Federated Learning for Breast Density Classification: A Real-World Implementation

Holger R. Roth[1(✉)], Ken Chang[2], Praveer Singh[2], Nir Neumark[2], Wenqi Li[1], Vikash Gupta[3], Sharut Gupta[2], Liangqiong Qu[4], Alvin Ihsani[1], Bernardo C. Bizzo[2], Yuhong Wen[1], Varun Buch[2], Meesam Shah[5], Felipe Kitamura[6], Matheus Mendonça[6], Vitor Lavor[6], Ahmed Harouni[1], Colin Compas[1], Jesse Tetreault[1], Prerna Dogra[1], Yan Cheng[1], Selnur Erdal[3], Richard White[3], Behrooz Hashemian[2], Thomas Schultz[2], Miao Zhang[4], Adam McCarthy[2], B. Min Yun[2], Elshaimaa Sharaf[2], Katharina V. Hoebel[7], Jay B. Patel[7], Bryan Chen[7], Sean Ko[7], Evan Leibovitz[2], Etta D. Pisano[2], Laura Coombs[5], Daguang Xu[1], Keith J. Dreyer[2], Ittai Dayan[2], Ram C. Naidu[2], Mona Flores[1], Daniel Rubin[4], and Jayashree Kalpathy-Cramer[2]

[1] NVIDIA, Santa Clara, USA
hroth@nvidia.com
[2] Massachusetts General Hospital, Boston, USA
[3] Mayo Clinic, Jacksonville, USA
[4] Stanford University, Stanford, USA
[5] American College of Radiology, Reston, USA
[6] Diagnósticos da América (DASA), São Paulo, Brazil
[7] Massachusetts Institute of Technology, Cambridge, USA

Abstract. Building robust deep learning-based models requires large quantities of diverse training data. In this study, we investigate the use of federated learning (FL) to build medical imaging classification models in a real-world collaborative setting. Seven clinical institutions from across the world joined this FL effort to train a model for breast density classification based on Breast Imaging, Reporting & Data System (BI-RADS). We show that despite substantial differences among the datasets from all sites (mammography system, class distribution, and data set size) and without centralizing data, we can successfully train AI models in federation. The results show that models trained using FL perform 6.3% on average better than their counterparts trained on an institute's local data alone. Furthermore, we show a 45.8% relative improvement in the models' generalizability when evaluated on the other participating sites' testing data.

Keywords: Federated learning · Breast density classification · BI-RADS · Mammography

S. Albarqouni et al. (Eds.): DART 2020/DCL 2020, LNCS 12444, pp. 181–191, 2020.
https://doi.org/10.1007/978-3-030-60548-3_18

1 Introduction

Advancements in medical image analysis over the last several years have been dominated by deep learning (DL) approaches. However, it is well known that DL requires large quantities of data to train robust and clinically useful models [5,6]. Often, hospitals and other medical institutes need to collaborate and host centralized databases for the development of clinically useful models. This overhead can quickly become a logistical challenge and usually requires a time-consuming approval process due to data privacy and ethical concerns associated with data sharing in healthcare [12]. Even when these challenges can be addressed, data is valuable, and institutions may prefer not to share full datasets. Furthermore, medical data can be large, and it may be prohibitively expensive to acquire storage for central hosting [4]. One approach to combat the data sharing hurdles is federated learning (FL) [16], where only model weights are shared between participating institutions without sharing the raw data.

To investigate the performance of FL in the real world, we conducted a study to develop a breast density classification model using mammography data. An international group of hospitals and medical imaging centers joined this collaborative effort to train the model in purely data-decentralized fashion without needing to share any data. This is in contrast to previous studies in which the FL environment was only simulated [14,21]. We do not have centralized training experiments as references before starting the FL tasks, which places higher requirements on the robustness of the algorithms and selection of hyperparameters.

1.1 Related Works

Breast Density Scoring: The classification of breast density is quintessential for breast imaging to estimate the extent of fibroglandular tissue related to the patient's risk of developing breast cancer [2,19]. Women with a high mammographic breast density (>75%) have a four- to five-fold increase in risk for breast cancer compared to women having a lower breast density [3,26]. This condition affects roughly half of American women between the ages of 40 to 74 [7,25]. Patients identified with dense breast tissue may have masked tumors and benefit from supplemental imaging such as MRI or ultrasound [13]. High mammographic breast density impairs the sensitivity and specificity of breast cancer screening, possibly because (small) malignant lesions are not detectable even when they are present [17]. The standard evaluation metric for reporting breast density is the Breast Imaging Reporting and Data System (BI-RADS), based on 2D mammography [22]. Scans are categorized into one of four classes: (a) fatty, (b) scattered, (c) heterogeneously dense, and (d) extremely dense.

Due to the subjective nature of the BI-RADS criteria, there can be substantial inter-rater variability between pairs of clinicians. Sprague et al. [24] found that the likelihood of a mammogram being read as dense varies from radiologist to radiologist between 6.3% to 84.5%. Ooms et al. find that the overall agreement between four observers (inter-rater agreement) in terms of the overall

weighted kappa was 0.77 [17]. Another study reported the inter-rater variability to be simple kappa = 0.58 among 34 community radiologists [23]. Even the intra-rater agreement in the assessment of BI-RADS breast density can be relatively low. Spayne et al. [23] showed that the intra-rater agreement was below 80% when evaluating the same mammography exam within a 3- to 24- month period. Recent work on applying DL for mammography breast density classification [13] achieved a linear kappa of 0.67 when comparing the DL model's predictions to the assessments of the original interpreting radiologist.

Federated Learning: Federated learning has recently been described as being instrumental for the future of digital health [20]. FL enables collaborative and decentralized DL training without sharing any raw patient data [16]. Each client in FL trains locally on their data and then submits their model parameters to a server that accumulates and aggregates the model updates from each client. Once a certain number of clients have submitted their updates, the aggregated model parameters are redistributed to the clients, and a new round of local training starts. While out of the scope of this work, FL can also be combined with additional privacy-preserving measures to avoid potential reconstruction of training data through model inversion if the model parameters would be exposed to an adversary [14]. Recent works have shown the applicability of FL to medical imaging tasks [14,21]. The security and privacy-preserving aspects of federated machine learning in medical imaging have been discussed in more detail by Kaissis et al. [9].

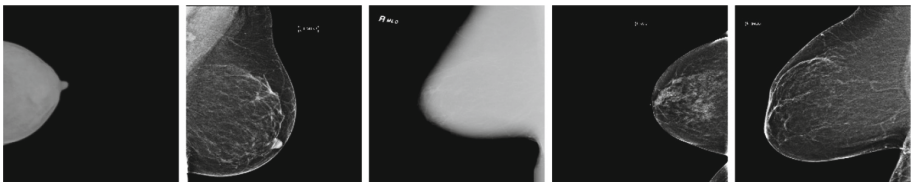

Fig. 1. Mammography data examples from different sites after resizing the original images to a resolution of 224 × 224. No special normalization was applied in order to keep the scanners' original intensity distribution that can be observed in 4.

2 Method

We implemented our FL approach in a real-world setting with participation from seven international clients.

Datasets: The mammography data was retrospectively selected after Institutional Review Board (IRB) approval as part of standard mammography screening protocols. The BI-RADS breast density class from the original interpreting radiologist was collected from the reports available in the participating hospitals' medical records and includes images from digital screening mammography

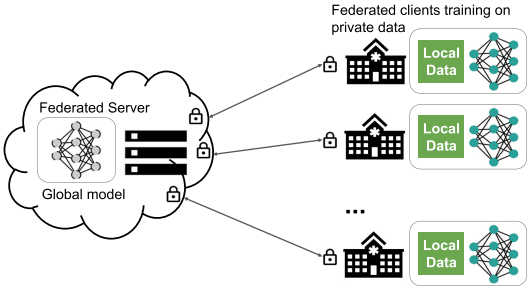

Fig. 2. Federated learning in medical imaging. The central server communicates with clients from multi-national institutions without exchanging any sensitive raw data. Still, the global model benefits from weights and gradients from clients' local models to achieve higher overall performance.

(Fig. 1). Clients 1 to 3 utilized the multi-institutional dataset previously described in [18], which was split by the digital mammography system used to acquire the image to account for different dataset sources.

Each client's data exhibited their own characteristics of detector type, image resolution, and mammography type. Furthermore, the number of training images varied significantly among clients, as shown in Table 1. The distributions of the different BI-RADS categories were markedly different at some clients but generally followed the distribution known from the literature, with more images in the categories b and c [18], see Fig. 3. Given these differences that are quite typical for real-world multi-institutional datasets, we can see that the data used in this study is non-independent and identically distributed (non-IID).

Intensity distributions among different sites also varied markedly, as can be observed in Fig. 4. This variance is due to the differences in imaging protocols and digital mammography systems used at each data contributing site. No attempt to consolidate these differences was made in our study to investigate the domain shift challenges proposed by this non-IID data distribution.

Table 1. Dataset characteristics at each client. Image resolution is shown in megapixels (MP).

Institution	Image resolution	Detector type	Image type	Bits	# Train	# Val.	# Test
client1	23.04	Direct	2D	12	22933	3366	6534
client2	.02 to 4.39	Direct	2D	12	8365	1216	2568
client3	4.39 to 13.63	Direct	2D	14	44115	6336	12676
client4	4 to 28	Direct/Scintillator	2D	12	7219	1030	2069
client5	8.48 to 13.63	Direct	2D	12	6023	983	1822
client6	8.6 to 13.63	Direct	2D	12	6874	853	1727
client7	1 to 136	Direct/Scintillator	2D/tomosynthesis	10/12	4021	664	1288

Fig. 3. Class distribution at different client sites as a fraction of their total data.

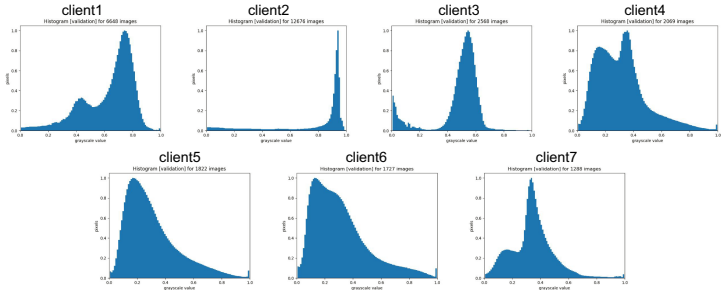

Fig. 4. Intensity distribution at different sites.

Client-Server-Based Federated Learning: In its typical form, FL utilizes a client-server setup (Fig. 2). Each client trains the same model architecture locally on their data. Once a certain number of clients finishes a round of local training, the updated model weights (or their gradients) are sent to the server for aggregation. After aggregation, the new weights on the server are re-distributed to the clients, and the next round of local model training begins. After a certain number of FL rounds, the models at each client converge. Each client is allowed to select their locally best model by monitoring a certain performance metric on a local hold-out validation set. The client can select either the global model returning from the server after averaging or any intermediate model considered best during local training based on their validation metric. In our experiments, we implement the `FederatedAveraging` algorithm proposed in [16]. While there exist variations of this algorithm to address particular learning tasks, in its most general form, FL tries to minimize a global loss function $\mathcal{L}$ which can be a weighted combination of K local losses $\{\mathcal{L}_k\}_{k=1}^{K}$, each of which is computed on a client k's local data. Hence, FL can be formulated as the task of finding the model parameters ϕ that minimize L given some local data $X_k \in X$, where X would be the combination of all local datasets.

$$\min_{\phi} \mathcal{L}(X;\phi) \quad \text{with} \quad \mathcal{L}(X;\phi) = \sum_{k=1}^{K} w_k \, \mathcal{L}_k(X_k;\phi), \tag{1}$$

where $w_k > 0$ denotes the weight coefficients for each client k, respectively. Note that the local data X_k is never shared among the different clients. Only

the model weight differences are accumulated and aggregated on the server as shown in Algorithm 1.

Algorithm 1 Client-server federated learning with `FederatedAveraging` [16,14]. T is the number of federated learning rounds and n_k is the number of `LocalTraining` iterations minimizing the local loss $\mathcal{L}_k(X_k; \phi^{(t-1)})$ for a client k.

procedure FEDERATED LEARNING
  Initialize weights: $\phi^{(0)}$
  for $t \leftarrow 1 \cdots T$ **do**
    for *client* $k \leftarrow 1 \cdots K$ **do** ▷ *Executed in parallel*
      Send $\phi^{(t-1)}$ to client k
      Receive $(\Delta\phi_k^{(t)}, n_k)$ from client's `LocalTraining`$(\phi^{(t-1)})$
    end for
    $\phi_k^{(t)} \leftarrow \phi^{(t-1)} + \Delta\phi_k^{(t)}$
    $\phi^{(t)} \leftarrow \frac{1}{\sum_k n_k} \sum_k (n_k \cdot \phi_k^{(t)})$
  end for
  return $\phi^{(t)}$
end procedure

In this work, we choose a softmax cross-entropy loss which is commonly used for multi-class classification tasks: $\mathcal{L}_0 = -\sum_{i=1}^{C} y_i \log(p_i)$; with $C = 4$ being the number of classes. Here, p_i is the predicted probability for a class i from the final softmax activated output layer of our neural network $f(x)$ and y is the one-hot encoded ground truth label for a given image.

Classification Model and Implementation: In this work, we do not focus on developing a new model architecture but instead focus on showing how FL works in a real-world collaborative training situation. We implement a DenseNet-121 [8] model as a backbone and append a fully-connected layer with four outputs to its last feature layer to classify a mammography image as one of the four BI-RADS categories. The FL framework is implemented in Tensorflow[1] and utilizes the NVIDIA Clara Train SDK[2] to enable the communication between server and clients as well as to standardize the training configuration among clients. Each client employed an NVIDIA GPU with at least 12 GB memory.

All mammography images were normalized to an intensity range of $[0 \ldots 1]$ and resampled to a resolution of 224×224. We include both left and right breast images and all available views (craniocaudal and mediolateral oblique) in training. Each client separated their dataset into training, validation, and testing sets on the patient level (see Table 1). At inference time, predictions from all images from a given patient were averaged together to give a patient-level prediction.

Each client trained for one epoch before sending their updated model weights to the server for aggregation, and the server waited for all clients before performing a weighted sum of the clients' weight differences. We used initial learning rates of 1e−4 with step-based learning rate decay, Adam optimization for each

[1] https://www.tensorflow.org/.
[2] https://developer.nvidia.com/clara.

client, and model weight decay. A mini-batch of size 32 was sampled from the dataset such that all categories were equally represented during training. Random spatial flips, rotations between ±45°, and intensity shifts were used as on-the-fly image augmentation to avoid overfitting to the training data. The FL training was run for 300 rounds of local training and weight aggregations, which took about 36 h. After the FL training is finished, each client's best local model is shared with all other clients and tested on their test data to evaluate the models' generalizability.

In an additional experiment, we use the locally best models each client receives after FL to execute a second round of local fine-tuning based on this model. This additional "adaptation" step can improve a client's model on their local data.

Evaluation Metric: We utilize Cohen's linear weighed kappa[3] to evaluate the locally best models' performance before and after federated learning in comparison with the radiologists' ground truth assessments. The kappa score is a number between −1 and 1. Scores above 0.8 are generally considered very good agreement, while zero or lower would mean no agreement (practically random assignment of labels). A kappa of 0.21 to 0.40, 0.41 to 0.60, and 0.61 to 0.80 represents fair, moderate, and substantial agreement, respectively [11]. The kappa measure has been chosen to be directly comparable to previous literature on breast density classification in mammography [13,17,23].

3 Results

In Table 2, we show the performance of locally best models (selected by best validation score on local data) using local training data alone as well as after federated learning. On average, a 6.3% relative improvement can be observed when the model is applied to a client's test data (diag. mean). We also observe a general improvement of these best local models applied to the different clients' test data. Here, the generalizability (off-diag. mean) of the models improved by 45.8% on average.

Figure 5 summarizes the kappa scores for local training, after FL, including after local fine-tuning, which improves a given model's performance on the client's local test data in all but one client.

[3] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.cohen_kappa_score.html.

Table 2. Performance of locally best models (selected by best validation score on local data) using (a) local training data alone and (b) after federated learning.

(a) Local:		Test						
	client	1	2	3	4	5	6	7
Train	1	0.62	0.59	0.44	0.02	0.02	-0.01	0.04
	2	0.15	0.56	0.02	-0.01	-0.00	0.00	-0.01
	3	0.19	0.01	0.64	0.02	0.07	0.00	0.05
	4	0.11	0.02	-0.00	0.63	0.52	0.61	0.50
	5	-0.00	-0.01	-0.03	0.54	0.62	0.65	0.31
	6	0.01	0.11	-0.02	0.49	0.59	0.71	0.32
	7	0.03	0.05	-0.05	0.40	0.37	0.46	0.69
	diag. mean							**0.64**
	off-diag. mean							**0.18**

(b) Federated:		Test						
	client	1	2	3	4	5	6	7
Train	1	**0.62**	0.62	0.48	0.15	0.23	0.24	0.11
	2	0.22	**0.65**	0.11	0.04	0.00	0.00	-0.01
	3	0.41	0.17	**0.63**	0.07	-0.00	0.01	-0.01
	4	0.06	0.48	-0.02	**0.69**	0.57	0.65	0.52
	5	0.24	0.13	0.02	0.64	**0.62**	0.69	0.52
	6	0.23	0.01	-0.00	0.53	0.68	**0.76**	0.31
	7	0.10	0.21	0.13	0.55	0.44	0.52	**0.77**
	Global	0.51	0.52	0.49	0.31	0.4852	0.31	0.0893
	diag. mean							**0.68**
	off-diag. mean							**0.26**

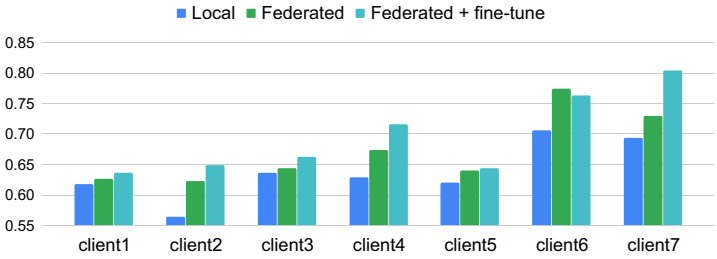

Fig. 5. Weighted linear kappa performance before and after federated learning, and after an additional round of local fine-tuning at each local site.

4 Discussion and Conclusions

Given our experimental results, we can see that federated learning (FL) in a real-world scenario can both achieve more accurate models locally as well as increase the generalizability of these models to data from other sources, such as test data from other clients. This improvement is due to the effectively larger training set made available through FL without the need to share any data directly. While we cannot directly compare to a centralized training setting due to the nature of performing FL in a real-world setting, we observed that the average performance of models is similar to values reported in the literature on centralized datasets. For example, Lehman et al. [13] reported a linear kappa value of 0.67 when applying DL for mammography breast density classification. We achieved an average performance of local models of 0.68 in the FL setting, confirming the ability of FL to achieve models comparable to models trained when the data is accumulated in a central database. However, while the generalizability is improved, it is still not comparable to the performance on local test sets. In particular, the final global model is not near any acceptable performance on any of the local test datasets. The heterogeneity in results across institutions illustrates the difficulties in training models that are generalizable. In practice, some local adaptation (fine-tuning, see Fig. 5) or at least model selection based on local validation data (see diagonal of Table 2) is needed.

In this work, we deliberately did not attempt any data harmonization methods to study the effect of different data domains. The marked differences in intensity distributions due to different mammography systems are observable in Fig. 4. Future work might explore the use of histogram equalization and other techniques [1,10] to harmonize non-IID data across different sites or investigate built-in strategies for domain adaptation within the FL framework [15]. Similarly, we did not fully address issues of data size heterogeneity and class imbalance within our FL framework. For example, client 7 had almost no category (b) samples due to their local labeling practices required by their clinical protocol. Future work could incorporate training strategies such as client-specific local training iterations, other mini-batch sampling strategies, and loss functions. We also did not attempt privacy-preservation techniques that would reduce the chance of model inversion and potential data leakage based on the trained models. Differential privacy could easily be applied to our framework, and it has been shown that it can achieve comparable results to the vanilla FL setting [14].

Despite these challenges, we were able to train mammography models in a real-world FL setting that improved the performance of locally trained models alone, illustrating the promise of FL for building clinically-applicable models and sidestepping the need for accumulating a centralized dataset.

Acknowledgements. Research reported in this publication was supported by a training grant from the National Institute of Biomedical Imaging and Bioengineering (NIBIB) of the National Institutes of Health (NIH) under award number 5T32EB1680 to K. Chang and J. B. Patel and by the National Cancer Institute (NCI) of the NIH under Award Number F30CA239407 to K. Chang. This study was supported by NIH grants U01CA154601, U24CA180927, U24CA180918, and U01CA242879, and National Science Foundation (NSF) grant NSF1622542 to J. Kalpathy-Cramer.

References

1. Baweja, C., Glocker, B., Kamnitsas, K.: Towards continual learning in medical imaging. arXiv preprint arXiv:1811.02496 (2018)
2. Boyd, N., et al.: Quantitative classification of mammographic densities and breast cancer risk: results from the Canadian national breast screening study. JNCI J. Nat. Cancer Inst. **87**(9), 670–675 (1995)
3. Boyd, N.F., et al.: Mammographic density and the risk and detection of breast cancer. N. Engl. J. Med. **356**(3), 227–236 (2007)
4. Chang, K., et al.: Distributed deep learning networks among institutions for medical imaging. J. Am. Med. Inform. Assoc. **25**(8), 945–954 (2018)
5. Chang, K., et al.: Multi-institutional assessment and crowdsourcing evaluation of deep learning for automated classification of breast density. J. Am. Coll. Radiol. (2020). https://doi.org/10.1016/j.jacr.2020.05.015
6. Dunnmon, J.A., Yi, D., Langlotz, C.P., Ré, C., Rubin, D.L., Lungren, M.P.: Assessment of convolutional neural networks for automated classification of chest radiographs. Radiology **290**(2), 537–544 (2019)
7. Ho, J.M., Jafferjee, N., Covarrubias, G.M., Ghesani, M., Handler, B.: Dense breasts: a review of reporting legislation and available supplemental screening options. AJR Am. J. Roentgenol. **203**(2), 449–456 (2014)

8. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708 (2017)
9. 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**, 1–7 (2020)
10. Karani, N., Chaitanya, K., Baumgartner, C., Konukoglu, E.: A lifelong learning approach to brain MR segmentation across scanners and protocols. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11070, pp. 476–484. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00928-1_54
11. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics **33**, 159–174 (1977)
12. Larson, D.B., Magnus, D.C., Lungren, M.P., Shah, N.H., Langlotz, C.P.: Ethics of using and sharing clinical imaging data for artificial intelligence: a proposed framework. Radiology **295**, 192536 (2020)
13. Lehman, C.D., et al.: Mammographic breast density assessment using deep learning: clinical implementation. Radiology **290**(1), 52–58 (2019)
14. Li, W., et al.: Privacy-preserving federated brain tumour segmentation. In: Suk, H.-I., Liu, M., Yan, P., Lian, C. (eds.) MLMI 2019. LNCS, vol. 11861, pp. 133–141. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32692-0_16
15. Li, X., Gu, Y., Dvornek, N., Staib, L., Ventola, P., Duncan, J.S.: Multi-site FMRI analysis using privacy-preserving federated learning and domain adaptation: Abide results. arXiv preprint arXiv:2001.05647 (2020)
16. McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: Artificial Intelligence and Statistics, pp. 1273–1282 (2017)
17. Ooms, E., et al.: Mammography: interobserver variability in breast density assessment. Breast **16**(6), 568–576 (2007)
18. Pisano, E.D., et al.: Diagnostic performance of digital versus film mammography for breast-cancer screening. N. Engl. J. Med. **353**(17), 1773–1783 (2005)
19. Razzaghi, H., Troester, M.A., Gierach, G.L., Olshan, A.F., Yankaskas, B.C., Millikan, R.C.: Mammographic density and breast cancer risk in white and African American women. Breast Cancer Res. Treat. **135**(2), 571–580 (2012)
20. Rieke, N., et al.: The future of digital health with federated learning. arXiv preprint arXiv:2003.08119 (2020)
21. 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: Crimi, A., Bakas, S., Kuijf, H., Keyvan, F., Reyes, M., van Walsum, T. (eds.) BrainLes 2018. LNCS, vol. 11383, pp. 92–104. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-11723-8_9
22. Sickles, E., d'Orsi, C., Bassett, L., Appleton, C., Berg, W., Burnside, E., et al.: ACR BI-RADS® mammography. In: ACR BI-RADS® Atlas, Breast Imaging Reporting and Data System, vol. 5, p. 2013 (2013)
23. Spayne, M.C., Gard, C.C., Skelly, J., Miglioretti, D.L., Vacek, P.M., Geller, B.M.: Reproducibility of bi-rads breast density measures among community radiologists: a prospective cohort study. Breast J. **18**(4), 326–333 (2012)
24. Sprague, B.L., et al.: Variation in mammographic breast density assessments among radiologists in clinical practice: a multicenter observational study. Ann. Intern. Med. **165**(7), 457–464 (2016)

25. Sprague, B.L., et al.: Prevalence of mammographically dense breasts in the United States. JNCI J. Nat. Cancer Inst. **106**(10), dju255 (2014)
26. Yaghjyan, L., et al.: Mammographic breast density and subsequent risk of breast cancer in postmenopausal women according to tumor characteristics. J. Nat. Cancer Inst. **103**(15), 1179–1189 (2011)