Journal of the American Medical Informatics Association, 30(1), 2023, 54–63
https://doi.org/10.1093/jamia/ocac188
Advance Access Publication Date: 10 October 2022
Research and Applications

Research and Applications

Evaluation of federated learning variations for COVID-19 diagnosis using chest radiographs from 42 US and European hospitals

Le Peng[1], Gaoxiang Luo[1], Andrew Walker[1], Zachary Zaiman[2], Emma K. Jones[3], Hemant Gupta[4], Kristopher Kersten[5], John L. Burns[6], Christopher A. Harle[7], Tanja Magoc[8], Benjamin Shickel[9,10], Scott D. Steenburg[11], Tyler Loftus[10,12], Genevieve B. Melton[3,4,13,14], Judy Wawira Gichoya[15], Ju Sun[1], and Christopher J. Tignanelli[3,13,14]

[1]Department of Computer Science and Engineering, University of Minnesota, Minneapolis, Minnesota, USA, [2]Department of Computer Science, Emory University, Atlanta, Georgia, USA, [3]Department of Surgery, University of Minnesota, Minneapolis, Minnesota, USA, [4]Fairview Health Services, Minneapolis, Minnesota, USA, [5]Nvidia Corporation, Santa Clara, California, USA, [6]The School of Medicine, Indiana University, Indianapolis, Indiana, USA, [7]Department of Health Outcomes and Biomedical Informatics, University of Florida, Gainesville, Florida, USA, [8]University of Florida College of Medicine, Gainesville, Florida, USA, [9]Department of Medicine, University of Florida, Gainesville, Florida, USA, [10]Intelligent Critical Care Center, University of Florida, Gainesville, Florida, USA, [11]Department of Radiology and Imaging Sciences, Indiana University School of Medicine, Indianapolis, Indiana, USA, [12]Department of Surgery, University of Florida, Gainesville, Florida, USA, [13]Center for Learning Health System Sciences, University of Minnesota, Minneapolis, Minnesota, USA, [14]Institute for Health Informatics, University of Minnesota, Minneapolis, Minnesota, USA, and [15]Department of Radiology, Emory University, Atlanta, Georgia, USA

Corresponding Author: Christopher J. Tignanelli, MD, Department of Surgery, University of Minnesota, 420 Delaware St. SE, Minneapolis, MN 55455, USA; ctignane@umn.edu

Received 25 April 2022; Revised 31 August 2022; Editorial Decision 24 September 2022; Accepted 7 October 2022

ABSTRACT

Objective: Federated learning (FL) allows multiple distributed data holders to collaboratively learn a shared model without data sharing. However, individual health system data are heterogeneous. "Personalized" FL variations have been developed to counter data heterogeneity, but few have been evaluated using real-world healthcare data. The purpose of this study is to investigate the performance of a single-site versus a 3-client federated model using a previously described Coronavirus Disease 19 (COVID-19) diagnostic model. Additionally, to investigate the effect of system heterogeneity, we evaluate the performance of 4 FL variations.

Materials and methods: We leverage a FL healthcare collaborative including data from 5 international healthcare systems (US and Europe) encompassing 42 hospitals. We implemented a COVID-19 computer vision diagnosis system using the Federated Averaging (FedAvg) algorithm implemented on Clara Train SDK 4.0. To study the effect of data heterogeneity, training data was pooled from 3 systems locally and federation was simulated. We compared a centralized/pooled model, versus FedAvg, and 3 personalized FL variations (FedProx, FedBN, and FedAMP).

Results: We observed comparable model performance with respect to internal validation (local model: AUROC 0.94 vs FedAvg: 0.95, $P = .5$) and improved model generalizability with the FedAvg model ($P < .05$). When investigating the effects of model heterogeneity, we observed poor performance with FedAvg on internal validation as compared to personalized FL algorithms. FedAvg did have improved generalizability compared to personalized FL algorithms. On average, FedBN had the best rank performance on internal and external validation.

Conclusion: FedAvg can significantly improve the generalization of the model compared to other personalization FL algorithms; however, at the cost of poor internal validity. Personalized FL may offer an opportunity to develop both internal and externally validated algorithms.

Key words: computer vision, federated learning, COVID-19, artificial intelligence

INTRODUCTION

Biomedical artificial intelligence (AI) models learn from data to predict future patterns. Typically, to train generalizable AI models, data from multiple institutions are pooled and a model is trained on the pooled (centralized) dataset. Such an approach suffers from multiple issues including concerns about data security, data privacy, and regulatory restrictions from Health Insurance Portability and Accountability Act (HIPAA)[1] and General Data Protection Regulation.[2] Federated learning (FL) overcomes these limitations by enabling multiple distributed data holders to collaboratively train a shared model without data sharing.

Individual health system data, however, are highly heterogeneous due to the differences in patient populations and clinical workflow. Thus, as previous patterns help predict future patterns, in the context of system heterogeneity, it is possible that AI models trained on single-institution historic data may have superior predictive capabilities locally than federated models trained with supplementation of external institutional data. This presents a unique problem that must be addressed by FL variations in healthcare. Federated Averaging (FedAvg), a well-known FL algorithm, may suffer in the presence of data heterogeneity.[3,4] Zhao et al[4] show that when the client's data are highly skewed, the accuracy of FedAvg reduces significantly, by up to ~55%. Similarly, Li et al[5] point out that FedAvg will not converge to the optimal solution when the training quantity is imbalanced among different clients. Recently, "personalized" variations of FL have been developed to counter this data heterogeneity issue, but most studies to date only consider a single type of data heterogeneity and experiment in a simulated environment.[5–9] Thus, a rigorous investigation of single site, centralized ("pooled"), FedAvg, and "personalized" FL models is needed to characterize the optimal solutions and current gaps. Additionally, an investigation is needed to characterize the generalizability of "personalized" FL variations using real-world clinical data from multiple heterogeneous hospitals. The purpose of this study was to investigate the performance of single site, centralized, and multiple FL models using a previously developed and validated computer vision diagnostic AI model using a real-world Coronavirus Disease 19 (COVID-19) chest radiograph (CXR) images from 42 United States (US) and European hospitals.

OBJECTIVE

This study characterized the problem posed by data heterogeneity using the well-known FedAvg algorithm across data from 5 heterogeneous international (US and Europe) healthcare systems encompassing 42 hospitals. Then, to identify an optimal solution for model training, we evaluated the performance of 3 personalized federated variations (FedBN [Fed batch normalization],[6] FedAMP [Fed attentive message passing],[8] and FebProx[5]) a locally trained model, FedAvg, and a centralized (data "pooled" together at a single site) model.

MATERIALS AND METHODS

Dataset collection

Model training and validation datasets

Five datasets were available for this analysis (Table 1). Three were pooled locally at the University of Minnesota and 2 were made available within a federated healthcare platform[10] which supports the FedAvg algorithm.[10] Figure 1 provides a schematic representation of the study design, available datasets, and the analysis conducted for this study.

All patients aged 18 years or older who presented to the emergency department (ED) of a participating site between February 1, 2019 and November 11, 2020 and received a CXR were included. Positive cases were defined as CXRs from patients who presented to the ED with PCR-confirmed COVID-19 (taken either 2 weeks prior to COVID-19 diagnosis or during a COVID-19-associated hospitalization). Negative controls were defined as CXRs from patients who were not PCR positive for COVID-19 and had a CXR obtained in the ED for any reason prior to the onset of COVID-19. This definition of a negative control ensures that negative controls were true negatives and not patients that had a false negative PCR or were not tested for COVID-19 due to limited availability of COVID-19 PCR testing early in the pandemic. There were no missing DICOM images for patients who had a CXR performed. To train the federated model, we used data collected from 5 sites: (1) 12-hospital M Health Fairview University of Minnesota system (MHFV), Minnesota, USA, (2) 16-hospital Indiana University (IU) system, Indiana, USA, (3) 12-hospital Emory University (EU) system, Georgia, USA, (4) University of Florida, Gainesville, Florida, USA, and (5) Valencian Region Medical ImageBank (BIMCV), Valencia, Spain (Table 1).

Data harmonization

The only data element used in model training was the CXR image itself. Each institution was responsible for obtaining data locally. Data from BIMCV and University of Florida, Gainesville were transferred directly to MHFV, whereas data from IU and EU were stored within each institution. For data harmonization, data from MHFV, BIMCV, and the University of Florida, Gainesville, were stored locally on MHFV's HIPAA-compliant cloud computing environment each within a site-specific directory. A spreadsheet was maintained that contained the path to each image as well as data labels. Data from IU and EU were stored locally within each partner institution's GPU computing environments.

Investigation of training heterogeneity impact on model performance

We hypothesized that a federated model would improve generalizability of the model, but at the result in diminished local performance due to increased training data heterogeneity. To investigate this, we proposed a 2-phase model to mitigate the effect of data inconsistency and data noise. In phase 1, we preprocessed the multisource (heterogeneous) data using a shared processing pipeline including lung segmentation, outlier detection, and image normal-

Table 1. Distribution of the data on each datasets

	COVID-19 status	N	Age in years, median (IQR: 25th–75th)	Male %	Racial distribution (if available)
MHFV	Positive	3997	62 (50–73)	57.4%	31.7% Black 20.4% Other 47.9% White
	Negative	41516	60 (44–72)	47.1%	9.7% Black 9.1% Other 81.2% White
IU	Positive	8231	62 (50–74)	57.3%	Not Available
	Negative	7668	71 (43–71)	48.8%	Not Available
EU	Positive	8602	61 (50–73)	51.5%	68.5% Black 10.8% Other 20.7% White
	Negative	11651	60 (45–72)	48.4%	50.4% Black 6.7% Other 42.9% White
BIMCV	Positive	2261	Not Available	Not Available	Not Available
	Negative	1561	Not Available	Not Available	Not Available
UF	Positive	1009	58 (39–69)	45.7%	41.8% Black 9.9% Other 48.3% White
	Negative	1460	59 (43–70)	52.7%	27.4% Black 5.8% Other 66.8% White

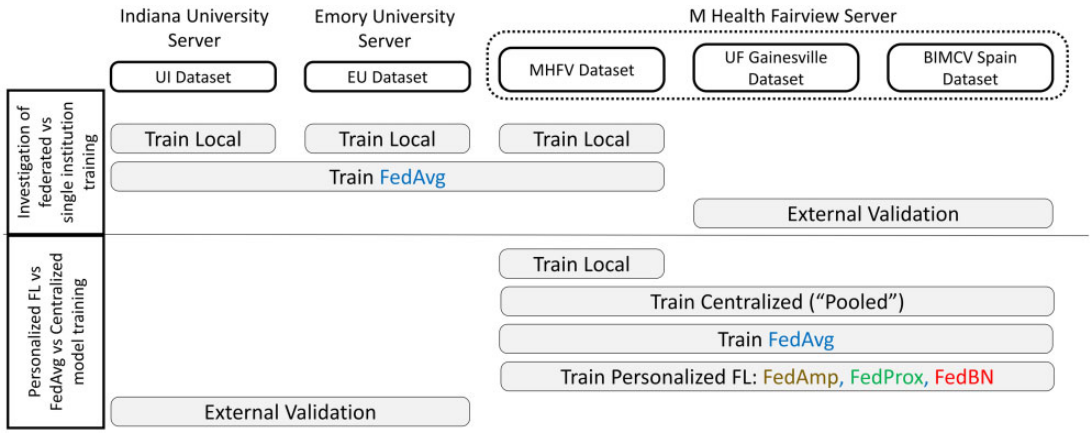

Figure 1. Schematic representation of the available datasets and the analysis conducted for this study. IU: Indiana University; EU: Emory University; MHFV: M Health Fairview; UF: University of Florida; BIMCV: Valencian Region Medical ImageBank.

ization; as we have developed previously (see Supplementary Methods).[11] This step helps to ensure that the data fed into the final classifier is qualitatively consistent across all sites, which is critical for successful federation. Additionally, these steps minimize the inclusion of data outside of the lung window thus minimizing the risk of "AI shortcuts".[12]

In phase 2, we used Clara Train SDK 4.0 and NVFlare, and built the FL model using MONAI (Medical Open Network for AI)[13] Densenet121 pre-trained on ImageNet.[14]

We used FedAvg, a standard FL algorithm widely used in many applications and areas. The main idea of FedAvg is to aggregate models to achieve a comparable performance to centralized learning without the need for data sharing. It aims to solve the problem as shown in Equation (1).

$$\min_{w} \sum_{i=1}^{K} \frac{n_i}{n} F_k(\boldsymbol{w}) \text{ where } F_k = \frac{1}{n_k} \sum_{i=1}^{n_k} L_w(X_i,\ Y_i), \qquad (1)$$

where n_k is the number of training samples in client k, and n is the sum of the training sample over all the clients. Note that only weights w are shared among different clients. There is no data sharing during the training.

The pipeline of FedAvg is shown in Figure 2, and presents a hub-and-spoke topology with the hub representing a central aggregator and spoke connecting to each client. Overall, FL consists of 3 steps: (1) local training: each client site can independently train the model locally for several epochs, (2) model aggregation: each client will send the model trained in the last step to a central server, where the model will be aggregated (weighted average), and (3) local update: the aggregated model at the central server will be sent back to the

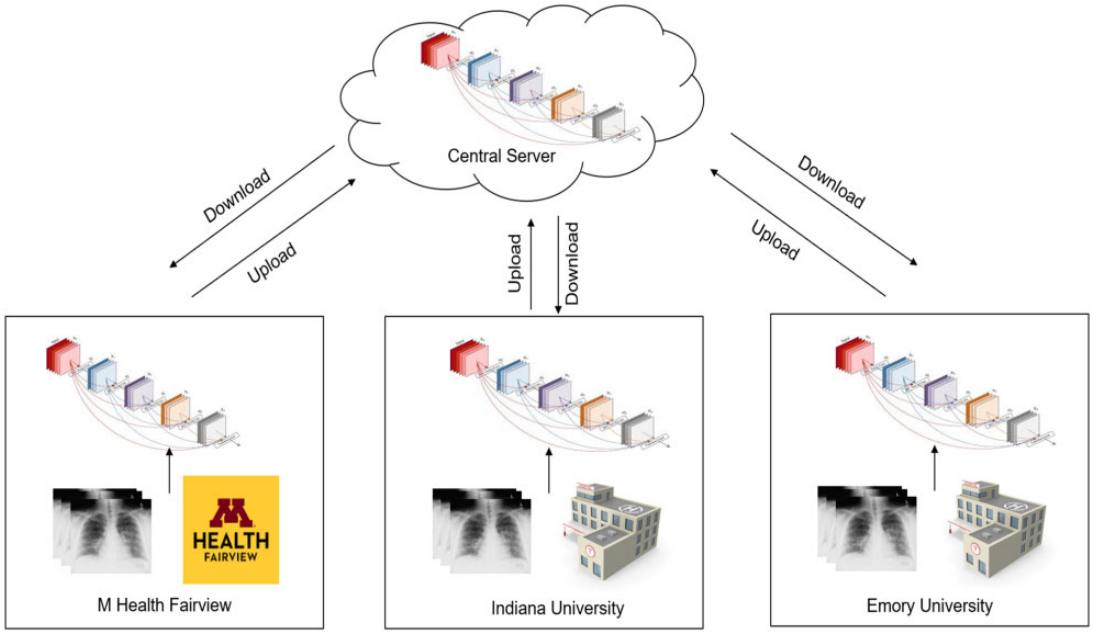

Figure 2. Overview of federated learning in Nvidia Clara Train.

clients for update in the next round. The model is converged by iterating the 3 steps through a finite round. We summarize the FedAvg in Algorithm 1.

Investigation of single institution model training verses federated learning

For this experiment, we leveraged the Nvidia Clara Train Platform which already contained data from 3 sites (MHFV, IU, and EU). For each site, we randomly split the data into training and test sets with a ratio of 8:2. Furthermore, we separated 20% of the training data into a validation set to avoid overfitting. For external model evaluation, we held data from University of Florida, Gainesville (UF, Florida, USA) and BIMCV[15] (Valencia, Spain) as external validation sets. Details on the label distribution can be found in Table 1.

Investigation of personalized federated variation performance

In order to investigate the performance of personalized federated variations, all training data was required to be held locally in the MHFV server for 2 reasons: (1) to facilitate the evaluation of a pooled model and (2) due to their novelty, personalized federated variations are not yet supported by Clara train. To facilitate this experiment, our training data consisted of data from MHFV, UF, and BIMCV, and we reserved EU and IU dataset for external validation.

For comparison, we selected 3 representative personalized FL algorithms, including FedProx, FedBN, and FedAMP, and compared their performance with FedAvg and centralized training. **FedProx** was initialized by Li et al[5] to address the heterogeneous local update issues in FedAvg. By adding a proximal term to the objective of the local update, it can suppress the impact of variable local updates. **FedBN** is proposed by Li et al[6] to solve the covariate shift over the clients' data. In its design, all weights are shared and aggregated except from batch normalization layers. **FedAMP** is an attention-based approach that automatically learns the aggregation weight for each client. Their design follows the personalized federation manner where all clients have a dedicated global model on the server and the goal is to train a personalized model such that the local performance is maximized. The comparison of these methods can be found in Algorithms 1 and 2.

Training details

We normalized the CXR images using the mean and standard deviation of ImageNet (see Supplementary Methods). Before being fed into the neural network, mild random rotation was used for data augmentation. For each round of federated training, we used cross-entropy loss to measure the prediction error (the difference between model prediction and ground truth label). To mitigate the adverse effect of class imbalance, we defined the training objective as minimizing the maximum of averaged loss over the positive and negative cases. We set a maximum training epoch of 100 and optimized our learning objective using Adam optimizer,[16] with an initial learning rate of 0.0001 and set other hyper-parameters as default in PyTorch 1.5.0.[17] A learning-rate scheduler was applied to achieve faster convergence; when the area under the precision recall curve (AUPRC) on the validation set stopped improving, the learning rate was decreased by a factor of 0.5. As different FL clients have installed GPUs of different processing capacity, we specified a batch size of 64, 128, and 256 for sites 2, 3, and 1, respectively. The learning rate and training epoch used above is determined by k-fold cross-validation. To be more specific, we defined a set for each hyper-parameter: training epoch {50, 100, 200} and learning rate {0.001, 0.0001, 0.00001, 0.000001}. Then, we divided the data into 5-fold, with each having the same number of images. For each iteration, we used 1 of the 5 folds as the validation set while remaining folds were used for training. The optimal combination was chosen if the highest average AUPRC on 5 iterations was observed on the validation set. Due to the complexity of tuning hyper-parameters in a FL setting, all the tuning process was performed locally on MHFV.

All analysis was completed on a system consisting of Intel(R) Xeon(R) Processor E5-2690 v4 + 4x Tesla V100 PCIe (MHFV),

Algorithm 1: Federated learning algorithms (FedAvg/FedBN/FedProx/FedAMP)

Notation: X_i indicate data from client i, K is the total number of client, T is maximum training round, n is the sum of n_1 to n_k, σ is the hyper-parameter in FedAMP

Initialize server model weights $\mathbf{w}(1)/w_i(1) \forall\ i = 1,\ 2, \ldots, K$

Initialize client model weights $w_i\ \forall\ i = 1,\ 2,\ \ldots, K$

For each round t = 1, 2, … T do

Send server model weight $w(t)$ to each client/send $w(t)_k$ to client k

For each client $k\ =\ 1,\ 2, \ldots,\ K$ do

Client k perform LocalUpdate$(X_k,\ Y_k,\ w_k) \leftarrow$ Algorithm 2

$$\gamma_k\ =\ \frac{n_k}{n}$$

End for

For each client $j\ =\ 1,\ 2, \ldots,\ K$ do

$$\gamma_{ij}\ =\ \beta_k\left(1 - e^{\frac{\|\ w_i - w_j\|^2}{\sigma}}\right) \text{ where } i \neq j$$

End for

For each global model i

$$w(t+1)_i = \sum_{i=1}^{k} \gamma_{ik} w_k$$

End for

For each layer j in model k if is not BatchNorm then

$$w(t+1) = \sum_{j=1}^{k} \gamma_k w_{kj}$$

End for

End for

Algorithm 2: Local model training using mini batch stochastic gradient descent (LocalUpdate) (FedAvg/FedBN/FedProx/FedAMP)

Notation: R is the local update round, B is the number of batches, f_{w_r} is the neural network parameterized by w_r, η is the learning rate, μ is the hyper-parameter in FedProx, λ and α_k are the hyper-parameters in FedAMP

For each round $r\ =\ 1,\ 2,\ \ldots,\ R$ do

Randomly shuffle X_k and create B batches $((X_1,\ Y_1),\ (X_2,\ \ Y_2), \ldots,\ (X_B,\ \ Y_B))$

$$L_{w_r}\ =\ BCELoss(f_{w_r}(X_b),\ Y_b)\ +\ \frac{\mu}{2}||w_k - w_k(t)||^2\ +\ \frac{\lambda}{2\alpha_k}||w_k - w_k(t)||^2$$

For each mini batch $b\ =\ 1,\ 2, \ldots,\ B$ do

$$w_{r+1}\ =\ w_r\ -\ \eta \nabla L_{w_r}(X_b,\ Y_b)$$

End for

Table 2. Internal and external validation of federated model

		N	AUROC	AUPRC	95% CI	Precision	Recall	F1 score
Internal	MHFV	9102	0.951	0.838	0.940–0.963	0.616	0.840	0.711
	IU	3179	0.871	0.886	0.857–0.885	0.828	0.748	0.786
	EU	4051	0.832	0.801	0.813–0.851	0.681	0.784	0.729
External	BIMCV	3822	0.601	0.611	0.585–0.617	0.616	0.471	0.533
	UF	2469	0.713	0.651	0.692–0.734	0.629	0.592	0.610

Table 3. Performance comparison between single institution model (SIM) and federated learning model (FLM)

	AUROC			Sensitivity			Specificity		
	SIM	**FLM**	***P* value**	**SIM**	**FLM**	***P* value**	**SIM**	**FLM**	***P* value**
MHFV	0.944	0.951	.492	0.870	0.840	.020	0.939	0.950	<.05
BIMCV	0.557	0.601	<.05	0.301	0.471	<.05	0.833	0.730	<.05
UF	0.667	0.713	<.05	0.548	0.592	<.05	0.721	0.759	<.05

Note: We use Delong's test to compare the difference of AUROC and McNemar's test to compare specificity and sensitivity.

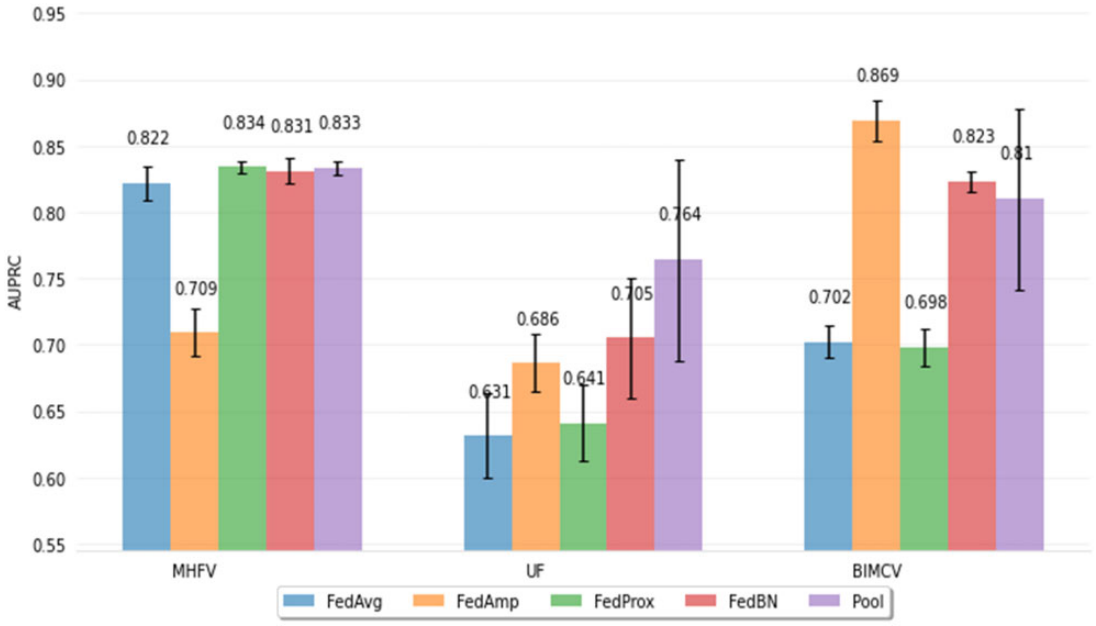

Figure 3. Performance of 4 federated learning algorithms on internal validation dataset measured by AUPRC. AUPRC: area under the precision recall curve; FL: federated learning; MHFV: M Health Fairview; UF: University of Florida, Gainesville; Pool: Centralized ("Pooled") model performance.

Intel(R) Xeon(R) Gold 6130 + NVIDIA V100X-16Q (IU), and Intel(R) Core(TM) i9-9940X + 2x GeForce RTX 2080Ti (EU). Python 3.8.5 and libraries Numpy 1.19.4, Pandas 1.0.0, Scikit-image 0.15.0, and Pydicom 1.4.2 were used for image conversion; torch 1.5.0, torchvision 0.8.0a0, and monai 0.5.3 were used for neural network training. The FL framework was built upon Clara train v4.0 and Docker 20.10.11 + azure-3. If not specifically mentioned, all settings were consistently aligned across 3 clients.

Statistical analysis

The sample size needed for adequate power varies with disease prevalence.[18] Our previous studies identified that the real-world prevalence of COVID-19 for ED patients who receive a CXR for any reason is 4.8%.[11] Assuming a 5% prevalence race, 4860 predictions would be needed for investigation with over 80% power. We measured the model performance using area under the receiver operating characteristic (AUROC), where the ROC curve is created by plotting the true positive rate against the false positive rate at varied thresholds. We used Youden index (J),[19] a common summary measure of the ROC curve, to determine the optimal threshold, from which precision, recall and F1 score were calculated. We used Delong's test[20] and Mcnemar's test[21] to evaluate the statistical significance of the difference between the single institute model and FL model on the reported metrics. For the 2 tests, a *P* value less than .05 was considered to indicate statistical significance.

The Checklist for Artificial Intelligence in Medical Imaging was used for reporting this study (see Supplementary files). This study was approved by the University of Minnesota Institutional Review Board (IRB) STUDY00014526. This study was approved by the University of Florida IRB Study 202101857. This study was approved by the Indiana University IRB Study 2010169012. This study was approved by the Emory University IRB Study STUDY00000506 ML-COVID19.

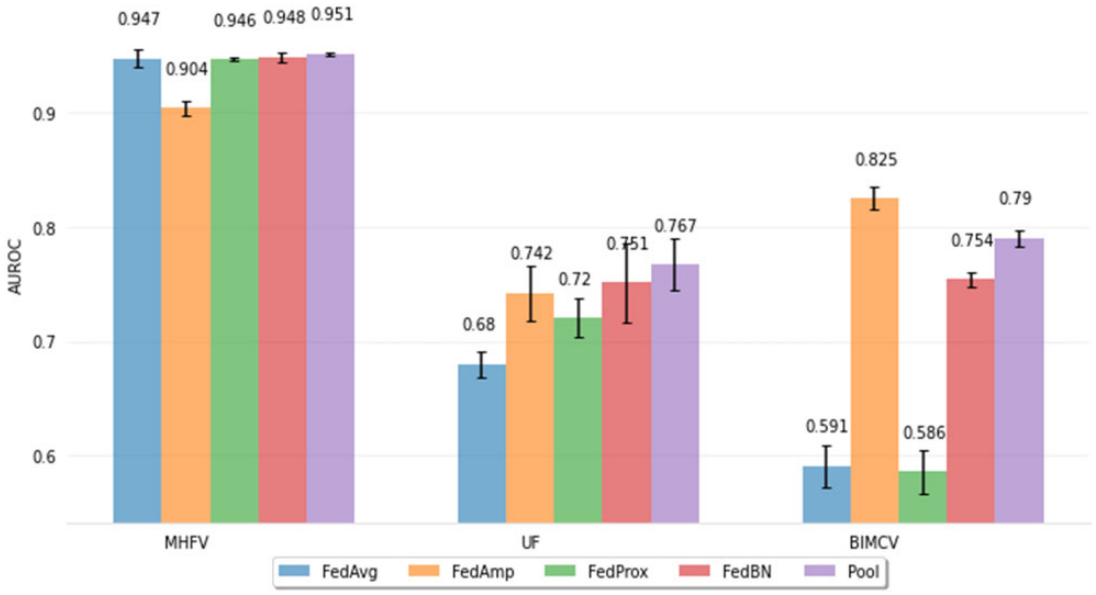

Figure 4. Performance of 4 federated learning algorithms on internal validation dataset measured by AUROC. AUROC: area under the receiver operating characteristic curve; FL: federated learning; MHFV: M Health Fairview; UF: University of Florida, Gainesville; Pool: Centralized ("Pooled") model performance.

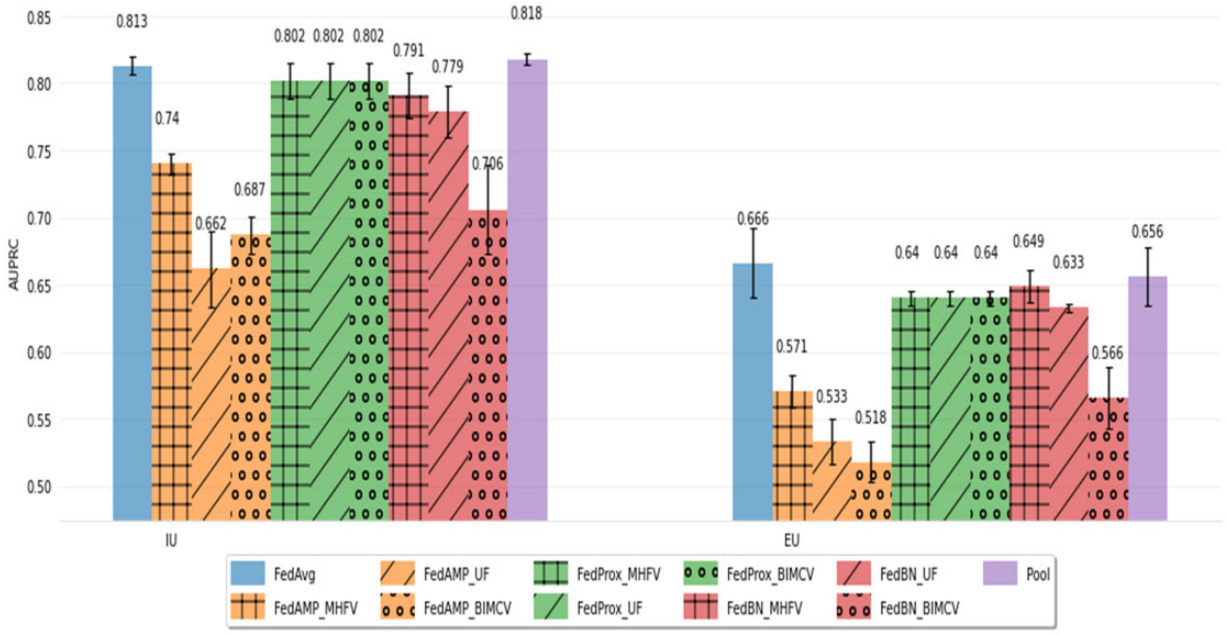

Figure 5. Performance of 4 federated learning algorithms on external validation dataset measured by AUPRC. AUPRC: area under the precision recall curve; FL: federated learning; IU: Indiana University; EU: Emory University; MHFV: M Health Fairview; UF: University of Florida, Gainesville; Pool: Centralized ("Pooled") model performance.

RESULTS

Investigation of federated training versus single institution training

The model trained with single-site MHFV data had an internal AUROC of 0.94 (95% confidence interval [CI], 0.93–0.96) and AUPRC of 0.85 and BIMCV (Spain) external validation of AUROC: 0.56 (95% CI, 0.54–0.57) and AUPRC: 0.57 and University of Florida external validation of AUROC: 0.67 (95% CI, 0.65–0.69) and AUPRC: 0.6 (Supplementary Table S1). To investigate if the FL model had improved performance versus the single institution model (SIM), a Delong's test was performed. The federated learning FedAvg model (FLM) trained using data from all 3 sites had similar performance as the MHFV SIM within MHFV (AUROC SIM: 0.94 vs AUROC FLM: 0.95, $P = .5$) (Tables 2 and 3). However, the federated model was associated with significant improvement in

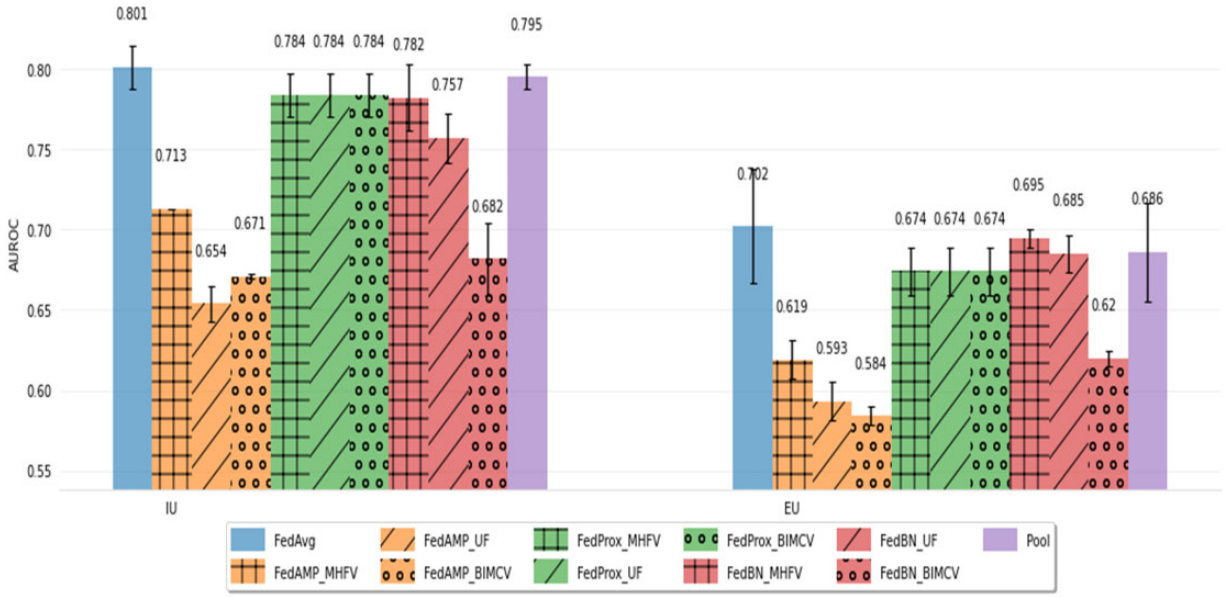

Figure 6. Performance of 4 federated learning algorithms on external validation dataset measured by AUROC. AUROC: area under the receiver operating characteristic curve; FL: federated learning; IU: Indiana University; EU: Emory University; MHFV: M Health Fairview; UF: University of Florida, Gainesville; Pool: Centralized ("Pooled") model performance.

Table 4. DeLong's test evaluating performance of pooled, FedAvg, and personalized FL variations

	IU		EU	
	AUROC	*P* value	AUROC	*P* value
FedAvg	0.801 ± 0.013	.408	0.702 ± 0.036	.213
FedAmp	0.713 ± 0.001	<.05	0.619 ± 0.012	<.05
FedProx	0.784 ± 0.013	.184	0.674 ± 0.015	.0536
FedBN	0.782 ± 0.021	.135	0.695 ± 0.006	<.05
Centralized/Pool	0.795 ± 0.008	reference	0.686 ± 0.031	reference

Note: *P* values are obtained using Delong's test to compare with the AUC of the centralized/pooled model.

external validation (BIMCV [Spain] AUROC FLM: 0.60 vs SIM: 0.56 and UF Gainesville AUROC FLM: 0.71 vs SIM: 0.67, $P < .05$). To investigate if the federated model was associated with improved sensitivity and specificity, a McNemar's test was performed. The federated model had significantly lower sensitivity (0.84 vs 0.87, $P = .02$) than the SIM; however, significantly higher specificity (0.95 vs 0.94, $P = .02$). On external validation, the federated model using FedAvg was associated with significantly improved AUROC (Table 3).

Personalized federated learning versus FedAvg and pooled model training

To evaluate the performance of personalized federated variations (as compared to pooled and FedAvg training), we conducted a federation using the original 2 external validation datasets and the MHFV dataset. As we did not have access to the IU and EU datasets, they were used as external validation datasets. On internal validation, pooled data training and FedBN provided more consistent results. FedAvg had superior performance on internal validation for the MH FV datasets which made up 87.9% of the training data; however, poor performance on UF and BIMCV datasets (Figures 3 and 4). Despite this skew in data, though FedBN maintained comparable training with pooled training. Of note, we identified significantly improved performance in BIMCV with FedAMP. We hypothesize this is because FedAMP is invulnerable to quantity skew as the aggregation weights are adaptive learned during training (Figures 3 and 4).

On external validation, we noted FedAvg performed comparable (Delong $P = .4$ [IU dataset] and $P = .2$ [EU dataset] with centralized training). This supports that FedAvg trains a generalizable model comparable to centralized training. In both cases, FedAMP significantly underperformed centralized training (Figures 5 and 6). Of the personalized FL variations, again FedBN had comparable (Delong $P = .1$ [IU]) or superior (Delong $P = .003$ [EU dataset]) performance as compared to centralized model training (Table 4, Figures 5 and 6).

DISCUSSION

FL is a novel approach originally developed with the intention to preserve data privacy while facilitating data availability for generalizable model training. To date, few studies have investigated the role of FL in the healthcare setting.[22] In this study, we characterized the problem posed by data heterogeneity using the FedAvg algorithm and evaluated the performance of 3 personalized federated variations (FedBN, FedAMP, and FedProx), a locally trained model, FedAvg, and a centralized model for a diagnostic model for COVID-19. This study identified that FL is feasible and improves model generalizability. Of the federated variations, FedAvg significantly improves external generalizability of models compared to SIMs; however, suffers from limitations on internal validation. While statistically significant improvements in external generalizability were observed, it is critical to point out that performance was only modestly improved (ie, UF external validation improved from AUROC 0.667 to 0.713). In both models, SIM and FLM, a large performance drop was noted between internal and external validation. We believe the drop in performance comes from the innate distribution shift

between the internal and external datasets. As shown in Table 1, the distributions of sex, race, and age in the UF dataset were different from MHFV and EU. Similarly, the BIMCV dataset, collected from Europe, has a distribution different from all other datasets in our study. These disparities were associated with differences in model performance. For example, in Table 2, the model was trained on data from MHFV, EU, and IU, all US centers, which may explain why the model performed better on UF data than on BIMCV data. Additionally, when the model was trained using data from MHFV, UF, and BIMCV (Table 4), we see highest performance in IU. This is unsurprising as FedAvg tends to perform optimally in sites that contribute the most data to model training. MHFV contributed the most data and IU has a similar patient population as MHFV (Table 1).

When analyzing 4 different FL algorithms on real-world medical image datasets with data heterogeneity, we found that FedAvg has a strong bias in favor of clients with large data quantities. MHFV, which contributes 87.9% training data, received the highest AUROC and AUPRC score among all clients. However, on internal validation (Figure 4) at lesser contributors (UF and BIMCV), FedAvg significantly underperformed centralized training. This presents a serious limitation to the widespread use of FedAvg. As pointed out by prior work,[23] the heterogeneity in local dataset size affects the number of local updates that the client will perform, which ultimately causes objective inconsistency and leads to a biased solution. Other types of heterogeneity, such as system heterogeneity[24] and data distribution heterogeneity,[25] can also bring bias to the federation.[26]

Some personalized FL algorithms claim to address the data distribution drift problem[5,8]; however, in this study, they did not generalize well on the real-world medical image data, eg, FedAMP and FedProx. We observed consistent good performance with FedBN on all evaluations which even outperformed centralized training in 1 of the 2 external validation datasets. As compared with centralized training, FedBN had comparable performance on internal validation (Figure 4), and the FedBN_MHFV model had comparable or significantly superior performance on external validation (Table 4). Our result suggests that when the goal is to maximize the generalization of the model, then FedAvg or FedBN is preferred whereas when local performance matters more, personalized FL such as FedBN should be prioritized.

One potential solution worth investigating could be the inclusion of an outer layer within the neural network which can identify the optimal FL for deployment given demographic and clinical similarities between the original training datasets and the validation dataset. In this scenario, a model would be trained with multiple FL variations (FedAvg, FedBN, etc.) with the inclusion of details regarding dataset demographic and clinical distribution. Hospital systems with similar demographic and clinical distributions may have superior performance when validated using personalized FL algorithms trained on similar health systems.

Limitations

This study suffers from the following limitations: due to data privacy limitations, we were unable to pool the training data across all institutions together and train a centralized model and compare its performance with the federated model. Additionally, the Nvidia Clara platform only supports the FedAvg algorithm, thus any federated training using personalized variations were done using simulation. The current federation is small in scale and thus running in a well-controlled environment. Previous studies by our group[11] investigated whether temporal changes (ie, change in patient population or radiographic features with different COVID-19 variants) resulted in COVID-19 diagnostic model bias and performance drift, thus this was not further investigated in this study, as the purpose of this study was to investigate institutional data heterogeneity and model federation.

CONCLUSIONS

Our results suggest that FedAvg can significantly improve the generalization of the model compared to other personalization FL algorithms; however, at the cost of poor internal validity. Personalized FL such as FedBN may offer an opportunity to develop both internal and externally validated algorithms. Future research should develop network layers that can characterize dataset distribution and identify the optimal FL algorithms based on dataset distribution to maximize returns.

FUNDING

This work was supported by the Agency for Healthcare Research and Quality (AHRQ) and Patient-Centered Outcomes Research Institute (PCORI), grant K12HS026379 (CJT) and the National Institutes of Health's National Center for Advancing Translational Sciences, grants KL2TR002492 (CJT) and UL1TR002494 (EKJ). The University of Minnesota Office of the Vice President of Research (OVPR) COVID-19 Impact Grant (JS and CJT). The National Institute of Biomedical Imaging and Bioengineering (NIBIB) MIDRC grants 75N92020C00008 and 75N92020C00021 and the US National Science Foundation #1928481 from the Division of Electrical, Communication & Cyber Systems (JWG).

AUTHOR CONTRIBUTIONS

LP, GL, AW, JLB, TL, GBM, JWG, and CJT—study design, data collection, data analysis, data interpretation, writing, and critical revision; ZZ, CAH, BS, and SDS—study design, data collection, data interpretation, writing, and critical revision; EKJ, HG, and TM—study design, data interpretation, writing, and critical revision; KK—study design, data interpretation, and critical revision; JS—study design, data analysis, data interpretation, writing, and critical revision.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

ACKNOWLEDGMENTS

We would like to acknowledge the Minnesota Supercomputing Institute, Fairview Health Services, as well as the University of Minnesota Center for Learning Health System Sciences—a partnership between the University of Minnesota Medical School and the School of Public Health—and its Program for Clinical Artificial Intelligence.

CONFLICT OF INTEREST STATEMENT

None declared.

DATA AVAILABILITY

The data underlying this article cannot be shared publicly due the inclusion of protected health information in the patient data.

REFERENCES

1. Health Insurance Portability and Accountability Act of 1996. https://www.cdc.gov/phlp/publications/topic/hipaa.html. Accessed April 4, 2022.
2. General Data Protection Regulation. https://gdpr-info.eu/. Accessed April 2, 2022.
3. Li X, Huang K, Yang W, Wang S, Zhang Z. On the convergence of FedAvg on non-IID data. 2019: arXiv:1907.02189. https://ui.adsabs.harvard.edu/abs/2019arXiv190702189L. Accessed July 1, 2019.
4. Zhao Y, Li M, Lai L, Suda N, Civin D, Chandra V. Federated learning with non-IID data. 2018: arXiv:1806.00582. https://ui.adsabs.harvard.edu/abs/2018arXiv180600582Z. Accessed June 1, 2018.
5. Li T, Sahu AK, Zaheer M, Sanjabi M, Talwalkar A, Smith V. Federated optimization in heterogeneous networks. 2018: arXiv:1812.06127. https://ui.adsabs.harvard.edu/abs/2018arXiv181206127L. Accessed December 1, 2018.
6. Li X, Jiang M, Zhang X, Kamp M, Dou Q. FedBN: Federated learning on non-IID features via local batch normalization. 2021: arXiv:2102.07623. https://ui.adsabs.harvard.edu/abs/2021arXiv210207623L. Accessed February 1, 2021.
7. Wang J, Liu Q, Liang H, Joshi G, Poor HV. Tackling the objective inconsistency problem in heterogeneous federated optimization. 2020: arXiv:2007.07481. https://ui.adsabs.harvard.edu/abs/2020arXiv200707481W. Accessed July 1, 2020.
8. Huang Y, Chu L, Zhou Z, *et al*. Personalized cross-silo federated learning on non-IID data. 2020: arXiv:2007.03797. https://ui.adsabs.harvard.edu/abs/2020arXiv200703797H. Accessed July 1, 2020.
9. Fallah A, Mokhtari A, Ozdaglar A. Personalized federated learning: a meta-learning approach. 2020:arXiv:2002.07948. https://ui.adsabs.harvard.edu/abs/2020arXiv200207948F. Accessed February 1, 2020.
10. NVIDIA Clara: An Application Framework Optimized for Healthcare and Life Sciences Developers. https://developer.nvidia.com/clara. Accessed April 11, 2022.
11. Sun J, Peng L, Li T, *et al*. Performance of a chest radiograph ai diagnostic tool for COVID-19: a prospective observational study. *Radiol Artif Intell* 2022; 4 (4): e210217.
12. DeGrave AJ, Janizek JD, Lee SI. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nat Mach Intell* 2021; 3 (7): 610–9.
13. MONAI (Medical Open Network for AI). https://monai.io/start.html. Accessed October 10, 2021.
14. ImageNet. https://www.image-net.org/index.php. Accessed October 10, 2021.
15. BIMCV-COVID19. https://bimcv.cipf.es/bimcv-projects/bimcv-covid19/. Accessed October 10, 2021.
16. Kingma DP, Ba J. Adam: a method for stochastic optimization. 2014: arXiv:1412.6980. https://ui.adsabs.harvard.edu/abs/2014arXiv1412.6980K. Accessed December 1, 2014.
17. PyTorch: An Imperative Style, High-Performance Deep Learning Library. https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html. Accessed on April 14, 2022.
18. Bujang MA, Adnan TH. Requirements for minimum sample size for sensitivity and specificity analysis. *J Clin Diagn Res* 2016; 10 (10): YE01–6.
19. Youden WJ. Index for rating diagnostic tests. *Cancer* 1950; 3 (1): 32–5.
20. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988; 44 (3): 837–45.
21. Mc NQ. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* 1947; 12 (2): 153–7.
22. Dayan I, Roth HR, Zhong A, *et al*. Federated learning for predicting clinical outcomes in patients with COVID-19. *Nat Med* 2021; 27 (10): 1735–43.
23. Tackling the Objective Inconsistency Problem in Heterogeneous Federated Optimization. https://proceedings.neurips.cc/paper/2020/hash/564127c03caab942e503ee6f810f54fd-Abstract.html. Accessed April 1, 2022.
24. Bonawitz K, Eichner H, Grieskamp W, *et al*. Towards federated learning at scale: system design. 2019: arXiv:1902.01046. https://ui.adsabs.harvard.edu/abs/2019arXiv190201046B. Accessed February 1, 2019.
25. Federated Multi-Task Learning. https://proceedings.neurips.cc/paper/2017/hash/6211080fa89981f66b1a0c9d55c61d0f-Abstract.html. Accessed April 2, 2022.
26. Li T, Sahu AK, Talwalkar A, Smith V. Federated learning: challenges, methods, and future directions. *IEEE Signal Process Mag* 2020; 37 (3): 50–60.