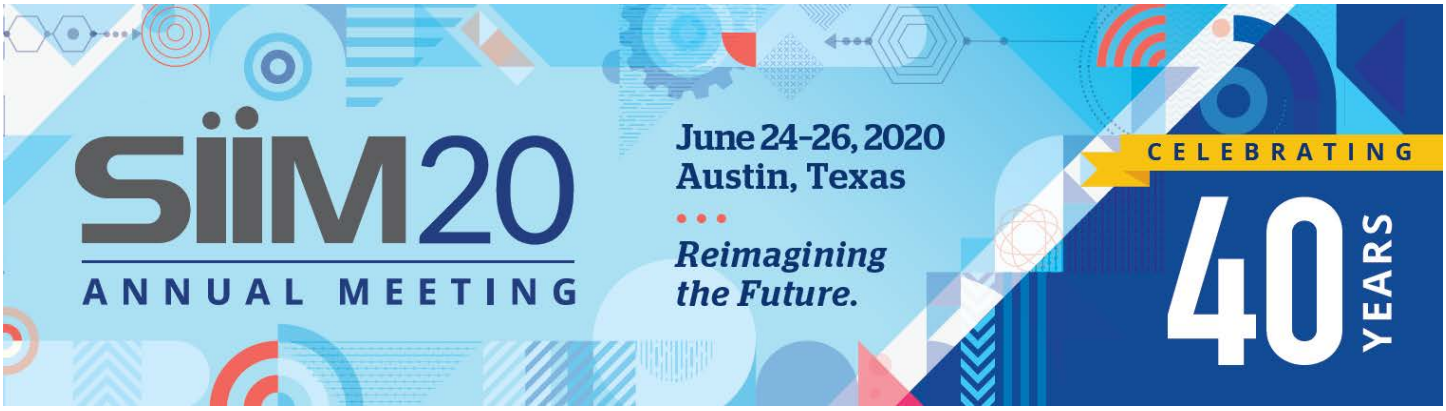

Federated Deep Learning Among Multiple Institutions for Automated Classification of Breast Density

Ken Chang, MSE, Massachusetts Institute of Technology; Praveer Singh, PhD; Wenqi Li, PhD; Holger Roth, PhD; Alvin Ihsani, PhD: Nir Neumark, MD; Bernardo C. Bizzo, MD; Yuhong Wen, PhD; Varun Buch, MD; Adam McCarthy, MS; B. Min Yun, PhD; Elshaimaa Sharaf, MBBCh; Katharina V. Hoebel, MD; Jay B. Patel; Bryan Chen; Sean Ko; Evan Leibovitz; Etta D. Pisano, MD; Laura Coombs, PhD; Daguang Xu, PhD; Keith J. Dreyer, DO, PhD; Ittai Dayan, MD; Ram C. Naidu, PhD; Jayashree Kalpathy-Cramer, PhD

Introduction

Deep learning has become the method of choice for automating medical imaging tasks. One requirement for training robust deep learning models is a large amount of diverse patient data. However, imaging data cannot always be shared due to patient privacy concerns and lack of infrastructure to store large data volumes at a centralized location. As such, there is need to collaboratively train deep neural networks on multi-institutional data without sharing patient data. One promising approach is federated learning, in which locally trained models are synchronously averaged on a central parameter server. While federated learning has been tested in simulation, it has yet been demonstrated in a real-world medical imaging setting. In this study, we evaluate federated deep learning for mammographic breast density on a large multi-institutional patient cohort.

Hypothesis

We hypothesize that deep learning models trained on a single institution will not generalize to other institutions and only a model trained on all institutions via centrally hosting data or federated learning will generalize well.

Methods

Digital screening mammograms were retrospectively obtained through Digital Mammographic Imaging Screening Trial (DMIST), which was divided three institutions based on scanner manufacturers. We also obtained data from our institution. Each examination was interpreted by a radiologist using ACR BI-RADS breast density lexicon (Category a-d). Our final cohort consisted of 32947, 12152, 63127, and 17549 images from institutions 1-4, respectively (Fig. 1A). Data from each institution was divided into training/validation/testing sets in a 7:2:1 ratio. Single institution models as well as a centrally hosted data model were trained using NVIDIA Clara Train with a ResNet50V2 neural network architecture on a single GPU for 40 epochs. For comparison, federated learning was implemented using Clara Federated and trained across four GPUs for 160 epochs with model averaging every 4 epochs. Agreement between radiologist and model predictions were assessed via linear κ coefficient across the four breast density categories in the testing set.

Results

The intensity distributions of mammograms from different institutions varied in their range as well as distribution shape (Fig. 1B). As a result, models trained at a single institution only performed well on the testing set of the same institution and had lower performance on testing sets from other institutions (Fig. 2). The centrally hosted model simulates the scenario in which institutions are capable of sharing their data to a central site. The centrally hosted model performed well across all institutions in the testing set. Similarly, federated training of models performed well across all institutions in the testing set (Fig. 3).

Conclusion

A deep learning model trained at a single institution is not a viable option for robust deployment across multiple institutions. Only centrally hosting data or federated training results in a model with relatively high generalization ability, the latter of which mitigates the need for sharing of patient data.

Statement of Impact

In this study, we show that federated learning achieves a model that is generalizable across multiple institutions. This serves as a demonstration of the viability of distributed learning in a real-world multi-institutional patient cohort.

Keywords

federated learning, distributed learning, screening mammography, breast density, BIRADS

A

	Institution 1 (n = 32947)	Institution 2 (n = 12152)
Radiologist-assessed BIRADS breast density		
a - Fatty	12.5%	10.8%
b - Scattered	41.4%	45.0%
c - Heterogeneously dense	39.3%	36.7%
d - Extremely dense	6.8%	7.5%

	Institution 3 (n = 63127)	Institution 4 (n = 17549)
Radiologist-assessed BIRADS breast density		
a - Fatty	9.1%	5.9%
b - Scattered	43.9%	47.4%
c - Heterogeneously dense	40.7%	40.7%
d - Extremely dense	6.3%	6.0%

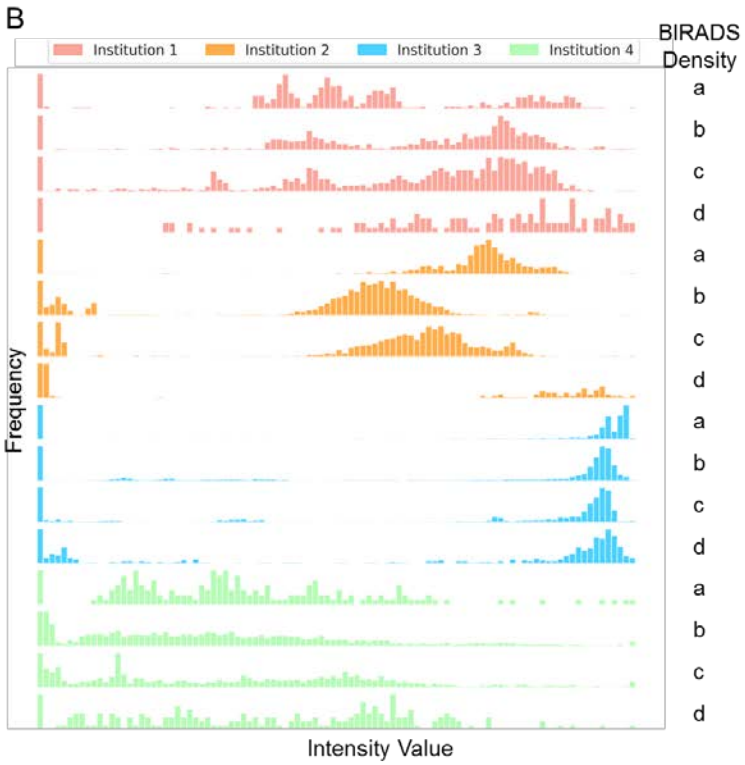

Figure 1. (A) Breast density class distribution at each institution (B) The intensity distribution of mammograms varies at each institution. Intensity distribution histogram (Frequency vs Intensity Value) of 100 randomly selected images from each institution.

Figure 2. Performance of models trained on single institutions, showing that for institution specific models, testing set performance was only high for the same institution and decreased for other institutions.

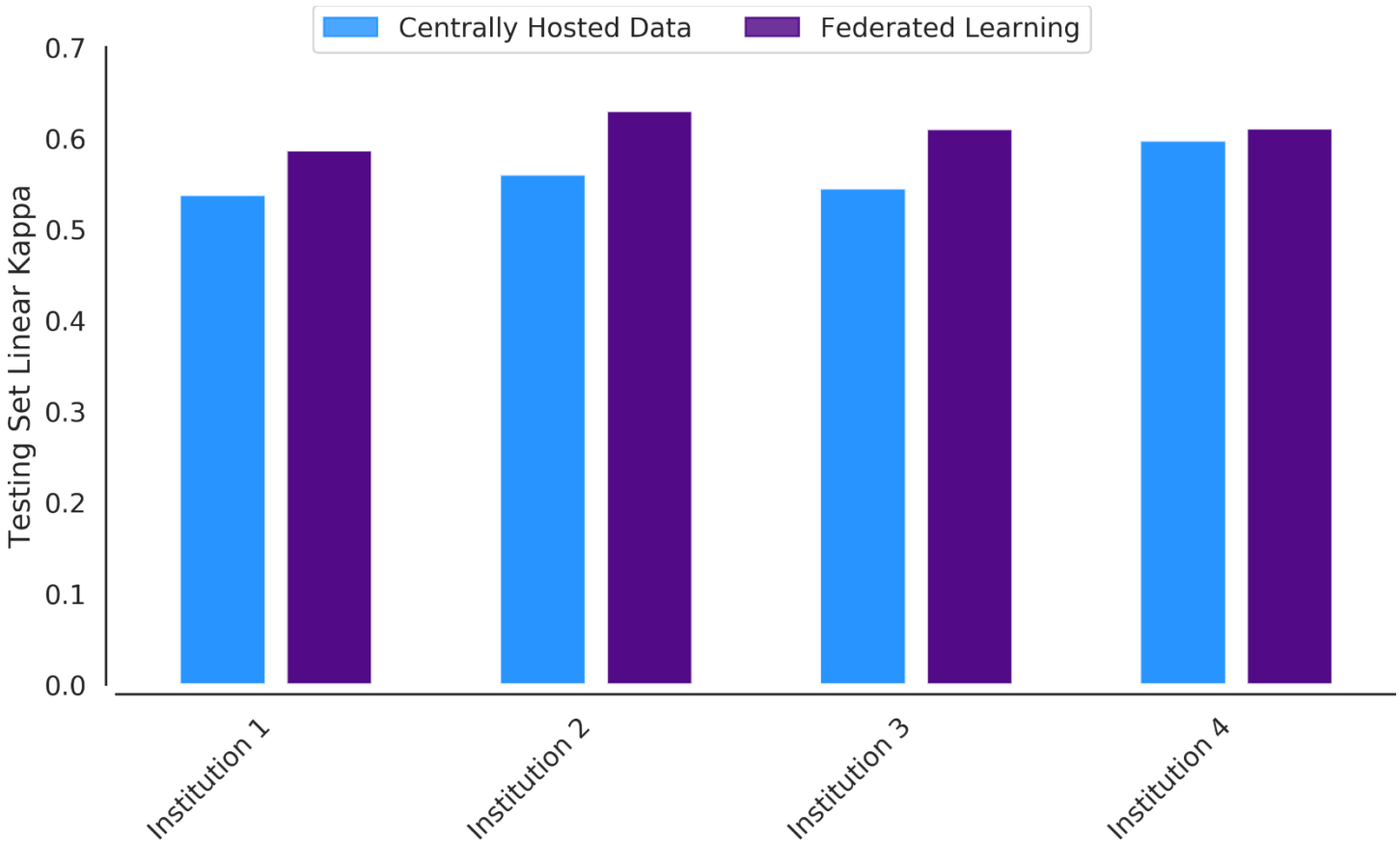

Figure 3. Bar plot of testing set performance at each institution for centrally hosted data and federated learning. Both approaches showed high performance across all institutions.