that cannot fully represent the complexity of real-life situations. For instance, threat modelling is typically undertaken on the level of FL participants representing entire hospitals, however this cannot take every individual person working for these hospitals and their specific motivations into account. Similarly, questions about participant reimbursement or model ownership in FL were outside the scope of our current investigation. Further studies in this developing field are required to fully illuminate such details. Lastly, as mentioned above, the utilization of DP causes a direct trade-off between model privacy and utility. Future work will need to address this trade-off through improved privacy analysis and training techniques, as the privacy guarantees of current studies, including the ϵ-value of around 6.0 seen in our study, are not yet sufficiently rigorous to be considered generally applicable.

Conclusion

We present a free, open-source software framework for privacy-preserving FL and end-to-end encrypted inference on medical imaging data, which we showcase in a clinically relevant real-life case study. 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 utilization of PPML techniques in healthcare and beyond.

Methods

Dataset collection. For model training, we used the previously proposed paediatric pneumonia dataset[20]. The dataset was reviewed by a specialist radiologist for image quality and representativeness and included 5,163 training images in the above-mentioned three categories, as well as a validation set of 624 images. For FL model development, the training set was randomly subsampled into three equally sized non-overlapping partitions. Class balance between nodes was not enforced.

For model testing on unseen data, we retrospectively collected 497 chest radiographs of the same classes of an age-matched cohort from two university hospitals (test set 1: 145 images (43 bacterial, 68 normal, 34 viral), test set 2: 352 images (120 bacterial, 126 normal, 106 viral)). Ethics committee and data protection votes for data collection and exchange were granted by all institutions waiving the requirement for informed consent in this retrospective study (protocol number 111/20 S-KH). All procedures were carried out in accordance with clinical best practices, applicable laws and regulations as well as the Declaration of Helsinki. Ground-truth labels for the dataset were generated from clinical records based on validated laboratory results and clinical parameters (c-reactive protein (CRP), body temperature, antibiotic response for bacterial, sputum or sweat polymerase chain reaction (PCR) and/or absence of bacterial infection signs for viral) as well as clinical assessment of specialist paediatricians/neonatologists not involved in image evaluation.

Model training. *Privacy-preserving processing of dataset statistics.* For the training of neural networks, data is typically pre-processed by mean subtraction and division by the standard deviation. In federated learning, dataset statistics from the local nodes or aggregated statistics from all nodes can be used. Additionally, the provision of the final model in an inference setting requires these statistics for rescaling incoming images. However, dataset statistics can contain private information that should not be shared, especially in case non-imaging data is included (for example, age in the case of clinical record data). Hence, we propose and implement differentially private secure aggregation of dataset statistics. Here, sensitivity-calibrated Laplacian noise is added to the statistics to satisfy a user-defined ϵ DP value before SMPC is used to average them, and they are then stored on the central server for later use. Before inference starts, the data is rescaled with the (differentially private) securely aggregated mean and standard deviation of the training set. For training, the nodes use their local dataset statistics. Thus, data leakage is prevented, especially in the case individual nodes contain few, or just one, dataset(s).

Model architecture, hyperparameters and augmentation. We used the ResNet18 architecture[19], pretrained on ImageNet[48], with the final average pooling layer replaced by a single linear layer with 512 units and randomly initialized with the Kaiming Uniform initializer[49]. Images were cropped to squares such that the entire chest section of the radiograph is preserved and resized to 224×224 pixels.

The following standard augmentation techniques were employed: random horizontal flips, random affine transformations, Gaussian noise injection. In extension, we used the Albumentations library[50] to apply the following transformations: random changes in the gamma value and brightness, blurring, optical distortions, grid shuffles/dropouts/distortions, elastic transforms, changes in hue-saturation-value (HSV) colour space, inverting images, cutouts of the image, artificial shadows, fog, solarizations and sun flares. We also provide the option for histogram equalization or contrast-limited adaptive histogram equalization (CLAHE), both as an augmentation and a standardization technique. The individual augmentations were introduced with a probability p_1 and augmentation was activated overall with a probability p_2. Furthermore, we applied a modified variant of MixUp augmentation[51] by which the mixing parameter (λ) is randomly sampled from a uniform distribution similar to that in ref. [36].

Training was performed for 40 epochs using the Adam optimizer[52] with a log-linearly decreasing learning rate initially set at 10^{-4}. PriMIA caches models automatically after each round, and selects the model with the highest validation set Matthews correlation coefficient (MCC). The centralized model was trained by pooling all data on a single machine and training the model on the accumulated dataset. Personalized models were trained on the respective nodes using only the local dataset. PriMIA implements the ability to carry out centrally coordinated automated hyperparameter tuning on the entire federation or locally, which was used to determine the best model in every case according to highest validation set MCC. An example is provided in Supplementary Section 1 and Supplementary Fig. 1. Model hyperparameters are centrally set for all nodes, but image augmentation, local early stopping and weighted gradient descent are performed locally and independently on the nodes. Federated training and inference experiments were conducted over the public Internet on cloud instances with 32 CPU cores at 3.1 GHz and 64 GB of random access memory (RAM). Centralized model training was performed on a server with 36 CPU cores at 2.4 GHz and 512 GB of RAM.

Differentially private model training. DP model training entails several additional considerations. We describe these alongside PriMIAs DP implementation and the process of training the final DP model at length in Supplementary Section 8. In brief, PriMIA implements DP gradient descent[23] based on clipping the gradient L_2-norm of each individual sample, then adding calibrated Gaussian noise. This process occurs on each node independently with independent noise sources (local DP). We considered the paediatric pneumonia dataset private, therefore did not perform hyperparameter optimization based on multiple training runs. Furthermore, due to the relatively small size of the dataset, we determined it would not be possible to train the model with sufficient utility while maintaining acceptable privacy guarantees. Hence, we used the pre-training technique described previously[23] and employed a publicly available dataset trained on a related task to determine the optimal parameters for the DP mechanism and pre-train the model. Details can be found in Supplementary Section 8.2.2 and Supplementary Fig. 4.

Training topology, gradient descent and secure aggregation. We selected the hub-and-spoke system topology due to its reported improved final model performance over techniques such as incremental or cyclical training[5,43] and its higher flexibility with respect to node availability and asynchronous training[53]. In PriMIA training is carried out asynchronously in rounds. Initially, the model is sent from the central server to all computation nodes. During each round, nodes locally perform a variant of gradient descent in which gradient updates are weighted inversely by the frequency of the individual dataset classes present on the node (class weighted gradient descent). After a number of batches (denoted by σ) have been processed on every node, the updated models are securely averaged (SecAgg[54]) using the FSS SMPC protocol (see below), before being distributed back to the nodes. For model averaging, we utilize class-weighted federated averaging[37] whereby the central model updates are weighted by the class frequency on the nodes before a new training round begins.

Model synchronization and the σ parameter. Previous work has investigated the federated synchronization rate parameter (σ) as central in controlling network input/output and training duration[55]. We found the choice of this parameter to also affect model performance and training time, and it has recently been described as an important open research target in FL with respect to the optimal trade-off between model accuracy and training time[47]. We provide further details on these findings in Supplementary Section 10 and Supplementary Fig. 6.

Measures against FL training deterioration. Literature findings and our own evidence indicate that, in case one of the federation's nodes contains less data than others, continuing training beyond convergence until other nodes have completed training can lead to overfitting or training collapse. Alternatively, not including the updates from this node can lead to catastrophic forgetting[56] of the node's data and reduced generalization performance. We empirically determined

that local early stopping, that is, terminating training on the local node once the node's local dataset is exhausted, then using the state of the node's local model for all future update steps until a full round of training is completed, led to improved training performance.

Secure multi-party computation protocols. *Function secret sharing.* FSS belongs to the family of SMPC protocols, in which several parties share a secret (for example, data or a model) to ensure privacy. A party alone holds a random share of the private value and cannot reconstruct the value on their own. A quorum of parties (sometimes all parties) need to collaborate to reconstruct the private data. The terms encrypted and obfuscated are used interchangeably in this scenario to denote secret-shared data.

Unlike classical data secret sharing schemes like SecureNN[29], where a shared input $[\![x]\!]$ is applied on a public function f, FSS applies a public input x on a private shared function $[\![f]\!]$. Shares or *keys* $([\![f]\!]_0, [\![f]\!]_1)$ of a function f satisfy $f(x) = [\![f]\!]_0(x) + [\![f]\!]_1(x)$. Both approaches output a secret shared result. In our case, assume two parties respectively own shares $[\![y]\!]_0$ and $[\![y]\!]_1$ of a private input y, and they want to compute $[\![y \geq 0]\!]$. They receive some cryptographic primitives (see below), namely each get a share of a random value (or *mask*) $[\![\alpha]\!]$ and a share of the shared function $[\![f_\alpha]\!]$ of $f_\alpha: x \rightarrow (x \geq \alpha)$. They first mask their shares of $[\![y]\!]$ using $[\![\alpha]\!]$, by computing $[\![y]\!]_0 + [\![\alpha]\!]_0$ and $[\![y]\!]_1 + [\![\alpha]\!]_1$ and then revealing these values to reconstruct $x = y + \alpha$. Next, they apply this public x on their function shares $[\![f_\alpha]\!]_{j=0,1}$, to obtain a shared output $([\![f_\alpha]\!]_0(x), [\![f_\alpha]\!]_1(x)) = [\![f_\alpha(y+\alpha)]\!] = [\![(y+\alpha) \geq \alpha]\!] = [\![y \geq 0]\!]$. Previous studies on FSS[57,58] have shown the existence of such function shares for comparison which perfectly hide y and the result. For more details about the concrete implementation of FSS we refer to our previous work[26]. SMPC and the FSS protocol provide theoretical security guarantees in the honest-but-curious regime. FSS offers high communication efficiency and can be thus employed to reduce transaction latency. FSS is based in part on the SPDZ protocol[25]. To increase efficiency for specific mathematical operations (for example multiplication) by reducing the rounds of communication required to perform the operation, protocols such as SPDZ partition encrypted operations into an offline phase, during which no communications between parties take place, and an online phase, where parties communicate. During the offline phase, a trusted third party, referred to in PriMIA as a cryptographic provider (and in ref. [25] as a trusted dealer), provides cryptographic primitives. In practice, it is not a requirement for parties to use the PriMIA cryptographic provider, as the framework can be modified to use a trusted third party of their own choosing. These primitives can be computed in advance as they require no knowledge of the exact functions evaluated during the online phase, and the cryptographic provider does not participate in the online phase in which these computations take place. A schematic representation of the two phases and further terminology are provided in Supplementary Section 9 and Supplementary Fig. 5.

Secure aggregation. The SecAgg operation, consisting of a private addition and a public multiplication is performed using the additive secret sharing scheme of the underlying SPDZ[25] protocol. The protocol is designed such that random shares are distributed between participants, which individually contain no usable information and only the sum of their contributions (that is, the aggregated model updates) are revealed. Collusion between up to $n-1$ out of n participants (in the case study, two out of three) is insufficient to disclose the other participant's private information. SecAgg is performed without a need for cryptographic primitives or the cryptographic provider.

Secure inference. Secure inference represents a transaction between two parties, by which the data owner wishes to receive the model's prediction without disclosing their data, and the model owner wishes to keep their model hidden. We adapt our previous work on AriaNN[26], based on FSS, for encrypted inference to leverage its high communication efficiency, which allows the evaluation of private comparisons with minimal communication overhead. Such comparison operations are important for example for the evaluation of maximum pooling layers or rectified linear units. The cryptographic primitives provider is again not required for the actual inference process (online phase), which occurs exclusively between the two parties. In our framework, the data owner initiates a request to the system, the data and model are obfuscated by secret sharing and inference takes place using SMPC. Secure inference scenario is thus—in the sense described above—an end-to-end encrypted transaction, whereby both the data and the model is obfuscated. This guarantees both parties single-use accountability, that is, the guarantee that the data and model can be used for no other purpose than the one explicitly designated by the involved parties.

We note that while the data enjoys information-theoretic secrecy guarantees, the party requesting inference has access to the model's predictions and can perform black-box membership inference[59] or model inversion attacks[60]. PriMIA's DP training procedure provides effective protection against such attacks[30,32,59] to the individuals whose data was used to train the model used for inference.

Classification performance assessment. Classification performance was evaluated as follows. For expert readers, accuracy, sensitivity/specificity (recall) and MCC[27] were calculated on test set 1. The model's performance was evaluated in terms of accuracy, sensitivity/specificity (recall), ROC-AUC MCC on the validation set and on both test sets. MCC was employed due to its invariance to class imbalance and its indication of prediction concordance alongside quality of classification, leading to recent recommendations for its use over the usually employed accuracy or F1-Score metrics[61]. McNemar's test was used to test for statistical significance in classification performance. Cohen's κ (kappa) was used to test inter-rater/-model agreement. Statistical significance is defined as $p < 0.05$.

Inference and training latency assessment. We compared the average ± standard deviation duration in seconds of 1 epoch of training over 100 epochs as well as the average ± standard deviation duration of one inference transaction over 100 transactions in three settings: utilizing inter-process communication locally (using the PySyft VirtualWorker abstraction (no latency), utilizing the websocket/HTTP protocol on the local network (LAN) (low latency) and utilizing the public Internet (WAN) (high latency) with a 10-round-trip ping latency of 100 ms. Student's t-test was used to assess statistical significance.

Model inversion utilizing gradient updates. To exemplify the susceptibility of models trained without privacy-enhancing techniques against adversarial agents that attempt to expose sensitive data, we employ the Improved Deep Leakage from Gradients, iDLG, method with modifications as proposed previously[32], itself a variant of previously shownn techniques[31,62]. iDLG was found highly successful against the ResNet18 architecture used in our case study. We additionally modified the attack following newer evidence from[63] by utilizing the AdamW optimizer and initializing images with uniform sampling to further improve its success. The overview of the attack is as follows:

1. Adversary generates a randomized pair of a dummy model update and a corresponding label
2. Adversary captures the gradient update submitted by an honest client
3. Using a suitable cost function, the adversary attempts to minimize the difference between the honest update and the dummy update
4. The algorithm is repeated until either the loss starts diverging or the final iteration is reached

In the original implementation of the protocol, the difference between gradients is calculated using

$$||\Delta W' - \Delta W||^2 = ||\frac{\delta l(F(x', W), y')}{\delta W} - \Delta W||^2$$

where x' and y' are the data point and its label respectively, while W and W' are the victim's and attacker's gradient respectively. Following Geiping et al.'s implementation, we used the cosine similarity metric and utilized images of size 224×224, as authors show that this is the upper bound for acceptable reconstruction quality[32]. The empirical evaluation of various batch sizes showed that larger batch sizes drastically reduce the success of the reconstruction. We indicate an averaged model update from n parties each trained with a batch size of k to have been trained with an effective batch size of $n \times k$. Our observation matches ref. [32] which shows batch sizes above eight to substantially deteriorate the attack. We furthermore found the L_2-norm of the gradient update to strongly influence attack success. Thus, attacks at the beginning of training, when the loss (and thus the gradient with respect to it) is largest, were most successful. A low MSE value did not always signify a successful attack, since a specific model update can be generated by more than one image, resulting in noise that is able to mimic the update, but not the corresponding data. To improve attack evaluation, we also supply signal-to-noise ratio and perceptual metrics which more robustly assess the reconstruction quality and human perception of image similarity as performed in[32,64–66]. As an active attack, iDLG can be executed by an adversarial client or central server. We note that in the case of an adversarial central server, the usage of SMPC prevents the disclosure of individual model updates, therefore only allowing the adversary to utilize averaged model updates instead. For the attacks on the FL system we assumed that one out of three data owners is an adversary. For the 'baseline' attack on the centralized model, we used a batch size of 1. Attacks were performed against 100 randomly selected images from the training set. For the gradient norm experiments, 100 gradient samples were taken at equispaced intervals during model training. Batch size experiments were carried out under identical circumstances only varying batch size. Model and dummy image initialization was deterministically set for all experiments. Each attack was performed in triplicate with at most 24,000 iterations per run and the instance with the highest cosine similarity was selected. One way analysis of variance (ANOVA) followed by the Student's t-test were used to assess statistical significance between the MSE, SNR and FID scores. Details of the attack against the MedNIST dataset can be found in Supplementary Section 6.

Data availability

The paediatric pneumonia dataset is publicly available from Mendeley Data at https://doi.org/10.17632/rscbjbr9sj.3. The MedNIST dataset was assembled by B. J. Erickson (Department of Radiology, Mayo Clinic) and is available at

https://github.com/Project-MONAI/MONAI/. The MSD Liver Segmentation Dataset is available at http://medicaldecathlon.com. test sets 1 and 2 contain confidential patient information and cannot be shared publicly. Source data are provided with this paper.

Code availability

The current version of the PriMIA source code is publicly available at https://github.com/gkaissis/PriMIA and permanently archived at https://doi.org/10.5281/zenodo.4545599[18]. PriMIA includes source code from PySyft (https://github.com/OpenMined/PySyft), PyGrid (https://github.com/OpenMined/PyGrid) and Opacus (https://github.com/pytorch/opacus) re-used under open-source licence terms.

Received: 4 October 2020; Accepted: 26 March 2021;
Published online: 24 May 2021

References

1. McKinney, S. M. et al. International evaluation of an AI system for breast cancer screening. *Nature* **577**, 89–94 (2020).
2. Ardila, D. et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nat. Med.* **25**, 954–961 (2019).
3. Patent Index 2019: Spotlight on digital technologies. *European Patent Office* https://www.epo.org/about-us/annual-reports-statistics/statistics/2019.html (accessed 10 March 2021).
4. Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* **25**, 44–56 (2019).
5. Sheller, M. J. et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Sci. Rep.* **10**, 12598 (2020).
6. Schwarz, C. G. et al. Identification of anonymous MRI research participants with face-recognition software. *N. Engl. J. Med.* **381**, 1684–1686 (2019).
7. Narayanan, A. & Shmatikov, V. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)* 111–125 (IEEE, 2008).
8. Price, W. N. & Cohen, I. G. Privacy in the age of medical big data. *Nat. Med.* **25**, 37–43 (2019).
9. Banerjee, S., Hemphill, T. & Longstreet, P. Wearable devices and healthcare: data sharing and privacy. *Inf. Soc.* **34**, 49–57 (2018).
10. Raisaro, J. L. et al. SCOR: a secure international informatics infrastructure to investigate COVID-19. *J. Am. Med. Inform. Assoc.* **27**, 1721–1726 (2020).
11. Vaid, A. et al. Federated learning of electronic health records to improve mortality prediction in hospitalized patients with COVID-19: machine learning approach. *JMIR Med. Inform.* **9**, e24207 (2021).
12. Roth, H. R. et al. Federated learning for breast density classification: a real-world implementation. In *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning* 181–191 (Springer, 2020).
13. Fredrikson, M., Jha, S. & Ristenpart, T. 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).
14. Wang, Z. et al. Beyond inferring class representatives: user-level privacy leakage from federated learning. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications* 2512–2520 (IEEE, 2019).
15. La, H. J., Kim, M. K. & Kim, S. D. A personal healthcare system with inference-as-a-service. In *2015 IEEE International Conference on Services Computing* 249–255 (IEEE, 2015).
16. Kaissis, G. A., Makowski, M. R., Rückert, D. & Braren, R. F. Secure, privacy-preserving and federated machine learning in medical imaging. *Nat. Mach. Intell.* **2**, 305–311 (2020).
17. Ryffel, T. et al. A generic framework for privacy preserving deep learning. Preprint at https://arxiv.org/abs/1811.04017 (2018).
18. Kaissis, G. & Ziller, A. PriMIA version 2021.02 https://doi.org/10.5281/zenodo.4545599 (2021).
19. 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).
20. Kermany, D. S. et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* **172**, 1122–1131 (2018).
21. Gupta, G. R. Tackling pneumonia and diarrhoea: the deadliest diseases for the world's poorest children. *Lancet* **379**, 2123–2124 (2012).
22. Evans, D., Kolesnikov, V. & Rosulek, M. A pragmatic introduction to secure multi-party computation. *Found. Trends Privacy Secur.* **2**, 70–246 (2018).
23. Abadi, M. et al. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security* (ACM, 2016).
24. Mironov, I., Talwar, K. & Zhang, L. Rényi differential privacy of the sampled gaussian mechanism. Preprint at https://arxiv.org/abs/1908.10530 (2019).
25. Damgård I., Pastro V., Smart N. & Zakarias S. Multiparty computation from somewhat homomorphic encryption. In *Advances in Cryptology – CRYPTO 2012* (eds. Safavi-Naini, R. & Canetti R.) (Springer, 2012).
26. Ryffel, T., Pointcheval, D. & Bach, F. ARIANN: low-interaction privacy-preserving deep learning via function secret sharing. Preprint at https://arxiv.org/abs/2006.04593 (2020).
27. Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta Protein Struct.* **405**, 442–451 (1975).
28. Boyle, E., Gilboa, N. & Ishai, Y. Function secret sharing. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques* 337–367 (Springer, 2015).
29. Wagh, S., Gupta, D., & Chandran, N. Securenn: 3-party secure computation for neural network training. In *Proc. Privacy Enhancing Technologies* 26–49 (Sciendo, 2019).
30. Carlini, N., Liu, C., Erlingsson, Ú., Kos, J. & Song, D. The secret sharer: evaluating and testing unintended memorization in neural networks. In *28th {USENIX} Security Symposium ({USENIX} Security 19)* 267–284 (2019).
31. Zhao, B., Mopuri, K. R. & Bilen, H. iDLG: improved deep leakage from gradients. Preprint at https://arxiv.org/abs/2001.02610 (2020).
32. Geiping, J., Bauermeister, H., Dröge, H. & Moeller, M. Inverting gradients. How easy is it to break privacy in federated learning? In *Advances in Neural Information Processing Systems* 16937–16947 (NeurIPS, 2020).
33. Bluemke, D. A. et al. Assessing radiology research on artificial intelligence: a brief guide for authors, reviewers, and readers—from the radiology editorial board. *Radiology* **294**, 487–489 (2020).
34. Wu, B. et al. P3SGD: patient privacy preserving SGD for regularizing deep CNNs in pathological image classification P3SGD. In *Proc. Conference on Computer Vision and Pattern Recognition* 2099–2108 (CVPR, 2019).
35. Reddi, S. et al. Adaptive federated optimization. Preprint at https://arxiv.org/abs/2003.00295 (2020).
36. Fu, Y., Wang, H., Xu, K., Mi, H. & Wang, Y. Mixup based privacy preserving mixed collaboration learning. In *2019 IEEE International Conference on Service-Oriented System Engineering (SOSE)* 275–2755 (IEEE, 2019).
37. McMahan, B., Moore, E., Ramage, D., Hampson, S. & y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics* 1273–1282 (PMLR, 2017).
38. Bergstra, J., Bardenet, R., Bengio, Y. & Kégl, B. Algorithms for hyper-parameter optimization. In *Proceedings of the 24th International Conference on Neural Information Processing Systems* 2546–2554 (Curran Associates, 2011).
39. Parks, C. L. & Monson, K. L. Automated facial recognition of computed tomography-derived facial images: patient privacy implications. *J. Digital Imaging* **30**, 204–214 (2016).
40. Qaisar Ahmad Al Badawi, A. et al. Towards the AlexNet moment for homomorphic encryption: HCNN, the first homomorphic CNN on encrypted data with GPUs. In *IEEE Transactions on Emerging Topics in Computing* (IEEE, 2020).
41. Wagh, S. et al. Falcon: honest-majority maliciously secure framework for private deep learning. In *Proc. Privacy Enhancing Technologies* 188–208 (Sciendo, 2021).
42. Silva, S., Altmann, A., Gutman, B. & Lorenzi, M. Fed-BioMed: a general open-source frontend framework for federated learning in healthcare. In *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning* (eds Albarqouni, S. et al.) 201–210 (Springer, 2020).
43. 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 (Springer, 2019).
44. Li, W. et al. Privacy-preserving federated brain tumour segmentation. In *International Workshop on Machine Learning in Medical Imaging* 133–141 (Springer, 2019).
45. Lu, M. Y. et al. Federated learning for computational pathology on gigapixel whole slide images. Preprint at https://arxiv.org/abs/2009.10190 (2020).
46. Li, X. et al. Multi-site fMRI analysis using privacy-preserving federated learning and domain adaptation: ABIDE results. *Med. Image Anal.* **65**, 101765 (2020).
47. Kairouz, P. & McMahan, H. B. *Advances and Open Problems in Federated Learning* (Now, 2021).
48. Deng, J. et al. ImageNet: a large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition (CVPR09)* (2009).
49. He, K., Zhang, X., Ren, S. & Sun, J. Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision* 1026–1034 (2015).
50. Buslaev, A. et al. Albumentations: fast and flexible image augmentations. *Information* **11**, 125 (2020).
51. Huang, L., Zhang, C. & Zhang, H. Self-adaptive training: beyond empirical risk minimization. In *Advances in Neural Information Processing Systems* Vol. 33 (NeurIPS, 2020).

52. Kingma, P. & Ba, J. Adam: a method for stochastic optimization. In *Proc. International Conference on Learning Representations* (ICLR, 2015).
53. Rieke, N. et al. The future of digital health with federated learning. *npj Digital Med.* **3**, 119 (2020).
54. Bonawitz, K. et al. Practical secure aggregation for federated learning on user-held data. In *NIPS Workshop on Private Multi-Party Machine Learning* (NIPS, 2016).
55. Wang, S. et al. Adaptive federated learning in resource constrained edge computing systems. *IEEE J. Sel. Areas Commun.* **37**, 1205–1221 (2019).
56. Kirkpatrick, J. et al. Overcoming catastrophic forgetting in neural networks. *Proc. Natl Acad. Sci. USA* **114**, 3521–3526 (2017).
57. Boyle, E., Gilboa, N. & Ishai, Y. Function secret sharing: improvements and extensions. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security* 1292–1303 (2016).
58. Boyle, E., Gilboa, N. & Ishai, Y. Secure computation with preprocessing via function secret sharing. In *Theory of Cryptography Conference* 341–371 (Springer, 2019).
59. 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).
60. He, Z., Zhang, T. & Lee, R. B. Model inversion attacks against collaborative inference. In *Proceedings of the 35th Annual Computer Security Applications Conference* (ACM, 2019).
61. Chicco, D. & Jurman, G. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC Genomics* **21**, 6 (2020).
62. Zhu, L., Liu, Z. & Han, S. Deep leakage from gradients. In *Advances in Neural Information Processing Systems* 14774–14784 (2019).
63. Wang, Y. et al. SAPAG: a self-adaptive privacy attack from gradients. Preprint at https://arxiv.org/abs/2009.06228 (2020).
64. Oh, H. & Lee, Y. Exploring image reconstruction attack in deep learning computation offloading. In *The 3rd International Workshop on Deep Learning for Mobile Systems and Applications: EMDL '19* (ACM, 2019).
65. Gao, W. et al. Privacy-preserving collaborative learning with automatic transformation search. In *Proc. Conference on Computer Vision and Pattern Recognition* (CVPR, 2021).
66. Yanchun, L. & Nanfeng, X. Generative adversarial networks based on denoising and reconstruction regularization. In *2019 IEEE 21st International Conference on High Performance Computing and Communications IEEE 17th International Conference on Smart City IEEE 5th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)* (IEEE, 2019).
67. Kermany, D., Zhang, K. & Goldbaum, M. Large dataset of labeled optical coherence tomography (OCT) and chest X-ray images. *Mendeley Data* https://doi.org/10.17632/rscbjbr9sj.3 (2018).

Acknowledgements

We acknowledge funding from the following sources, funders played no role in the design of the study, the preparation of the manuscript or the decision to publish. The Technical University of Munich, School of Medicine Clinician Scientist Programme (KKF), project reference H14 (G.K.). German Research Foundation, SPP2177/1, German Cancer Consortium (DKTK) and TUM Foundation, Technical University of Munich (R.B. and G.K.). European Community Seventh Framework Programme (FP7/2007-2013 grant no. 339563 – CryptoCloud) and FUI ANBLIC Project (T.R.). UK Research and Innovation London Medical Imaging & Artificial Intelligence Centre for Value Based Healthcare (D.R. and G.K.). Technical University Munich/ Imperial College London Joint Academy of Doctoral Studies (D.U.). We thank B. Farkas for creating Figures 1 and 2 as well as the PriMIA logo; P. Cason and H. Emanuel for assisting with PyGrid debugging, M. Lau for his input, D. Testuggine for his input on differentially private gradient descent, M. Jay for helping with PySyft debugging, N. Remerscheid for his work on the liver segmentation case study, the PySyft and PyGrid development teams for their foundational work and the OpenMined community for their scientific input, contributions and discussion.

Author contributions

G.K. conceived and coordinated the project, evaluated the chest radiography data, helped with PriMIA programming and wrote the paper. A.Z. conceived and developed PriMIA, oversaw PriMIA development, trained the models, performed data analysis and wrote the paper. J.P.-P. helped with project management, oversaw the security and cryptography aspects of PriMIA, supervised model inversion attacks and helped write the paper. T.R. designed and developed PySyft and PyGrid, designed and implemented the AriaNN FSS protocol, helped with PriMIA programming and performed inference latency assessment. D.U. performed the model inversion attacks and helped write the paper. A.T. conceived the OpenMined project, designed and developed PySyft and PyGrid, provided project guidance and assistance and helped with PriMIA development. I.D.L.C.J. developed PyGrid and helped with PriMIA programming. J.M. provided project guidance and prototype code. F.J. performed data curation and helped with data analysis on the chest radiographs. M.-M.S. performed data curation and evaluated the chest radiography data. A.S. and M.M. provided project management, support and guidance. D.R. provided oversight, project management, support, guidance and scientific input, and helped write the paper. R.B. provided oversight, project management, support and guidance, helped with data procurement, provided scientific input and helped write the paper. All authors proof-read and accepted the final version of the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-021-00337-8.

Correspondence and requests for materials should be addressed to R.B.

Peer review information *Nature Machine Intelligence* thanks Haixu Tang, Holger Roth and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.