scientific reports

Check for updates

OPEN

Collaborative training of medical artificial intelligence models with non-uniform labels

Soroosh Tayebi Arasteh[1], Peter Isfort[1], Marwin Saehn[1], Gustav Mueller-Franzes[1], Firas Khader[1], Jakob Nikolas Kather[2,3,4,5], Christiane Kuhl[1], Sven Nebelung[1,6] & Daniel Truhn[1,6]✉

Due to the rapid advancements in recent years, medical image analysis is largely dominated by deep learning (DL). However, building powerful and robust DL models requires training with large multi-party datasets. While multiple stakeholders have provided publicly available datasets, the ways in which these data are labeled vary widely. For Instance, an institution might provide a dataset of chest radiographs containing labels denoting the presence of pneumonia, while another institution might have a focus on determining the presence of metastases in the lung. Training a single AI model utilizing all these data is not feasible with conventional federated learning (FL). This prompts us to propose an extension to the widespread FL process, namely flexible federated learning (FFL) for collaborative training on such data. Using 695,000 chest radiographs from five institutions from across the globe—each with differing labels—we demonstrate that having heterogeneously labeled datasets, FFL-based training leads to significant performance increase compared to conventional FL training, where only the uniformly annotated images are utilized. We believe that our proposed algorithm could accelerate the process of bringing collaborative training methods from research and simulation phase to the real-world applications in healthcare.

Artificial intelligence (AI) is widely expected to reshape medicine in the next decade[1]. The development of robust and clinically useable AI models hinges however on the availability of large and multi-institutional datasets as illustrated by recent publications that have advanced the field in many different areas covering diagnosis and prognosis of diseases in radiological[2,3] and histopathological[4–6] use-cases. One solution to use multi-institutional datasets is conventional federated learning (FL)[7–9] in which the AI model is sent to multiple collaborating centers for training. However, this paradigm requires that the model sees data that is labeled in exactly the same way at each center, i.e. if one center has labeled the presence of pneumonia in its dataset, all the other participating centers also need to label their data with the presence of pneumonia[2,10–13]. While these requirements can be met if the study is carefully planned before the start of data acquisition, in more realistic scenarios, centers often already possess large data that has been individually labeled. In medicine in particular, labels might differ quite dramatically, since the labeling process is complex and since there is no standardized way of labeling the presence of a disease[14–16]. Labels might often be created by two different centers and might be closely related yet appear completely separate to the algorithm that is to be trained. For example, center A might have annotated a dataset of thoracic radiographs with binary labels about the presence of cardiomegaly, while center B might have decided to label another dataset of thoracic radiographs with binary labels about the presence of lung congestion. Both labeling schemes are related and there is mutual information in the labels, since patients with an enlarged heart are more prone to lung congestion, however, conventional FL does not allow to jointly train a model with these data[17].

In this study, we propose flexible federated learning (FFL) as a solution to this impediment on collaboration. In our architecture we divide the classification network into a classification head and a feature extraction

[1]Department of Diagnostic and Interventional Radiology, University Hospital RWTH Aachen, Pauwelsstr. 30, 52074 Aachen, Germany. [2]Department of Medicine III, University Hospital RWTH Aachen, Aachen, Germany. [3]Medical Faculty Carl Gustav Carus, Else Kroener Fresenius Center for Digital Health, Technical University Dresden, Dresden, Germany. [4]Division of Pathology and Data Analytics, Leeds Institute of Medical Research at St James's, University of Leeds, Leeds, UK. [5]Medical Oncology, National Center for Tumor Diseases (NCT), University Hospital Heidelberg, Heidelberg, Germany. [6]These authors contributed equally: Sven Nebelung and Daniel Truhn. ✉email: dtruhn@ukaachen.de

backbone. The backbone is shared between all sites and weights are jointly trained in a FL scheme. The classification head on the other hand can be tailored to the local data with an individual loss function, see Fig. 1. Our goal was to collaboratively and securely train a common backbone network using all data from separate data owners utilizing all available labels. Our hypothesis was that this backbone network learns to extract features that are relevant for the classification of related, but different tasks and that using such a common—and jointly trained—backbone improves classification accuracy for each participating center. We tested this hypothesis on five multicentric datasets comprising a total of over 695,000 thoracic radiographs. The labels assigned to the radiographs from each of the five centers differed, but were related and carried similar information content, thus providing the ideal testing ground for our paradigm.

Results

FFL trains medical classification models with superior performance on non-overlapping labels. We first test our hypothesis that FFL performs superior to conventional FL in a prototypical setting with high-quality data. We utilized two datasets that were both manually labeled by expert radiologists: VinDr-CXR[18,19], a public dataset of thoracic radiographs and UKA-CXR, a private dataset of intensive care thoracic radiographs[20]. Labels for both datasets were different, such that training in a conventional FL setting was not possible. In particular, UKA-CXR has labels for a dedicated set of pathologies for each patient side (e.g., left lung and right lung), while VinDr-CXR utilizes a different set of pathologies and global labels (indicating the presence of a disease in the left *or* the right lung), see Table 1 and Fig. 2A. We chose two distinct label categories in each dataset that have overlapping information content: *cardiomegaly* and *pleural effusion* for VinDr-CXR and *right pleural effusion* and *left pneumonic infiltrates* for the UKA-CXR dataset. Subsequently we trained a ResNet[21] within our FFL scheme on the full UKA-CXR dataset (n = 122,294 training images) and on varying amounts of data from VinDr-CXR (n = 2000, 5000, and 15,000). When tested on a held-out benchmark test set of VinDr-CXR, the average area under the receiver-operator-curve (AUROC) was significantly higher when applying FFL as compared to local training (0.90 ± 0.02 vs. 0.86 ± 0.04; p = 0.001). We observed a similar trend when increasing the training set to n = 5000 (0.92 ± 0.02 vs. 0.90 ± 0.01; p = 0.003) and n = 15,000, i.e., the full dataset (0.95 ± 0.01

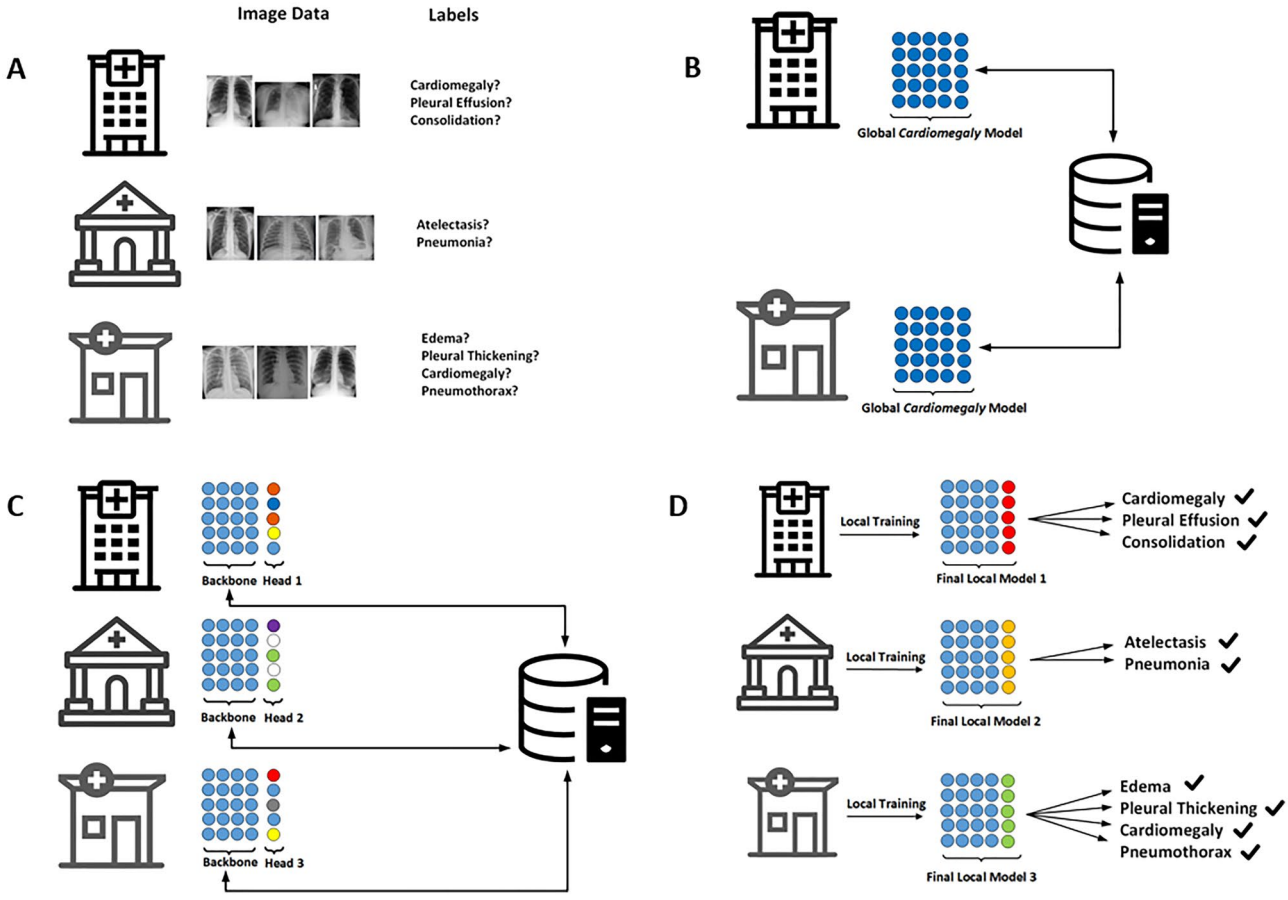

Figure 1. Overview of the flexible federated learning (FFL) process. (**A**) Three separate data centers intend to train AI models for the prediction of different diseases. (**B**) Conventional federated learning: only center 1 and center 3 who have overlapping objectives can collaborate on training a neural network for the detection of cardiomegaly only. (**C**) FFL: all centers collaborate to train a common backbone network and individual classification heads using all their data. (**D**) For classification, each center employs the common backbone and the local classification head.

	Local VinDr 2K	FFL VinDr 2K	Local VinDr 5K	FFL VinDr 5K	Local VinDr 15K	FFL VinDr 15K
AUROC	0.86 ± 0.04	0.90 ± 0.02	0.90 ± 0.01	0.92 ± 0.02	0.94 ± 0.02	0.95 ± 0.01
P-value	0.001		0.003		0.035	

Table 1. Results of the comparison between local and FFL-based training of VinDr-CXR dataset with non-overlapping labels for different training set sizes, tested on the VinDr-CXR benchmark. Average area under the receiver-operator-curve (AUROC) over *cardiomegaly* and *pleural effusion*. The FFL was performed in combination with UKA-CXR dataset of n = 122,294 images with two different labels including *pleural effusion right* and *pneumonic infiltrates right*.

vs. 0.94 ± 0.02; p = 0.035). Thus, in all of these experiments, FFL improved performance as compared to local training.

FFL trains medical classification models with superior performance on partly overlapping labels. Next, we extended the available classification labels to comprise seven categories in each dataset. Part of these labels overlap, e.g., *cardiomegaly*, while others again denote distinct categories. This reflects a more realistic scenario in which both sites have independently labeled their data on common pathologies, but differ in the details of their labeling approach. In particular, for the VinDr-CXR dataset we employ the labels *no finding*, *aortic enlargement*, *pleural thickening*, *cardiomegaly*, *pleural effusion*, *pneumothorax*, and *atelectasis* and for the UKA-CXR dataset we employ *cardiomegaly*, *pleural effusion right*, *pleural effusion left*, *pneumonic infiltrates right*, *pneumonic infiltrates left*, *atelectasis right*, and *atelectasis left*.

By analogy with the first experiment we compared local training to FFL-based training for subsets of n = 2000, n = 5000, and n = 15,000 labeled radiographs. Again, when tested on the held-out benchmark test set of VinDr-CXR, the average AUROC was higher when applying FFL as compared to local training (0.78 ± 0.06 vs. 0.77 ± 0.08; p = 0.340). Similar results were observed when increasing the training set to n = 5000 (0.82 ± 0.05 vs. 0.79 ± 0.07; p = 0.010) and n = 15,000, i.e., the full dataset (0.84 ± 0.05 vs. 0.83 ± 0.09; p = 0.180), see Table 2 and Fig. 2B. Thus, FFL improves performance of classification models on partly overlapping data as compared to local training.

FFL is scalable. To examine if FFL keeps its advantageous properties when trained on truly large and diverse multi-centric datasets, we perform the following experiment: we employ five independent cohorts of thoracic radiographs who each are trained on five labels: (1) the VinDr-CXR dataset (n = 15,000) with labels including *no finding, aortic enlargement, pleural thickening, cardiomegaly,* and *pleural effusion*; (2) the ChestX-ray14[22] dataset (n = 86,524) with labels including *cardiomegaly, effusion, pneumonia, consolidation,* and *no finding;* (3) the CheXpert[23] dataset (n = 128,356) with labels including *cardiomegaly, lung opacity, lung lesion, pneumonia,* and *edema*; (4) the MIMIC-CXR[24,25] dataset (n = 210,652) with labels including *enlarged cardiomediastinum, consolidation, pleural effusion, pneumothorax,* and *atelectasis*; and (5) the UKA-CXR dataset (n = 122,294) with labels including *pleural effusion left, pleural effusion right, cardiomegaly, pneumonic infiltrates left,* and *pneumonic infiltrates right*. It should be noted that only the UKA-CXR and the VinDr-CXR dataset have labels that were manually set by medical experts, while the remaining three datasets have labels extracted from natural language processing of radiological reports. For each of the five cohorts, we performed local training and compared it to training within our FFL framework for hold-out test set of each cohort. In all cohorts, FFL-based training outperformed local training in terms of the average AUROC (VinDr-CXR: 0.885 ± 0.049 vs. 0.867 ± 0.045, p = 0.001; ChestX-ray14: 0.744 ± 0.080 vs. 0.744 ± 0.076, p = 0.363; CheXpert: 0.797 ± 0.061 vs. 0.796 ± 0.064, p = 0.243; MIMIC-CXR: 0.786 ± 0.066 vs. 0.772 ± 0.072, p = 0.004; UKA-CXR 0.918 ± 0.031 vs. 0.916 ± 0.031; p = 0.001, respectively), see Table 3 and Fig. 2C. Thus, even though we observe a saturation effect if the local data comprises thousands of thoracic radiographs, FFL improves performance as compared to local training and can still be used if the data is labeled with vastly different labeling regimes.

Discussion

AI models are becoming increasingly important in modern medicine and are currently reaching a stage in which they can improve patient care and render medical processes more efficient[26–35]. However, the biggest limitation in the development of such data-driven AI models, is their need to access large amounts of annotated data for training. For this, stakeholders need to be able to collaborate on a large scale without jeopardizing patient privacy[36]. Only through such multi-institutional collaboration can robust AI models be trained that make the transition from bench to bedside[36]. Federated learning has been proposed as a solution that allows multiple institutions, individuals, or data providers to collaborate in training AI models without sharing any data with each other[2,37]. This paradigm works well if the data is homogeneously labeled, i.e., if all participating institutions use the same labeling procedure. However, it is the norm rather than the exception that different data providers have similar data but have labeled the data in a seemingly incompatible fashion. Conventional federated learning cannot deal with this situation and new solutions are required. We provide this solution by proposing FFL as a framework for the training on data that is not uniformly labeled. We test this paradigm on a big multi-institutional database of over 680,000 thoracic radiographs from five different hospitals covering the US, Asia and Europe and we find that FFL consistently improves the performance of deep learning models over a wide variety of pathologies.

Our study has limitations. First, we performed all the experiments in a proof-of-concept setup, i.e., within one institutional network, thus the setup is only a simulation of the real situation. However, the setting in which

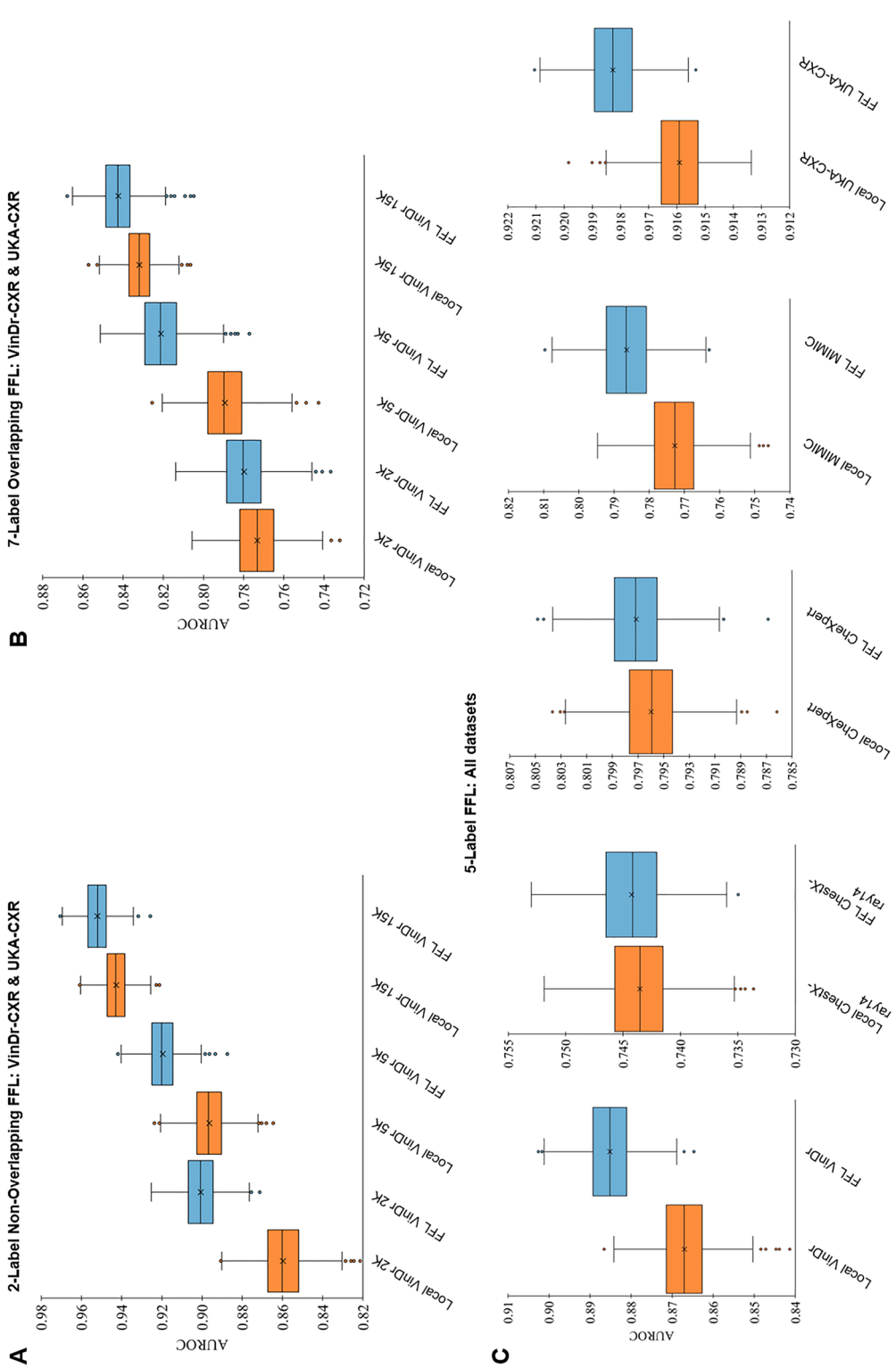

Figure 2. Comparison between flexible federated learning (FFL)-based training and local training of classification models. (**A**) FFL-based training on UKA-CXR data (n = 122,294, labels: *pleural effusion right* and *pneumonic infiltrates right*) and on VinDr-CXR data (on 2K, 5K and 15K images, labels: *cardiomegaly* and *pleural effusion*) if there is no overlap between labels. Performance tested on an independent VinDr-CXR test set. (**B**) Same setup as in (**A**), but training is performed with partially overlapping labels on UKA-CXR (n = 122,294, labels: *cardiomegaly*, *pleural effusion right*, *pleural effusion left*, *pneumonic infiltrates right*, *pneumonic infiltrates left*, *atelectasis right*, and *atelectasis left*) and on VinDr-CXR (on 2K, 5K and 15K images, labels: over *no finding*, *aortic enlargement*, *pleural thickening*, *cardiomegaly*, *pleural effusion*, *pneumothorax*, and *atelectasis*). (**C**) FFL-based training on five different datasets (VinDr-CXR, n = 15,000; ChestX-ray14, n = 86,524; CheXpert, n = 128,356; MIMIC-CXR, n = 210,652; and UKA-CXR, n = 122,294). Testing is performed on the respective held-out test data.

multiple institutions—each with their own network—perform FFL was simulated realistically, by keeping the datasets strictly separate and distributing them to different computing entities. Second, we only tested convolutional neural networks, in particular a ResNet50 architecture. We made that choice to demonstrate our proof-of-concept on one of the most widely used architectures[38–41]. Recently, more general network architectures such as transformers[42–44] have been proposed and may become more important in the future. However, it can be assumed that Transformer architectures may similarly profit from FFL, potentially even stronger than convolutional neural networks since they usually require even bigger data to converge. Third, we only demonstrated FFL for the case of chest radiographs. This is due to the unique availability of public datasets that allow for the study to be performed and to be repeated by other researchers. FFL is not specific to chest X-ray analysis, though. Future works will employ FFL in different domains such as gigapixel imaging in pathology[45,46], and in 3-dimensional volumetric medical imaging such as magnetic resonance imaging and computed tomography.

Our proposed flexible federated learning scheme provides a new way of thinking about collaborative learning. With FFL data does not need to be labeled in an identical fashion at every institution. Rather, machine learning researchers can tap into the vast amount of data that has been labeled heterogeneously and utilize it to train their models on truly big data. This brings secure and privacy-preserving multi-institutional collaboration to the next level and allows the training of models on truly big data.

Methods

Ethics statement. The methods were performed in accordance with relevant guidelines and regulations and approved by ethical committee of the Medical Faculty of RWTH Aachen University. Where necessary, informed consent was obtained from all subjects and/or their legal guardian(s).

Patient cohorts. VinDr-CXR[18,19] is a cohort containing a total of n = 18,000 frontal chest X-ray (CXR) images manually labeled by radiologists. The official training and the benchmark test sets include n = 15,000 and n = 3000 images, respectively. The available labels consist of 27 different diseases including *aortic enlargement, atelectasis, calcification, cardiomegaly, clavicle fracture, consolidation, edema, emphysema, enlarged pulmonary artery, interstitial lung disease, infiltration, lung opacity, lung cavity, lung cyst, mediastinal shift, nodule/mass, pleural effusion, pleural thickening, pneumothorax, pulmonary fibrosis, rib fracture, other lesion, chronic obstructive pulmonary disease, lung tumor, pneumonia, tuberculosis, other diseases* as well as the *no finding* label.

ChestX-ray14[22] dataset contains a total of n = 112,120 frontal X-ray images from 30,805 unique patients[47]. The dataset contains labels for 14 diseases including *atelectasis, cardiomegaly, effusion, infiltration, mass, nodule, pneumonia, pneumothorax, consolidation, edema, emphysema, fibrosis, pleural thickening, hernia* and also for *no finding*. The labels were automatically generated from radiology reports using natural language processing

	Local VinDr 2K	FFL VinDr 2K	Local VinDr 5K	FFL VinDr 5K	Local VinDr 15K	FFL VinDr 15K
AUROC	0.77 ± 0.08	0.78 ± 0.06	0.79 ± 0.07	0.82 ± 0.05	0.83 ± 0.09	0.84 ± 0.05
P-value	0.340		0.010		0.180	

Table 2. Results of the comparison between local and FFL-based training of VinDr-CXR dataset with overlapping labels for different training set sizes, tested on the VinDr-CXR benchmark. Average AUROC values over *no finding*, *aortic enlargement*, *pleural thickening*, *cardiomegaly*, *pleural effusion*, *pneumothorax*, and *atelectasis*. The FFL was performed in combination with UKA-CXR dataset of n = 122,294 images with 7 other labels including *cardiomegaly*, *pleural effusion right*, *pleural effusion left*, *pneumonic infiltrates right*, *pneumonic infiltrates left*, *atelectasis right*, and *atelectasis left*.

Dataset name	Training set size	Included labels	Training setup	AUROC	P-value
VinDr-CXR	n = 15,000	*No finding, aortic enlargement, pleural thickening, cardiomegaly, pleural effusion*	Local	0.867 ± 0.045	0.001
			FFL	0.885 ± 0.049	
ChestX-ray14	n = 83,525	*Cardiomegaly, lung opacity, lung lesion, pneumonia, edema*	Local	0.744 ± 0.076	0.363
			FFL	0.744 ± 0.080	
CheXpert	n = 126,141	*Cardiomegaly, lung opacity, lung lesion, pneumonia, edema*	Local	0.796 ± 0.064	0.243
			FFL	0.797 ± 0.061	
MIMIC-CXR-JPG-v2.0	n = 237,972	*Enlarged cardiomediastinum, consolidation, pleural effusion, pneumothorax, atelectasis*	Local	0.772 ± 0.072	0.004
			FFL	0.786 ± 0.066	
UKA-CXR	n = 122,297	*Pleural effusion left, pleural effusion right, cardiomegaly, pneumonic infiltrates left, pneumonic infiltrates right*	Local	0.916 ± 0.031	0.001
			FFL	0.918 ± 0.031	

Table 3. Results of the comparison between local and FFL-based training for 5 different datasets. Average AUROC values over all included labels for each dataset, tested on the test benchmark of the corresponding dataset. The FFL process for each dataset was performed in combination with the other 4 datasets including 5 different labels for each dataset.

techniques. We adopted the original proposed benchmark test subset including n = 25,596 images and utilized the rest of the n = 86,524 images as training.

CheXpert[23] dataset v1.0 contains n = 224,316 chest radiographs of 65,240 patients. Out of these, 157,676 images are frontal chest radiographs. All the images are automatically labeled based on radiology reports utilizing a natural-language-processing-based labeler. The available labels include *atelectasis, cardiomegaly, consolidation, edema, enlarged cardiomediastinum, fracture, lung lesion, lung opacity, pleural effusion, pleural other, pneumonia, pneumothorax, support devices,* and *no finding*. Unlike ChestX-ray14 and VinDr-CXR datasets which consist of binary labels, CheXpert labels include 4 different classes of "positive", "negative", "uncertain", and "not mentioned in the reports". The "uncertain"' label can capture both the uncertainty of a radiologist in the diagnosis as well as ambiguity inherent in the report[23]. We divided the dataset to 80% training and 20% test.

MIMIC Chest X-ray JPG (MIMIC-CXR-JPG) database v2.0.0[24,25] consists of 377,110 CXR images including n = 210,652 frontal images for training, 1691 frontal images for validation, and 2844 frontal images for test. MIMIC-CXR-JPG provides free-text radiology reports associated with the images. Furthermore, 2 separate sets of labels generated using the labelers from CheXpert[23] and NegBio[48], an open-source rule based tool for negation and uncertain detection in radiology reports, are provided. We used the labels generated based on the CheXpert labeler in order to be consistent with the CheXpert dataset.

Finally, we employed UKA-CXR[20], a large internal dataset of chest radiographs from RWTH Aachen University Hospital. The dataset consists of n = 193,361 frontal CXR images, all manually labeled by the radiologists. The available labels include *pleural effusion*, *pneumonic infiltrates*, *atelectasis*, and *pneumothorax*, each one separately for *right* and *left* parts, and *cardiomegaly*. The labeling system for *cardiomegaly* included 5 classes of "normal", "uncertain", "borderline", "enlarged", and "massively enlarged". For the rest of the labels, 5 classes of "negative", "uncertain", "mild", "moderate", and "severe" were used. Data were split into 75% training and 25% testing data using patient-wise stratification, but otherwise completely random allocation. It is worth noting that, in none of the datasets, there was any overlap between training and test cohorts.

Data pre-processing. ChestX-ray14, CheXpert, and MIMIC-CXR-JPG-v2.0 datasets were readily available in PNG standard formats. All the image pixels of the datasets which were only available in digital imaging and communications in medicine (DICOM) format, i.e., VinDr-CXR and UKA-CXR, were extracted and converted into PNG. The DICOM field PhotometricInterpretation was used to determine whether the pixel values were inverted, and if necessary images were inverted[24]. Only the frontal images were used during the experiments. We followed the same pre-processing scheme for all datasets. All the images were resized to (512×512) resolution. Afterwards, a normalization scheme as described before by Johnson et al.[24] was utilized by subtracting the lowest value in the image, dividing by the highest value in the shifted image, truncating values, and converting the result to an unsigned integer, i.e., the range of [0, 255]. Finally, using Python's OpenCV library, histogram equalization was performed by shifting pixel values towards 0 or towards 255[24].

A binary diagnosis paradigm was chosen for all the experiments. ChestX-ray14 and VinDr-CXR datasets included binary labels by design. For the CheXpert dataset (and subsequently for the MIMIC-CXR-JPG-v2.0 dataset), all the 3 classes of "negative", "uncertain", and "not mentioned in the reports" were treated as the negative class and only the original "positive" class was treated as the positive class. For the UKA-CXR dataset, the "negative" and "uncertain" classes ("normal" and "uncertain" for *cardiomegaly*) were treated as negative, while the "mild", "moderate", and "severe" classes ("borderline", "enlarged", and "massively enlarged" for *cardiomegaly*) were treated as positive.

Flexible federated learning (FFL) scheme. The backbone architecture of all networks at all sites was identical by using shared weights of a ResNet50[21]. After each iteration, the locally updated weights were pooled and averaged and the updated backbone weights were sent back to the sites for the next iteration.

The network head, i.e., the classification layer, was individual to each site and its updates were not aggregated during FFL. This allowed for different classification problems to be backpropagated at each site and made it possible to use data with labels that are unique to each site. For the classification head we employed a fully connected neural network layer as described below. After convergence each site was allowed to perform additional training rounds without central aggregation (i.e., neither of the backbone, nor the classification head) for fine-tuning.

The situation with multiple separate data centers was simulated by isolating each center on a virtual machine within the same network and on the same bare-metal computer. This is slightly different from the real situation in which virtual machines would be set up in different networks, but linked through a common virtual private network. However, there is no principal difference to the real setup.

Deep learning training procedure. We performed data augmentation during training by applying medio-lateral flipping with a probability of 0.5 and random rotation in the range of [0, 10] degrees. The ResNet50 architecture was employed as a backbone architecture. We followed the same 50-layer implementation proposed by He et al.[21], where the first layer included a (7×7) convolution producing an output image with 64 channels. The inputs to the network were $(512 \times 512 \times 3)$ images in batches of size 16. Last layer included a linear layer which reduced the (2048×1) output feature vectors to the desired number of diseases to be predicted for each case. The sigmoid function was utilized to convert the output predictions to individual class probabilities. The full network contained a total of 23,512,130 trainable parameters.

All models were optimized using the Adam[49] optimizer. During FFL training of the backbone, a learning rate of 5×10^{-5} was chosen. Whereas a learning rate of 9×10^{-5} was selected for the training of individual classification heads. As loss function, we chose the binary weighted cross-entropy with inverted class frequencies of the training data as loss weights. It is worth mentioning that even though in our implementation the choice

of the loss function type was the same in all networks, as the objectives were not the same, every classification head had an independent loss function.

Quantitative evaluation. The area under the receiver-operator-curve (AUROC) was used as the primary evaluation metric. Accuracy, sensitivity, and specificity were utilized as further evaluation metrics. We reported the average AUROC over all the labels for each experiment, while the individual AUROC of different labels, as well as accuracy, sensitivity, and specificity are reported in the supplemental material (see Tables S3–S5). It should be noted that we followed a multilabel classification paradigm, where multiple diseases could have positive labels given an image. Therefore, we optimized the average performance of the networks over all the diseases, as opposed to optimizing per disease.

Statistical analysis. Bootstrapping was employed with 1000 redraws for each measure to determine the statistical spread and calculate p-values for differences[50]. For the calculation of sensitivity and specificity scores, a threshold was chosen according to Youden's criterion[51], i.e., a threshold that maximized (true positive rate–false positive rate).

Hardware. The hardware used in our experiments were Intel CPUs with 18 cores and 32 GB RAM and Nvidia RTX 6000 GPUs with 24 GB memory.

Data availability

ChestX-ray14 data is publicly available under https://www.v7labs.com/open-datasets/chestx-ray14. VinDr-CXR and MIMIC-CXR-JPG data are restricted-access resources, which can be accessed from PhysioNet by agreeing to its data protection requirements under https://physionet.org/content/vindr-cxr/1.0.0/ and https://physionet.org/content/mimic-cxr-jpg/2.0.0/, respectively. CheXpert data can be requested from Stanford University at https://stanfordmlgroup.github.io/competitions/chexpert/. The UKA-CXR data is not publicly accessible as it is internal data of patients of the University Hospital RWTH Aachen. A reasonable request from the corresponding author is required for accessing the data.

Code availability

All source codes for training and evaluation of the deep neural networks, collaborative learning, data augmentation, CXR image analysis, and preprocessing is publicly available at https://github.com/tayebiarasteh/chestx. All code for the experiments was developed in Python 3.8 using the PyTorch 1.4 framework. The collaboratibe learning process was developed using PySyft 0.2.9[52].

Received: 11 January 2023; Accepted: 11 April 2023
Published online: 13 April 2023

References

1. Rajpurkar, P., Chen, E., Banerjee, O. & Topol, E. J. AI in health and medicine. *Nat. Med.* **28**, 31–38 (2022).
2. Dayan, I. *et al.* Federated learning for predicting clinical outcomes in patients with COVID-19. *Nat. Med.* **27**, 1735–1743 (2021).
3. Han, T. *et al.* Image prediction of disease progression for osteoarthritis by style-based manifold extrapolation. *Nat. Mach. Intell.* **4**, 1029–1039 (2022).
4. Saldanha, O. L. *et al.* Swarm learning for decentralized artificial intelligence in cancer histopathology. *Nat. Med.* **28**, 1232–1239 (2022).
5. Schrammen, P. L. *et al.* Weakly supervised annotation-free cancer detection and prediction of genotype in routine histopathology. *J. Pathol.* **256**, 50–60 (2022).
6. Ghaffari Laleh, N. *et al.* Benchmarking weakly-supervised deep learning pipelines for whole slide classification in computational pathology. *Med. Image Anal.* **79**, 102474 (2022).
7. Konečný, J., McMahan, H. B., Ramage, D. & Richtárik, P. Federated optimization: Distributed machine learning for on-device intelligence. Preprint at http://arxiv.org/abs/1610.02527 (2016).
8. Konečný, J. *et al.* Federated learning: Strategies for improving communication efficiency. Preprint at http://arxiv.org/abs/1610.05492 (2017).
9. McMahan, H. B., Moore, E., Ramage, D., Hampson, S. & Arcas, B. A. Y. Communication-efficient learning of deep networks from decentralized data. Preprint at http://arxiv.org/abs/1602.05629 (2017).
10. Banabilah, S., Aloqaily, M., Alsayed, E., Malik, N. & Jararweh, Y. Federated learning review: Fundamentals, enabling technologies, and future applications. *Inf. Process. Manag.* **59**, 103061 (2022).
11. Kairouz, P. *et al.* Advances and open problems in federated learning. *Found. Trends Mach. Learn.* **14**, 1–210 (2021).
12. Kaissis, G. *et al.* End-to-end privacy preserving deep learning on multi-institutional medical imaging. *Nat. Mach. Intell.* **3**, 473–484 (2021).
13. 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).
14. Qayyum, A., Ahmad, K., Ahsan, M. A., Al-Fuqaha, A. & Qadir, J. Collaborative federated learning for healthcare: Multi-modal COVID-19 diagnosis at the edge. *IEEE Open J. Comput. Soc.* **3**, 172–184 (2022).
15. Sheller, M. J. *et al.* Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data. *Sci. Rep.* **10**, 12598 (2020).
16. Xu, J. *et al.* Federated learning for healthcare informatics. *J. Healthc. Inform. Res.* **5**, 1–19 (2021).
17. Ruan, Y., Zhang, X., Liang, S.-C. & Joe-Wong, C. Towards flexible device participation in federated learning. Preprint at http://arxiv.org/abs/2006.06954 (2021).
18. Nguyen, H. Q., Pham, H. H., Tuan Linh, L., Dao, M. & Khanh, L. VinDr-CXR: An open dataset of chest X-rays with radiologist annotations.
19. Nguyen, H. Q. *et al.* VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations. *Sci. Data* **9**, 429 (2022).
20. Khader, F. *et al.* Artificial intelligence for clinical interpretation of bedside chest radiographs. *Radiology.* https://doi.org/10.1148/radiol.220510 (2022).

21. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 770–778. https://doi.org/10.1109/CVPR.2016.90 (IEEE, 2016).
22. Wang, X. *et al.* ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 3462–3471. https://doi.org/10.1109/CVPR.2017.369 (2017).
23. Irvin, J. *et al.* CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proc. AAAI Conf. Artif. Intell.* **33**, 590–597 (2019).
24. Johnson, A. E. W. *et al.* MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci. Data* **6**, 317 (2019).
25. Johnson, A. E. W. *et al.* MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. Preprint at http://arxiv.org/abs/1901.07042 (2019).
26. Truhn, D. *et al.* Encrypted federated learning for secure decentralized collaboration in cancer image analysis. *MedRxiv.* https://doi.org/10.1101/2022.07.28.22277288 (2022).
27. McKinney, S. M. *et al.* International evaluation of an AI system for breast cancer screening. *Nature* **577**, 89–94 (2020).
28. Killock, D. AI outperforms radiologists in mammographic screening. *Nat. Rev. Clin. Oncol.* **17**, 134–134 (2020).
29. Kleppe, A. *et al.* Designing deep learning studies in cancer diagnostics. *Nat. Rev. Cancer* **21**, 199–211 (2021).
30. Boehm, K. M., Khosravi, P., Vanguri, R., Gao, J. & Shah, S. P. Harnessing multimodal data integration to advance precision oncology. *Nat. Rev. Cancer* **22**, 114–126 (2022).
31. Elemento, O., Leslie, C., Lundin, J. & Tourassi, G. Artificial intelligence in cancer research, diagnosis and therapy. *Nat. Rev. Cancer* **21**, 747–752 (2021).
32. Echle, A. *et al.* Deep learning in cancer pathology: A new generation of clinical biomarkers. *Br. J. Cancer* **124**, 686–696 (2021).
33. Yao, T. *et al.* Compound figure separation of biomedical images with side loss. In *Deep Generative Models, and Data Augmentation, Labelling, and Imperfections* Vol. 13003 (eds Engelhardt, S. *et al.*) 173–183 (Springer, 2021).
34. Zhao, M. *et al.* VoxelEmbed: 3D instance segmentation and tracking with voxel embedding based deep learning. In *Machine Learning in Medical Imaging* Vol. 12966 (eds Lian, C. *et al.*) 437–446 (Springer, 2021).
35. Jin, B., Cruz, L. & Goncalves, N. Deep facial diagnosis: Deep transfer learning from face recognition to facial diagnosis. *IEEE Access* **8**, 123649–123661 (2020).
36. Bhinder, B., Gilvary, C., Madhukar, N. S. & Elemento, O. Artificial intelligence in cancer research and precision medicine. *Cancer Discov.* **11**, 900–915 (2021).
37. Ng, D., Lan, X., Yao, M.M.-S., Chan, W. P. & Feng, M. Federated learning: A collaborative effort to achieve better medical imaging models for individual sites that have small labelled datasets. *Quant. Imaging Med. Surg.* **11**, 852–857 (2021).
38. Victor Ikechukwu, A., Murali, S., Deepu, R. & Shivamurthy, R. C. ResNet-50 vs VGG-19 vs training from scratch: A comparative analysis of the segmentation and classification of Pneumonia from chest X-ray images. *Glob. Transit. Proc.* **2**, 375–381 (2021).
39. Kora, P. *et al.* Transfer learning techniques for medical image analysis: A review. *Biocybern. Biomed. Eng.* **42**, 79–107 (2022).
40. Nabavi, S. *et al.* Medical imaging and computational image analysis in COVID-19 diagnosis: A review. *Comput. Biol. Med.* **135**, 104605 (2021).
41. Yang, J., Shi, R. & Ni, B. MedMNIST classification decathlon: A lightweight AutoML benchmark for medical image analysis. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)* 191–195. https://doi.org/10.1109/ISBI48211.2021.9434062 (IEEE, 2021).
42. Dosovitskiy, A. *et al.* An image is worth 16 × 16 words: Transformers for image recognition at scale. Preprint at http://arxiv.org/abs/2010.11929 (2021).
43. Liu, Z. *et al.* Swin transformer: Hierarchical vision transformer using shifted windows. Preprint at http://arxiv.org/abs/2103.14030 (2021).
44. Han, K. *et al.* A survey on vision transformer. *IEEE Trans. Pattern Anal. Mach. Intell.* https://doi.org/10.1109/TPAMI.2022.3152247 (2022).
45. Kather, J. N. *et al.* Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLoS Med.* **16**, e1002730 (2019).
46. Kather, J. N. *et al.* Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nat. Med.* **25**, 1054–1056 (2019).
47. Kumar, P., Grewal, M. & Srivastava, M. M. Boosted cascaded convnets for multilabel classification of thoracic diseases in chest radiographs. In *Image Analysis and Recognition* Vol. 10882 (eds Campilho, A. *et al.*) 546–552 (Springer, 2018).
48. Peng, Y. *et al.* NegBio: A high-performance tool for negation and uncertainty detection in radiology reports. *AMIA Jt. Summits Transl. Sci.* **2017**, 188–196 (2018).
49. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. Preprint at http://arxiv.org/abs/1412.6980 (2017).
50. Konietschke, F. & Pauly, M. Bootstrapping and permuting paired t-test type statistics. *Stat. Comput.* **24**, 283–296 (2014).
51. Unal, I. Defining an optimal cut-point value in ROC analysis: An alternative approach. *Comput. Math. Methods Med.* **2017**, 3762651 (2017).
52. Ziller, A. *et al.* PySyft: A library for easy federated learning. In *Federated Learning Systems* Vol. 9659 (eds ur Rehman, M. H. & Gaber, M. M.) 111–113 (Springer, 2021).

Acknowledgements

The authors would like to thank the data providers of VinDr-CXR, ChestX-ray14, CheXpert, and MIMIC-CXR-JPG for providing them access to their public datasets. They additionally acknowledge the support by NVIDIA who provided our group with two RTX6000 GPUs.

Author contributions

The formal analysis was conducted by S.T.A. and D.T. and the original draft was written by S.T.A. and reviewed and corrected by D.T. and S.N. The software was mainly developed by S.T.A.; F.K., G.M.F., J.N.K. and D.T. provided technical expertise, P.I., M.S., J.N.K., C.K., S.N. and D.T. provided clinical expertise. The experiments were performed by S.T.A. All authors read the manuscript and contributed to the interpretation of the results and agreed to the submission of this paper.

Funding

Open Access funding enabled and organized by Projekt DEAL. This work was (partially) funded/supported by the RACOON network under BMBF Grant Number 01KX2021. JNK is supported by the German Federal Ministry of Health (DEEP LIVER, ZMVI1-2520DAT111) and the Max-Eder-Programme of the German Cancer

Aid (Grant #70113864), the German Federal Ministry of Education and Research (PEARL, 01KD2104C), and the German Academic Exchange Service (SECAI, 57616814).

Competing interests

JNK reports consulting services for Owkin, France, Panakeia, UK and DoMore Diagnostics, Norway and has received honoraria for lectures by MSD, Eisai and Fresenius. The other authors do not have any competing interests to disclose.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-33303-y.

Correspondence and requests for materials should be addressed to D.T.

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.