RESPECTING PATIENT PRIVACY WITH FEDERATED ARTIFICIAL INTELLIGENCE

MOHD ADLI MD ALI*[1], EDRE MOHAMMAD AIDID[2], HAFIDZUL ABDULLAH[1]

[1]Dept. Physics, International Islamic University of Malaysia, Kuantan, Malaysia
[2]Dept. Community Medicine, International Islamic University of Malaysia, Kuantan, Malaysia

**Corresponding author: qunox@iium.edu.my*

(Received: 18th June 2021; Accepted: 20th August 2021; Published on-line: 30th September 2021)

ABSTRACT: Multiple research has shown that deep artificial neural networks (ANN) can assist physicians in diagnosing a patient with greater accuracy and sensitivity. Nonetheless, the grand march of success by ANN is only possible by the availability of an open medical dataset. However, at the time of writing, there is no open medical dataset from the Malaysian population. The local dataset is crucial to validate the performance of any ANN modal on the local populations. The lack of any local dataset may be due to local medical institution's hesitance to release any medical images and records to respect patient's confidentiality. One way around this is to adopt the Federated Learning system, in which no sharing of patient data is required. Our experiment tested the capability of 25 ANN models to classify chest radiograph images into three classes: normal, bacterial pneumonia, and viral pneumonia. Each ANN model is given a training dataset that is random in size and class ratio. The result obtained from the experiment shows that the federated system obtains the highest score in all measured metrics. It obtained a score of 0.76, 0.72, and 0.72 for average weighted precision, weight sensitivity, and F1, respectively. It also has the lowest standard deviation in all performance metrics compared to other learning systems. The result obtained here further strengthens the notion that if Malaysia wants to adopt a national-level artificial intelligent system for medical purposes, it should utilize the federated learning system at its core. It ensures Malaysia has an artificial intelligence system that respects patient's privacy while maintaining its robustness.

KEY WORDS: *Artificial Neural Network, Federated Learning, Ensemble Learning, Chest Radiograph*

1. INTRODUCTION

In recent years, a significant number of research studies have shown that deep learning algorithms can assist physicians in providing high-level automation of rudimentary tasks and analysis. Such task includes image classification, tumor localization, region of interest segmentation and radiation treatment planning. Certain research, such as by Hollon et al. (2020) and McKinney et al. (2020), has even shown that the deep learning model outperforms professional judgment.

The remarkable progress of deep learning in medicine can be attributed to 1) the availability of cheap cloud computing resources for model development, 2) a

variety of model architecture selections, ranging from U-net (image segmentation, Ronneberger et al., 2015), to Yolo and Faster R-CNN (localization, Redmon et al., 2016; Ren et al., 2015), and 3) most importantly the growing number of open-access medical datasets.

1.1 Validating on Local Dataset

It is crucial to point out that most deep learning developments are done using open-access datasets. These datasets are usually acquired by a medical institute that serves a specific local population. For example, the three most commonly used chest radiograph datasets are the JRST, Shenzhen, and Google-NIH datasets, which comes from the Japanese, Chinese and unspecified populations (Shiraishi et al., 2000; Jaeger et al., 2014; Wang et al., 2017).

Since each dataset represents a specific population, the class ratio in the dataset also closely reflects the incidence rate of a particular disease at a specific region and time. For example, according to Ferlay (n.d.), the incidence of a male having lung cancer per 100,000 population in Japan, China, the United Kingdom, and Malaysia are 41.4, 47.8, 35.5, and 22.5, respectively. Since the incidence rate among nations is varied and shifts through time, it implies that most clinical dataset is a non-independent and identically distributed (Non-IID) dataset. Thus, deep learning model developers should not naively assume that the obtained model performance from a specific dataset is valid for all populations globally (Kelly et al., 2019).

In this paper, we demonstrated how different class ratios and sample counts in different training datasets could greatly fluctuate the performance of a classification model. Similar results also have been shown by Zhao et al. (2018) and Roth et al. (2020). Due to differences in incidence rate (class ratio), it must be strongly pointed that any deployment of machine learning solutions for clinical purposes must be first validated using a local dataset. However, there is yet any open clinical dataset that comes from Malaysian populations.

There are several reasons why local medical institutes hesitate to release any clinical dataset, ranging from lack of computing resources to host such datasets to simply want to respect patient privacy. Despite these constraints, Malaysia's medical institute should not ignore the many proven benefit of deep learning solutions. Therefore, a middle ground between respecting patient privacy while allowing the integration of deep learning into local clinical practice should be proposed. One such method is the federated deep learning method.

2. LITERATURE REVIEW

Federated learning is a form of decentralized learning in which models are trained locally in the participant edge devices using their own local dataset. There are many forms of federated learning, though the majority of them do not require any centralized data gathering or cross-sharing between participants. Co-operative learning is done through model sharing instead. Thus, medical institutes are not required to disclose any patient information to another party, including other medical institutes. For a more rigorous explanation on the concept and application of federated learning, interested readers should read Yang et al. (2019). While Rieke et al. (2020) focused their study on federated learning specifically for medical and health purposes.

2.1 Example of Federated Learning in Medicine

The majority of federated learning literature focuses more on proposing new architecture and frameworks, integrating various technologies such as block-chain (Kumar et al. 2021; Rahman et al. 2020) and peer-to-peer (Roy et al., 2019) to make federated learning more secure, lightweight, and flexible. Consequently, there is an insufficient number of comparative studies that highlight the difference in performance between the various federated learning framework and conventional centralized learning. For example, research done by Lee and Shin (2020) showed federated learning only obtains an F1-score of 0.807, less than the centralized learning score of 0.814 in classifying the ECG dataset. Similarly, Lu et al.'s (2020) research showed the AUC-score for classifying breast cancer histology slide is 0.932 for the federated, less than the 0.946 scores obtained by centralized learning. The results from these two experiments imply that the current federated learning method delivers classification performance that is lesser than the centralized learning method. Therefore, it highlights the need for further research to develop a federated system framework capable of providing classification performance on par with centralized learning. It is worth noting that, due to the recent pandemic, several research studies demonstrated the application of federated learning for COVID-19 diagnostic, for example, Kumar et al. (2020), Qayyum et al. (2021), Qian and Zhang (2021), Liu et al. (2020) and Zhang et al. (2021).

2.2 Malaysia Context

In the writing of Kelly et al. (2019), the authors stated several challenges of implementing deep learning solution into everyday clinical practice; it includes algorithm bias, usage of clinically non-reliable metrics, logistical difficulties and integration difficulties. In the context of Malaysia and other similar developing countries, the lack of computational resource and sparse network infrastructure pose other critical challenges. Even at the time of writing, a significant number of government hospitals and clinics still do not have an electronic health records (EHR) system, i.e., computerize documentations. Without such infrastructure, it is difficult for the medical institute to benefit from the recent progress in EHR and clinical artificial intelligence.

The lack of computing infrastructure in many medical institutes in Malaysia does put federated learning at a disadvantage compared to conventional centralized learning. In centralized learning, the computationally intensive part of training a classification model is outsourced to a single centralized training server. Therefore, local medical institutes do not require to have their own server to train their models. In contrast, federated learning requires participants to train models locally, thus having a certain standard of computing hardware.

A remedy to this problem is to develop lightweight model that does not require extensive computing resource. Instead of opting for a model with a large number of trainable parameters such as ResNet50 (He et al., 2016), Xception (Chollet et al., 2017) and VGG16 (Simonyan et al., 2014). A model with a smaller number of trainable parameters can be opted. Such lightweight models are DenseNet121 (Huang et al., 2017) and MobileNet (Howard et al., 2017). In this paper, we have opted to use DenseNet121 for this very reason as it is deemed to be more practical to be run in a Malaysian medical institute that lacks powerful computing resources.

3. METHODOLOGY

In this paper, we demonstrate two ways of implementing federated learning. In the first approach, each local model was trained using a local dataset without any external input. After training was finished, all models were the ensemble to vote the most likely class for a particular sample. This method, akin to the bagging method, is ensembled learning; we call this approach the Federated Ensemble Learning.

The second method is called Federated Averaging Learning, in which each participant cooperatively learned during model training. In this implementation, after each training epoch, each model's weight was gathered. A new model was created, in which the model's weight is the average of the previously gathered models. The newly created model was then distributed back to learn the local dataset until the next epoch. The cycle of training locally and centrally averaging the weight continued till the end of the training. In the end, each participant would have one model in which its weight is akin to the other.

3.1 Dataset

The chosen dataset was taken from Kermany et al. (2018), which contains 5,856 chest radiograph images. The images were divided into three classes: normal, bacterial pneumonia, and viral pneumonia. To recreate Non-IDD datasets, the training samples were randomly divided into 25 smaller training datasets, each with a different size and class ratio. The largest dataset contained 430 samples while the smallest contained only 29 samples; the ratio of pneumothorax sample to a normal sample is between 75.51% to 67.93%, this is shown in Fig. 1.

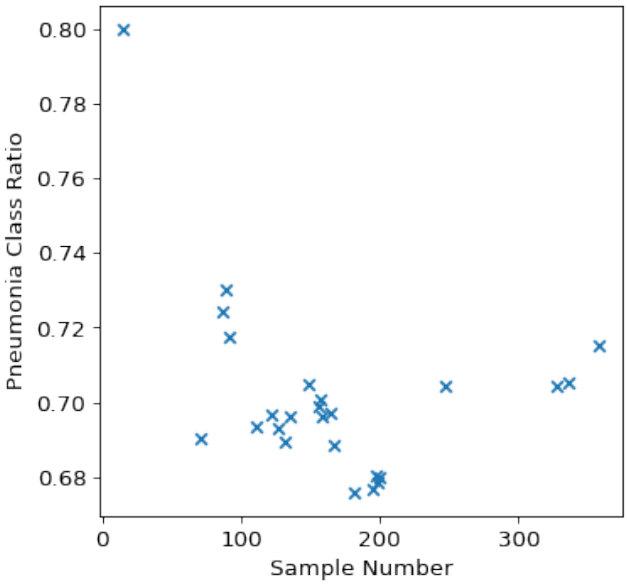

Fig. 1. The sample number Vs the class ratio for the 25 Non-IDD datasets.

3.2 The Simulations

Four different kinds of deep learning frameworks are simulated: centralize learning, fragmented learning, federated ensemble learning, and federated average learning. The detail of each simulation is given below.

Centralize Learning (CL): the condition in which a single server gathers samples from all medical centers. Then a single classification model is created and trained using all available samples. The trained model will then be replicated and distributed back to all medical centers. To simulate this condition, a single classification model was created and learned from all samples in the training dataset.

Fragmented Learning (FRL): opposite to the CL, this framework does not have sample sharing nor cooperative learning. Each medical center trains its model only using its training dataset. To simulate this, twenty-five identical classification models were created, and each trained with one of the non-IDD training datasets created earlier.

Federated Ensemble Learning (FEL): to simulate this condition, an N number of models that were previously created in the FL simulation were randomly taken to create a single ensemble model. 30 different combinations of ensemble models were created, each using different individual models. The simulation started with N = 5 then repeated with N = 10, 15, and 30.

Federated Average Learning (FAL): To simulate this approach, N number of models were created and initialized to have identical weights. Each model was trained with separate dataset. For each training epoch, the average weight of each model's trainable neurons was calculated. Then each neuron's weights would be adjusted to take the new calculated average weight. This process continued until the end of the training. The simulation was repeated with the initial number of participating models, N = 5, then 10 and lastly 15.

3.3 The Classification Model

All classification models created in all four frameworks are identical in architecture and initial weight. Dense121 was chosen as the base model, with its weight taken from ImageNet (Deng et al., 2009). Models were trained with the epoch set at 25, though early stopping is in place with the condition that the validation-loss does not change after ten epochs. The TensorFlow (Abadi et al., 2016) library version 2.4.1 was used for the model implementation and executed using the Google Cloud Computing infrastructure. Model classification performance evaluations were done using the sklearn library (Pedregosa et al., 2011) and the same testing dataset. Since the testing dataset was imbalanced, the metrics chosen were precision, sensitivity, and F1 score.

4. RESULT AND DISCUSSION

Fig. 2 shows the classification performance score of individual FRL models when tested on the same testing dataset. The score was plotted based on their testing dataset size and class ratio. Additionally, the dashed line shows the score obtained by a CL model. The figure shows that most FRL models score well below the CL score for every metric. More importantly, several FRL models obtained scores below the 0.5 thresholds, suggesting the model failed to learn pneumonia classifications.

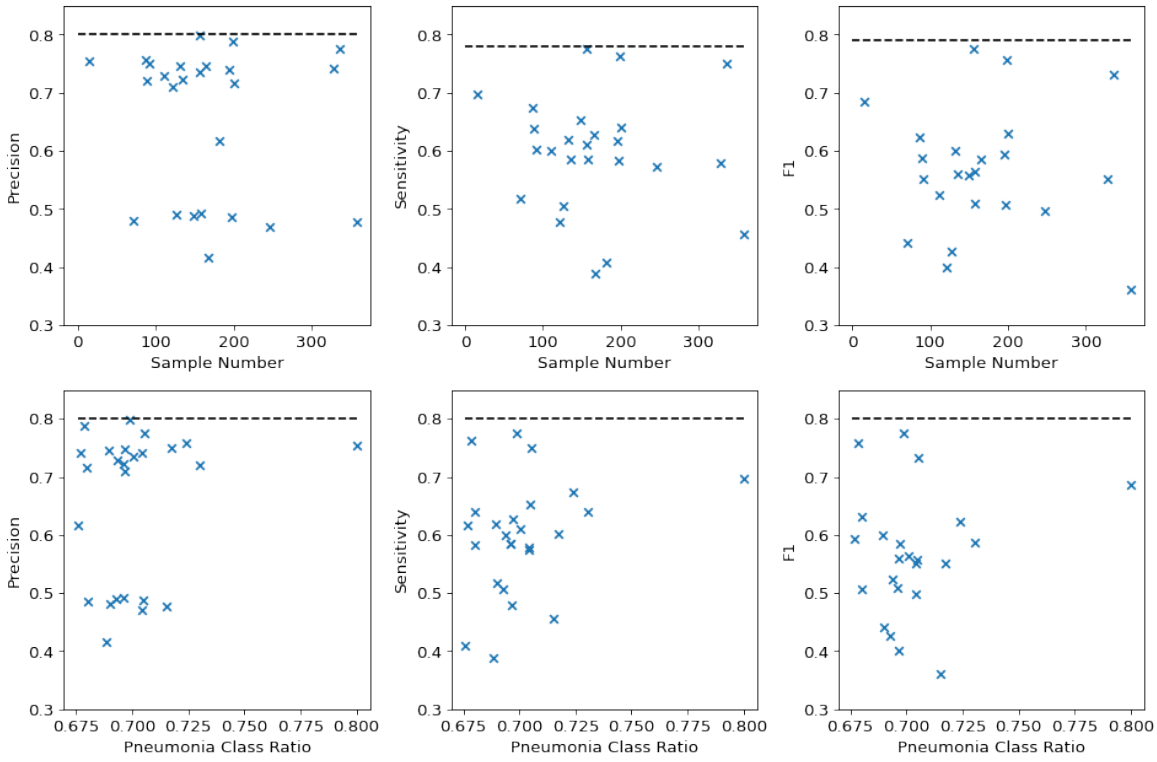

Fig. 2. The Classification performance metrics versus training sample number and pneumonia class ratio.

As stated before, in this experiment, we also want to investigate how Non-IDD datasets affect the classification performance of the deep learning model. Table 1 shows the Pearson correlation between a training dataset sample size and its class ratio to the resulting model classification performance score. The result shows that the classification result is more strongly correlated to training dataset sample size rather than the class ratio. This measurement validates the common assumption in the machine learning community that dataset size is more important than sample variety.

The result shown in Fig. 2 suggests that, in the FRL framework, there are possibilities that a deep learning model will fail to learn the appropriate class classifications. If such a failed model is deployed and used in a medical center, it will only annoy physicians over consistency misclassification in the best-case scenario. However, in the worst-case scenario, it can lead to misdiagnosis. Prolonged misclassification can also cause distrust and abandonment of the deep learning system as a whole.

Table 1. Pearson correlation score between a classification metrics and the training dataset size and pneumonia Class Ratio

Metrics	Dataset Size	Pneumonia Class Ratio
Precision	0.534	0.356
Sensitivity	0.741	0.205
F1	0.790	0.216

To avoid such circumstances, it is strongly recommended that medical center does not adopt the fragmented learning framework. Although the FRL framework ensures patient privacy, it does not lead to the creation of the most optimized deep learning model. Adopting a federated learning system will ensure only an accurate model is used. Nevertheless, if the FRL framework is not available for the medical center, the result from Table 1 suggests that model classification can be improved by simply expanding the training dataset size; variety in the training dataset comes second.

Table 2 shows the classification score obtained by the different learning framework. From the result, it is very clear that the centralized still obtained the highest score in every metric. The second-best learning framework is the FAL with 5 participants. The FEL framework only has a higher precision score than FRL but no significant improvement in sensitivity or the F1 score. The result also shows that by increasing the number of participating models, the standard deviation in FEL and FAL scores will decrease. This reduction in score's deviation is an important feature, as it suggests that both FAL and FEL produce models with consistent performance. Unlike in the case of the FRL framework, where there is a possibility that its model will fail to learn, it is unlikely the FAL and FEL produce a failed model. By having this consistency in performance, it encourages physicians to have confidence and reliability towards the system.

It is worth noting that for FAL, its average score decreases as the number of participants increases, the exact reason for this trend is unknown. One hypothesis is that as the number of participants model increases, it is harder for the model to converge to a 'one-fit all' solution. Further research is needed to validate this effect and also to find ways to mitigate it.

Despite FEL classification performance is less than FAL, FEL has its advantage of not requiring any centralize system to coordinate the learning. Medical institute can independently pick and choose trained model without the need to actively participate in the cooperative learning, this also reduce the overall network usage.

Table 2. Comparison of classification performance between different types of deep learning framework. Bold score shows the highest score recorded.

	Precision		Sensitivity		F1	
	Average	STD	Average	STD	Average	STD
SCL	0.801		0.78		0.791	
FRL	0.654	0.128	0.597	0.097	0.542	0.129
FEL (5)	0.716	0.079	0.597	0.063	0.546	0.083
FEL (10)	0.738	0.048	0.599	0.042	0.549	0.057
FEL (15)	0.739	0.047	0.599	0.032	0.549	0.047
FEL (20)	0.747	**0.010**	0.600	**0.029**	0.550	**0.043**
FAL (5)	**0.757**	0.023	**0.724**	0.041	**0.716**	0.051
FAL (10)	0.753	0.018	0.709	0.042	0.700	0.049
FAL (15)	0.739	0.015	0.674	0.038	0.664	0.043

5. CONCLUSION

In this paper, we have introduced the concept of federated learning and how it can be a middle ground solution for medical institutes that want to integrate deep learning into their everyday clinical task without jeopardizing patient data privacy. The result from our experiment shows that the federated learning method provides a better classification of pneumonia images than a fragmented learning system. However, centralized learning systems still provide are more accurate classification than federated learning. We also observed that for FAL, its classification performance decreases as the number of participant increase.

ACKNOWLEDGEMENT

This research is done under the Fundamental Research Grant Scheme, provided by the Malaysian Higher Education Fund, (FRGS19-181-0790).

REFERENCES

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., ... & Zheng, X. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467.

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1251-1258).

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition (pp. 248–255). Leee.

Ferlay, J., Ervik, M., Lam, F., Colombet, M., Mery, L., Piñeros, M., ... & Bray, F. (2018). Global cancer observatory: cancer today. Lyon, France: international agency for research on cancer, 1-6.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

Hollon, T. C., Pandian, B., Adapa, A. R., Urias, E., Save, A. V., Khalsa, S. S. S., ... & Orringer, D. A. (2020). Near real-time intraoperative brain tumor diagnosis using stimulated Raman histology and deep neural networks. Nature medicine, 26(1), 52-58.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., ... & Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4700-4708).

Jaeger, S., Candemir, S., Antani, S., Wáng, Y. X. J., Lu, P. X., & Thoma, G. (2014). Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. Quantitative imaging in medicine and surgery, 4(6), 475.

Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. BMC medicine, 17(1), 1-9.

Kermany, D., Zhang, K., & Goldbaum, M. (2018). Labeled optical coherence tomography (OCT) and Chest X-Ray images for classification. Mendeley data, 2(2).

Kumar, R., Khan, A. A., Kumar, J., Zakria, A., Golilarz, N. A., Zhang, S., ... & Wang, W. (2021). Blockchain-federated-learning and deep learning models for covid-19 detection using ct imaging. IEEE Sensors Journal.

Lee, G. H., & Shin, S. Y. (2020). Federated Learning on Clinical Benchmark Data: Performance Assessment. Journal of medical Internet research, 22(10), e20891.

Liu, B., Yan, B., Zhou, Y., Yang, Y., & Zhang, Y. (2020). Experiments of federated learning for covid-19 chest x-ray images. arXiv preprint arXiv:2007.05592.

Lu, M. Y., Kong, D., Lipkova, J., Chen, R. J., Singh, R., Williamsona, D. F., ... & Mahmood, F. (2020). Federated Learning for Computational Pathology on Gigapixel Whole Slide Images. arXiv preprint arXiv:2009.10190.

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., ... & Shetty, S. (2020). International evaluation of an AI system for breast cancer screening. Nature, 577(7788), 89-94.

Rahman, M. A., Hossain, M. S., Islam, M. S., Alrajeh, N. A., & Muhammad, G. (2020). Secure and provenance enhanced internet of health things framework: A blockchain managed federated learning approach. Ieee Access, 8, 205071-205087.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 779-788).

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems, 28, 91-99.

Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., ... & Cardoso, M. J. (2020). The future of digital health with federated learning. NPJ digital medicine, 3(1), 1-7.

Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention (pp. 234-241). Springer, Cham.

Roth, H. R., Chang, K., Singh, P., Neumark, N., Li, W., Gupta, V., ... & Kalpathy-Cramer, J. (2020). Federated Learning for Breast Density Classification: A Real-World Implementation. In Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning (pp. 181-191). Springer, Cham.

Roy, A. G., Siddiqui, S., Pölsterl, S., Navab, N., & Wachinger, C. (2019). Braintorrent: A peer-to-peer environment for decentralized federated learning. arXiv preprint arXiv:1905.06731.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.

Qayyum, A., Ahmad, K., Ahsan, M. A., Al-Fuqaha, A., & Qadir, J. (2021). Collaborative federated learning for healthcare: Multi-modal covid-19 diagnosis at the edge. arXiv preprint arXiv:2101.07511.

Qian, F., & Zhang, A. (2021). The value of federated learning during and post-COVID-19. International Journal for Quality in Health Care, 33(1), mzab010.

Shiraishi, J., Katsuragawa, S., Ikezoe, J., Matsumoto, T., Kobayashi, T., Komatsu, K. I., ... & Doi, K. (2000). Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. American Journal of Roentgenology, 174(1), 71-74.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., & Summers, R. M. (2017). Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2097-2106).

Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated machine learning: Concept and applications. ACM Transactions on Intelligent Systems and Technology (TIST), 10(2), 1-19.

Zhang, W., Zhou, T., Lu, Q., Wang, X., Zhu, C., Sun, H., ... & Wang, F. Y. (2021). Dynamic fusion-based federated learning for COVID-19 detection. IEEE Internet of Things Journal.

Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., & Chandra, V. (2018). Federated learning with non-iid data. arXiv preprint arXiv:1806.00582.