Comparison of federated learning models for detection of COVID-19

Erfan Darzidehkalani[1], Nanna M. Sijtsema[1], P.M.A van Ooijen[1]
[1]*Machine learning lab, Data Science Center in Health (DASH)*
University of Groningen, Hanzeplein 1, Groningen, The Netherlands
e.darzidehkalani@rug.nl

***Abstract*—Deep learning is effective in diagnosing COVID-19 and requires a large amount of data to be effectively trained. Due to data and privacy regulations, hospitals generally have no access to data from other hospitals. Federated learning has been used to solve this problem, where it utilizes a distributed setting to train models in hospitals in a privacy-preserving manner. Deploying federated learning is not always feasible as it requires high computation and network communication resources. This paper evaluates five federated learning algorithms' performance and resource efficiency for Covid-19 detection. A decentralized setting with CNN networks is set up, and the performance of federated learning algorithms is compared with a centralized environment. We examined the algorithms with varying numbers of participants, federated rounds, and selection algorithms.**

Our results show that cyclic weight transfer can have better overall performance, and results are better with fewer participating hospitals. Our results demonstrate good performance for detecting COVID-19 patients and might be useful in deploying federated learning algorithms for covid-19 detection and medical image analysis in general.

***Index Terms*—federated learning, medical image analysis, COVID-19, privacy preserving machine learning**

I. Introduction

Coronaviruses are a family of viruses that cause respiratory and intestinal illnesses in humans and animals. The best known variants are those responsible for the COVID-19, SARS and MERS epidemics. is known as viral pneumonia, which can be grouped into COVID-19, SARS, and MERS. Research has shown the effectiveness of chest imaging in diagnosing COVID-19 infected people. Chest imaging can be used as a powerful tool to diagnose COVID-19. Deep learning methods, such as Convolutional neural networks (CNN), can help radiologists diagnose COVID-19 with severe symptoms for various image analysis tasks [1]. For Covid-19 deep learning, models have shown great promise in spotting infected areas in CT-Scans and X-ray images.

The training of deep learning models requires sufficient and diverse medical datasets gathered from multiple data holders. And most of the existing solutions rely on a central entity in charge of collecting data from different hospitals. However, medical images may contain confidential and sensitive information about patients that often cannot be shared outside the institutions of their origin. One potential solution to this problem is federated deep learning. Federated learning aims to decentralize the whole process of training by leaving the data in the sites. In FL, the algorithm training is performed in a decentralized manner by different nodes, or clients, that use local data. In this scenario, each decentralized node trains an individual model using its data and shares the model parameters (instead of the data) with the rest.

Federated learning can differ from centralized data sharing in a number of ways. While both approaches aim to optimize their learning objective, federated learning algorithms have to account for the fact that communication with edge devices takes place over unreliable networks with very limited upload speeds. So unlike the centralized setting in which computation is generally a bottleneck, in federated learning communication might be the bottleneck.

In this paper, we developed a framework that enables collaboration between hospitals and uses multiple data sources to detect COVID-19 infection using federated learning The decentralized way of distributing data among different centers guarantees privacy and data is kept locally [2].

II. Background and Related works

Federated learning has been used for various imaging modalities such as MRI [3] [4], X-ray [5] retinotherapy [5] and for tasks such as brain tumor segmentation [6] [7] diagnosis [8] and treatment selection [7]. FL has shown great promise in developing models to support doctors in making treatment decisions for COVID-19 patients; it was investigated and reported that FL had a clear impact on patient care in a large-scale study on COVID-19 patients across 20 centers on five continents [9]. They used chest X-Ray imaging data in addition to clinical data to determine hospital triage for level of care and oxygen requirement in COVID-19 patients. They demonstrated that federated learning improved model performance for clients with limited datasets, compared to when they were trained on their local data. Another finding was that medical centers with smaller datasets had some classes with only a few patients resulting in underrepresented categories. These clients saw a significant improvement in prediction for those patient categories, which is especially important because, in COVID-19, patients with severe symptoms are generally in categories with fewer samples than a larger pool of patients with moderate symptoms. However, their care is more critical and requires more attention.

This paper is being prepared with IEEE standards This work was funded in part by NWO under project AMICUS

Erfan Darzidehkalani is with UMCG (e-mail: e.darzidehkalani@umcg.nl).

Several recent studies have been done to classify the COVID-19 images from healthy scans and to locate the lesion areas. The primary focus of AI tooling in the management of COVID-19 patients is interpreting radiology images, mainly chest CT, which has been widely applied for detecting lung changes to inform patient management, and guide treatment decisions [10] [11] [12]. Other studies have investigated 3D classification networks, [13] or Covid-19 detection with limited training samples. Most of the above studies achieved good accuracy and assumed a centralized environment where one data center has access to all the data. Researches that studied COVID-19 in a distributed setting, include [14] [15] [16].

The global aggregation models used in above researches were limited to model averaging in federated [16], [14], or blockchain setting [15]. However, other researchers have shown that the existing algorithms suffer from communication overhead [17] Also, increasing the number of participating hospitals can lead to convergence issues or catastrophic forgetting [3] [18]. As a result, comparing multiple federated learning algorithms under a standard setting could be informative in evaluating their applicability in practice. To evaluate the existing methods from multiple perspectives, we have implemented the most popular models and compared them in terms of performance, communication overhead, and computation burden.

III. Algorithms

Centralized data sharing In Centralized data sharing (CDS), data is stored in a central location and can be accessed by all clients. This is in contrast to federated and decentralized data sharing methods, where data is stored in multiple locations and accessed by multiple users. We use CDS as a comparison baseline to other algorithms.

Federated averaging: The learning procedure for federated averaging is an iterative process containing local and global steps. Each data owner trains a model received from a global server on its local dataset in local iterations. [19]The global server updates the global model by aggregating the updated local models. Then it sends it back to clients for the next round. The optimization problem for federated averaging could be formulated as $w = \sum_{i=1}^{N} p_i w_i$, $w_i = \arg\min_{w_i} (\mathcal{L}(\mathcal{D}_i; w))$ where N is the number of data owners, $\mathcal{L}(\mathcal{D}_i; w)$ is a loss function indicating global model parameters w of local datasets. The global server selects a subset of clients at each global round and sends the most recent global model to them. Then each client performs local training over its dataset for a selected number of epochs. The updated local models are calculated on selected batches. Local optimization can be formulated as $w_i \leftarrow w - \eta \cdot \nabla \mathcal{L}(w; \mathcal{D}_i)$, where η is the learning rate. Several local iterations might be required to go over all the local data. Local training procedures can be done for several local epochs. (3) The global model can be updated based on the local models w_i and is shared for aggregation: $w = \sum_{i=1}^{N} p_i w_i$, to update the global model for the next FL round.

Federated stochastic gradient descent: FedSGD is a federated approach to SGD. It uses a large-batch synchronous approach to multi-client learning, which performs better than naive asynchronous SGD training [3]. A subset is selected from the total number of clients, C defines the ratio of selected clients to the total number of clients. In federated stochastic gradient descent $C < 1$, for $C = 1$ the training would be non stochastic (full batch) since all the clients are involved. The gradient will be computed over the selected batch of clients [20] [21]. Similar to federated averaging, the selected subset of clients perform local training and the global server aggregates the updates.

Cyclic weight transfer: Recently, data-distributed deep learning methods have been successfully developed for training medical image classification models with cyclic weight transfer (CWT) approach. [5] In the CWT approach, models are trained at one institution at a time for several iterations before transferring the updated weights to the next institution cyclically until model convergence. A fundamental limitation with the existing implementation of CWT is that it is not optimized to handle variability in sample sizes, label distributions, resolution, and acquisition settings across institutions' training data. CWT performance decreases when these variabilities are introduced. For CWT to be utilized in practice, it must be capable of handling variabilities that would be found in most real-world medical imaging datasets. [2]

Single weight transfer: Single weight transfer (SWT) is another federated learning model widely used in the medical imaging domain. In Single weight transfer, models are trained in each institution for a number of iteration and then the updated model is transferred to the next institution. The difference between this method and CWT is that here the model passes each institution only once.

Stochastic weight transfer: In stochastic weight transfer (SWTS), we select a subsample of clients and train them in a cyclic manner. Similar to FedSGD, a ratio defines the number of selected clients to the total number of clients in each federated round.

IV. Experiment

Dataset Our experiments used two publicly available data sources, the Tongji hospital dataset and Brazil's SARS-CoV-2 dataset. Tongji dataset consists of 349 chest scans of COVID-19 positive and 397 scans of healthy subjects, all low-resolution CT modalities. Brazil's SARS-CoV-2 dataset consists of 2482 samples, 1252 scans of COVID-19-infected patients, and 1230 healthy subjects collected from multiple hospitals in Sao Paulo, Brazil. Train and test sets were obtained from random permutations of two aggregated datasets. Table I, shows data distribution.

TABLE I: Data distribution

Data	Class	Dataset	Train	Test
Covid	Brazil	1252	1451	150
	Tongji	349		
Non-Covid	Brazil	1230	1477	150
	Tongji	397		

Preprocessing Images were selected as 2D slices in greyscale. Preprocessing included randomly cropping between

0.5 to full size, random horizontal flipping, and intensity normalization. Images were all resized to 224×224. Figure 1 shows samples of processed images.

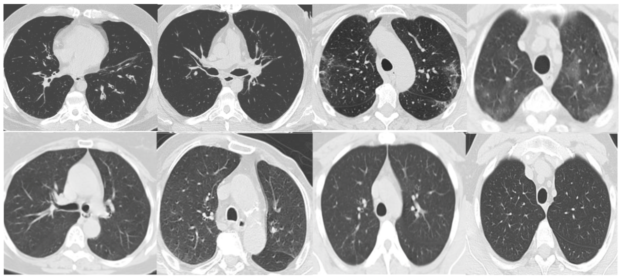

Fig. 1: Sample CT-scan images of Covid-19 images (top row) and Non-covid images (bottom row)

Training ResNet-18 is used as the backbone deep learning model. ResNet-18 comprises one initial block cascaded to four middle blocks. The initial block is made of convolutional, batch normalization, ReLU, and pooling layers. Middle blocks have the same layers, connected with straight and skip connections. The model is pre-trained on ImageNet dataset [22] with a CrossEntropy loss function and learning rate of 0.05. Each federated round consisted of 20 internal epochs for each client and batches of 16 samples in each iteration. For models which use minibatch training, like STWT and FedSGD, a subset of clients is randomly selected. We performed training with various participating clients and federated rounds to evaluate their effect on final performance. Models are also trained in a centralized, non-federated setting to build a comparison baseline.

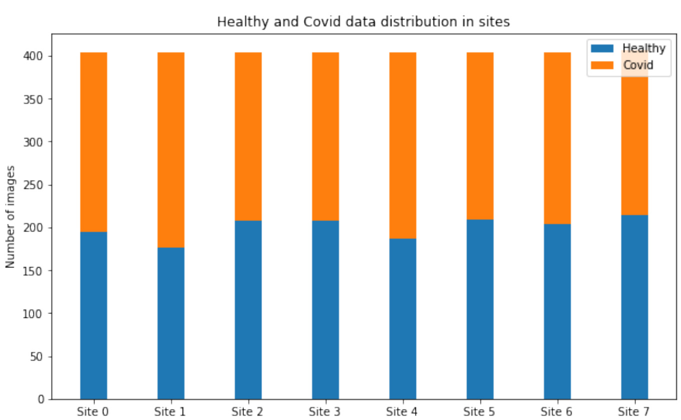

Fig. 2: Data distribution of each client in the simulated federated setting

Evaluation Standard classification metrics, accuracy, recall, precision, and F1 score, were used as our evaluation criteria. We also evaluate the level of communication, the amount of transferred data in each algorithm, and the computational complexity of the models.

V. Results

Here, the result for the setting with 10 participating clients and a maximum of 10 rounds is presented. The results are average performance among clients for all the federated rounds. Table II shows the results.

TABLE II: Comparison of federated learning algorithms on classification of COVID-19 data for 10 clients, averaged performance in all the 10 rounds.

Method	Accuracy	Recall	Precision	F1 score
CDS	87.75%	89.57%	87.93%	87.19%
FedAVG	66.72%	70.02%	43.80%	51.7%
FedSGD	65.17%	68.24%	43.86%	47.75%
CWT	87.75%	89.00%	88.67%	87.52%
SWT	64.60%	74.33%	65.55%	59.66%
STWT	84.21%	84.09%	83.33%	81.71%

Effect number of federated rounds

To evaluate the effect of number of rounds, models with 3,5,10 and 15 rounds are tested. The test results are shown for both centralized and federated learning algorithms. Table III shows the results of our experiment. The increasing number of rounds correlates with higher accuracy of the global model.

TABLE III: Effect number of rounds on accuracy of federated learning algorithms for 10 clients, 20 internal epochs.

Method	3 rounds	5 rounds	10 rounds	15 rounds
CDS	85.06%	81.56%	91.06%	91.04%
FedAVG	56.05%	63.78%	69.64%	70.73%
FedSGD	50.88%	55.9%	75.59%	76.94%
CWT	80.77%	**89.78%**	**91.27%**	**93.56%**
STWT	**90.73%**	83.97%	89.44%	93.01%

Effect number of participating clients To evaluate the number of clients on the FL network, we examined scenarios with 3,5, and 8 participating institutions. We trained each of the clients in 20 internal epochs. The number of Federated rounds for all the algorithms (except SWT) was 10. The average test results are shown in the Figure 3.

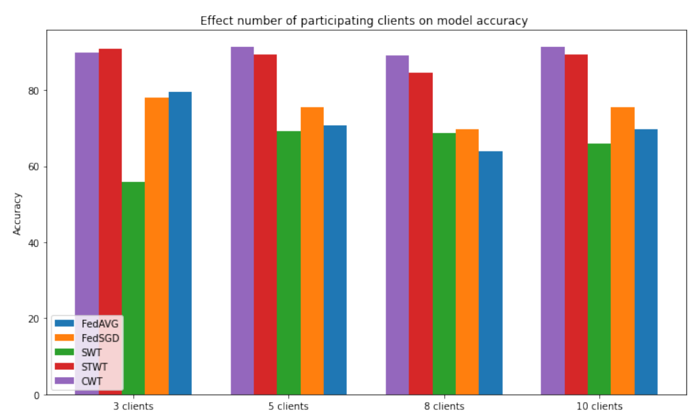

Fig. 3: Accuracy of federated learning algorithms with different number of clients.

Communication can also be a bottleneck in this setting. In methods like federated averaging, the lower bounds for total communicated data are proportional to $\sim 2NT$ where T is the number of rounds and N is the number of participating clients. In CWT, this lower bound is $\sim NT$. In our setting, we use a ResNet 101 model. We calculated the overall transferred

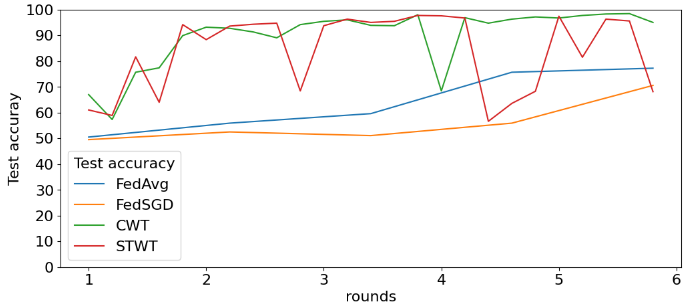

Fig. 4: Test accuracy of passing rounds

TABLE IV: Computation time (seconds) for FL algorithms for standardized setting

Method	3 clients	5 clients	8 clients	10 clients
FedAVG	8934 sec	8975 sec	9002 sec	9030 sec
FedSGD	8810 sec	8853 sec	9013 sec	9052 sec
CWT	5119 sec	5450 sec	5383 sec	5556 sec
STWT	2805 sec	5243 sec	6101 sec	6129 sec
SWT	**543 sec**	**547 sec**	**589 sec**	**618 sec**

data for the different number of rounds. As expected, in settings where clients are chosen stochastically, the number of communication is lower than full-client participation. Also, the experiments for computations cost show that non-sequential models generally have higher computational requirement than sequential models.

TABLE V: Comparison of total transferred data in a normalized setting in (GB)

Method	3 rounds	5 rounds	10 rounds	15 rounds
FedAVG	1.371	2.286	4.571	6.857
FedSGD	0.823	1.371	2.743	4.114
CWT	0.686	1.143	2.286	3.428
STWT	0.411	0.686	1.371	2.057

VI. Discussion

Our results show that federated learning has comparable performance to centralized data sharing, with the advantage of keeping data private. With large volumes of data and after high number of rounds, centralized data sharing and cyclic weight transfer have the highest accuracy.

Sequential models are susceptible to catastrophic forgetting, where a global model performs well on the latest client it has seen while having poor performance in other clients. Conversely, in algorithms like FedAvg and FedSGD, the models are averaged asynchronously after all the clients have finished their training. So the trajectory is smoother and overall improving with more communication rounds. As shown in Figure 4, local test results can have a high variance when passing through clients sequentially, indicating the catastrophic forgetting effect.

Models like FedAvg, and FedSGD, in which all the clients have identical copies of one global model, are slower and more challenging to converge compared to sequential models like CWT and STWT. Also, FedAvG and FedSGD require more training resources due to active server participation, resulting in more computation and network consumption. Stochastic selection of participants is an efficient way of training. Stochastic models save significant time and resources while having similar performance to full client participation. Overall, CWT has better results. These findings could be practical in further federated deployments in medical institutions.

Sequential models like CWT and STWT perform better than non-sequential models on fewer training rounds. For example, after three rounds of training, STWT and CWT both reach 96% accuracy, while FedAvG reaches 66%, and FedSGD performs equally to a random classifier. As the training proceeds, FedAVG and FedSGD gradually improve with more global rounds. The concept of sequential models is similar to fine-tuning [23] in centralized deep learning, so in cases where a hospital temporarily joins an FL network, or there is an urgency in training, sequential models are a better option.

More training rounds do not always lead to a better global model. Although average performance on all clients improves, more global rounds lead to worse performance for some clients. The global model can overfit some clients, leading to lower performance on others. [24] Some studies suggested early stopping and fine-tuning to local dataset after global training is finished [25]. In all the algorithms, more clients resulted in slower convergence. This effect is stronger in the FedAvg algorithm. In FedAVG, the Global model must compromise between potentially disparate local minima. [26] Methods such as adaptive or stochastic selection of clients and momentum-based models help faster convergence. [27] Our results suggest that stochastic client participation is close to full client participation. The average results of four trials with varying rounds, shown in Table III indicate that stochastic client participation in FedSGD results in 5.23% performance loss and 40% less bandwidth consumption. In STWT, it results in only 1.25% less accuracy but saves 40% of communication and 11.3% of computation. Also, SWT improves local clients' performance up to 80% accuracy with 8 clients, and it requires extremely low bandwidth requirements. These results are in accordance with prior studies, showing that, in theory, stochastic and full client participation have similar global minimums [28]. Stochastic client selection can be advantageous when there are limited resources, or in larger networks with occasionally unavailable clients.

We did not assume any shift in clients' data. A more comprehensive analysis should consider the effect of the domain and distribution shifts on the performance of the algorithms. Also, inter-client data variability and the effect of heterogenous clients could be a future line of research.

VII. Conclusion

Federated learning enables extensive collaborations of hospitals to address medical imaging problems while keeping data private. Real-world implementation requires consideration of efficiency and hardware requirements in addition to model performance, especially in the healthcare field, which generally has limited infrastructure. We implemented five federated learning algorithms for COVID-19 detection and analyzed their efficiency and accuracy. Our results suggest that federated learning algorithms have comparable performance to centralized data sharing, with the advantage of keeping data private. They also show that the sequential methods are a better option in most of the scenarios. This study can be helpful in the deployment of federated learning systems in COVID-19 detection and medical image analysis in general.

Acknowledgement

This research is supported by KWF Kankerbestrijding and the Netherlands Organisation for Scientific Research (NWO) Domain AES, as part of their joint strategic research programme: Technology for Oncology IL. The collaboration project is co-funded by the PPP allowance made available by Health Holland, Top Sector Life Sciences Health, to stimulate public-private partnerships.

References

[1] S. Kogilavani, J. Prabhu, R. Sandhiya, M. S. Kumar, U. Subramaniam, A. Karthick, M. Muhibbullah, and S. B. S. Imam, "Covid-19 detection based on lung ct scan using deep learning techniques," *Computational and Mathematical Methods in Medicine*, vol. 2022, 2022.

[2] E. Darzidehkalani, M. Ghasemi-Rad, and P. van Ooijen, "Federated learning in medical imaging: Part ii: Methods, challenges, and considerations," *Journal of the American College of Radiology: JACR*, pp. S1546–1440.

[3] M. J. Sheller, B. Edwards, G. A. Reina, J. Martin, S. Pati, A. Kotrotsou, M. Milchenko, W. Xu, D. Marcus, R. R. Colen, *et al.*, "Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data," *Scientific reports*, vol. 10, no. 1, pp. 1–12, 2020.

[4] S. Silva, B. A. Gutman, E. Romero, P. M. Thompson, A. Altmann, and M. Lorenzi, "Federated learning in distributed medical databases: Meta-analysis of large-scale subcortical brain data," in *2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019)*. IEEE, 2019, pp. 270–274.

[5] N. Balachandar, K. Chang, J. Kalpathy-Cramer, and D. L. Rubin, "Accounting for data variability in multi-institutional distributed deep learning for medical imaging," *Journal of the American Medical Informatics Association*, vol. 27, no. 5, pp. 700–708, 2020.

[6] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. S. Kirby, J. B. Freymann, K. Farahani, and C. Davatzikos, "Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features," *Scientific data*, vol. 4, no. 1, pp. 1–13, 2017.

[7] J. Lee, J. Sun, F. Wang, S. Wang, C.-H. Jun, and X. Jiang, "Privacy-preserving patient similarity learning in a federated environment: development and analysis," *JMIR medical informatics*, vol. 6, no. 2, p. e7744, 2018.

[8] I. Pan, H. H. Thodberg, S. S. Halabi, J. Kalpathy-Cramer, and D. B. Larson, "Improving automated pediatric bone age estimation using ensembles of models from the 2017 rsna machine learning challenge," *Radiology: Artificial Intelligence*, vol. 1, no. 6, p. e190053, 2019.

[9] M. Flores, I. Dayan, H. Roth, A. Zhong, A. Harouni, A. Gentili, A. Abidin, A. Liu, A. Costa, B. Wood, *et al.*, "Federated learning used for predicting outcomes in sars-cov-2 patients," *Research Square*, 2021.

[10] L. Yan, H.-T. Zhang, J. Goncalves, Y. Xiao, M. Wang, Y. Guo, C. Sun, X. Tang, L. Jing, M. Zhang, *et al.*, "An interpretable mortality prediction model for covid-19 patients," *Nature machine intelligence*, vol. 2, no. 5, pp. 283–288, 2020.

[11] Y. Hu, J. Jacob, G. J. Parker, D. J. Hawkes, J. R. Hurst, and D. Stoyanov, "The challenges of deploying artificial intelligence models in a rapidly evolving pandemic," *Nature Machine Intelligence*, vol. 2, no. 6, pp. 298–300, 2020.

[12] E. Burian, F. Jungmann, G. A. Kaissis, F. K. Lohöfer, C. D. Spinner, T. Lahmer, M. Treiber, M. Dommasch, G. Schneider, F. Geisler, *et al.*, "Intensive care risk estimation in covid-19 pneumonia based on clinical and imaging parameters: experiences from the munich cohort," *Journal of clinical medicine*, vol. 9, no. 5, p. 1514, 2020.

[13] X. Wang, X. Deng, Q. Fu, Q. Zhou, J. Feng, H. Ma, W. Liu, and C. Zheng, "A weakly-supervised framework for covid-19 classification and lesion localization from chest ct," *IEEE transactions on medical imaging*, vol. 39, no. 8, pp. 2615–2625, 2020.

[14] W. Zhang, T. Zhou, Q. Lu, X. Wang, C. Zhu, H. Sun, Z. Wang, S. K. Lo, and F.-Y. Wang, "Dynamic-fusion-based federated learning for covid-19 detection," *IEEE Internet of Things Journal*, vol. 8, no. 21, pp. 15 884–15 891, 2021.

[15] R. Kumar, A. A. Khan, J. Kumar, N. A. Golilarz, S. Zhang, Y. Ting, C. Zheng, W. Wang, *et al.*, "Blockchain-federated-learning and deep learning models for covid-19 detection using ct imaging," *IEEE Sensors Journal*, vol. 21, no. 14, pp. 16 301–16 314, 2021.

[16] T.-T. Ho, K.-D. Tran, and Y. Huang, "Fedsgdcovid: Federated sgd covid-19 detection under local differential privacy using chest x-ray images and symptom information," *Sensors*, vol. 22, no. 10, p. 3728, 2022.

[17] S. W. Remedios, J. A. Butman, B. A. Landman, and D. L. Pham, "Federated gradient averaging for multi-site training with momentum-based optimizers," in *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning*. Springer, 2020, pp. 170–180.

[18] K. Chang, N. Balachandar, C. Lam, D. Yi, J. Brown, A. Beers, B. Rosen, D. L. Rubin, and J. Kalpathy-Cramer, "Distributed deep learning networks among institutions for medical imaging," *Journal of the American Medical Informatics Association*, vol. 25, no. 8, pp. 945–954, 2018.

[19] E. Darzidehkalani, M. Ghasemi-Rad, and P. van Ooijen, "Federated learning in medical imaging: Part i: Toward multicentral health care ecosystems," *Journal of the American College of Radiology: JACR*, pp. S1546–1440.

[20] D. Chai, L. Wang, K. Chen, and Q. Yang, "Fedeval: A benchmark system with a comprehensive evaluation model for federated learning," *arXiv preprint arXiv:2011.09655*, 2020.

[21] Z. Charles, Z. Garrett, Z. Huo, S. Shmulyian, and V. Smith, "On large-cohort training for federated learning," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[23] Y.-T. Chen, Y.-C. Chuang, and A.-Y. Wu, "Online extreme learning machine design for the application of federated learning," in *2020 2nd IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS)*. IEEE, 2020, pp. 188–192.

[24] M. Mohri, G. Sivek, and A. T. Suresh, "Agnostic federated learning," in *International Conference on Machine Learning*. PMLR, 2019, pp. 4615–4625.

[25] T. Yu, E. Bagdasaryan, and V. Shmatikov, "Salvaging federated learning by local adaptation," *arXiv preprint arXiv:2002.04758*, 2020.

[26] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of fedavg on non-iid data," *arXiv preprint arXiv:1907.02189*, 2019.

[27] W. Liu, L. Chen, Y. Chen, and W. Zhang, "Accelerating federated learning via momentum gradient descent," *IEEE Transactions on Parallel and Distributed Systems*, vol. 31, no. 8, pp. 1754–1766, 2020.

[28] Y. J. Cho, J. Wang, and G. Joshi, "Client selection in federated learning: Convergence analysis and power-of-choice selection strategies," *arXiv preprint arXiv:2010.01243*, 2020.