Exploring the Potential of Federated Learning for Medical Image Analysis in Non-IID Settings

Abhinav Gupta
AG7948@NYU.EDU
New York University, New York, NY, USA

Tejas Mahajan
TM3647@NYU.EDU
New York University, New York, NY, USA

Sai Charitha Akula
SCA321@NYU.EDU
New York University, New York, NY, USA

Abstract

Building medical image analysis solutions using deep learning has shown progress over the past few years but has not been able to effectively leverage unlabelled datasets and even not extensively study on how will the models have to be tuned with different data distributions while maintaining privacy. Many architectures have been proposed to pre-train image encoder using multi-modal data and evaluate their performance on downstream medical image classfication tasks, but those are all done in a centralized setting. In this work, we study the impact of pre-training a multi modal model on an in-domain medical imaging dataset i.e. MIMIC-CXR and finetuning the model on a medical imaging classification dataset in a federated learning setup under varying training data overall quantities and for different data distribution strategies namely label distribution, volume distribution and attribute distribution skews. While utilizing 100% of the trained data we get similar AUC score on the CheXpert test set for volume and attribute distribution settings in comparison with centralized trained model but observe a $\sim 3\%$ decrease when using the label distribution. In case of 1% and 10% training data we are getting a maximum of $\sim 10\%$ decrease in AUC scores for the label distribution settings.

1. Introduction

With the ever growing number of studies in medical imaging technologies, deep learning (especially self-supervised) algorithms have shown promise for learning stronger representations. However, access to datasets from different health institutions is infeasible due to privacy concerns. To address the privacy concerns, federated learning has been one of the proposed solution enabling a collaborative distributed training setup among different organizations without sharing direct access to raw data. Industrially, this would take us one step further towards deploying more efficient deep learning models for health care by accessing data from across the world. Academically, this would help us understand the generalization gap between self supervised federated learning models vs centralized models. Additionally, past studies have shown that learning from unlabelled data in pretraining tasks will help in scaling models towards future zero or fewer labelled fine tuning tasks. To this front, we build[1]

[1] https://github.com/tjdevWorks/ConVIRT-Federated

on these existing multi-modal medical image datasets oriented architectures by studying the impacts of training such system in a federated learning environment. This will not only help preserve the privacy of the hospitals data but also help in analyzing the impacts of scaling the solution in for different data distribution settings and number of clients.

2. Related Work

Multi-modal Learning While medical image understanding has seen rapid progress with deep learning, collecting large scale high quality annotated datasets or extracting annotations from textual reports using heuristics have been major bottlenecks in realization of these systems. The major limitations for this is primarily the cost, time and privacy concerns associated in collecting these annotations and other being the handcrafted rules for annotating from textual reports are suboptimal and limited to few categories. Recent studies by Chen et al. (2020), He et al. (2020) and Caron et al. (2021) in general purpose deep learning have shown that we can leverage from unlabeled data through contrastive representation learning from natural images and gain comparable performance on downstream tasks. But in context of domain adaptation towards medical image analysis, studies have shown that pretraining on out of domain tasks lead to poor generalization performance. To address this, several recent studies like Zhang et al. (2020), Tiu et al. (2022) and Huang et al. (2021) have shown that we can pretrain in domain image-text encoder networks through contrastive learning and obtain better generalization capabilities on various downstream medical image analysis tasks. However, these models can't be deployed directly in the health care industry because of the lack of data availability due to privacy concerns.

Federated Learning in Healthcare Federated Learning (FL) is a collaborative learning paradigm where multiple participants parallelly learn a model without sharing any kind of raw data and then a global model is updated with the participant models' parameters according to a agreed upon aggregation method. Federate Learning is a trending decentralized approach especially in the field of healthcare as it provides privacy and security of the hospital data, and also many economic and regulatory benefits.

In recent times, there have many examples where federated learning is adopted in the medical domain. Brisimi et al. (2018) presented a patient hospitalization chance prediction FL model for heart diseases. Liu et al. (2020) proposed an Aligning, Integrating and Mapping Network (aimNet) to learn image representations. They experimented with image captioning and VQA tasks, and validated their model on horizontal, vertical and transfer based federated learning settings. Qayyum et al. (2021) proposed a collaborative learning framework for an automatic multi-modal (X-ray and Ultrasound) diagnosis of COVID-19. Ju et al. (2020) presented a FL inspired technique to classify Electroencephalography (EEG) signals. Sheller et al. (2019) performed a study regarding brain tumor segmentations.

Importance of Data Distribution When the local training data in various hospitals is not independent and identically distributed (Non-IID), models trained in federated learning setting can perform worse than those trained in the centralized setting. To address this, Wim Casteels (2020) introduces a regularization scheme that can be used during training to ensure that the model captures universal causal correlations i.e., correlations that are present across all subsets. It's main idea is to divide the non-i.i.d training data into different sub populations and during training, introduce a regularization term to penalize correlations that are not universal. Hangyu Zhu (2021) has done a generic survey of FL on Non-IID data describing various data distributions, including attribute skew, label skew and quantity skew. They provide a detailed analysis of the influence of Non-IID data on both parametric and non-parametric machine learning models in both horizontal and vertical federated learning.

In this work, we extend the multi-modal ConVIRT [20] architecture to support training in federated Learning setting, both during pre-training and fine-tuning for the downstream task and study the impact of number of clients, different strategies of splitting data among these clients to simulate IID and Non-IID settings, and different aggregation algorithms, including FedAvg.

3. Experiment Setup

3.1. Dataset Description

We use the same pre-training and fine-tuning datasets as described in our baseline ConVIRT [20] paper to benchmark and compare the results properly.

Pre-Training: We use the MIMIC-CXR-JPG [10] dataset for the pre-training self-supervised task. MIMIC-CXR-JPG [10] is a large publicly available dataset of chest radiographs in JPG format, totally derived from MIMIC-CXR [9] which contains high resolution DICOM images with free-text radiology reports. As the JPG [10] dataset does not contain the free-text reports, we implemented a python web-crawler to directly download the reports from the physionet MIMIC-CXR [9] website and place them in the proper directory. Our aggregated dataset contains about 377,110 JPG format images and 227,827 free-text radiology reports associated with these images. This dataset is completely de-identified to satisfy the US HIPAA requirements.

Fine-Training: We evaluate our self-supervised model on a medical image classification task based on CheXpert [8] dataset. This large public dataset contains 224,316 chest radiographs of 65,240 patients along with their associated radiology reports. The task involves multi-label binary classification of chest radiographs for the following 5 labels: cardiomegaly, pleural effusion, atelectasis, consolidation, and edema. We randomly sampled 5000 images from the original training dataset to use as validation set, and used the original validation set as our test set. The resulting dataset contained 218414/5000/234 images in the respective train/val/test splits.

3.2. Method

Building on the ConVIRT architecture, we take a paired input (x_v, x_u), where x_v represents a group of images and x_u represents a text sequence describing the radiologists findings & impressions for the images in x_v. The goal here is to learn an image encoder f_v (transforming an image to a fixed dimensional vector) which will later be used to fine tune the model for a downstream classification task. After pre-training and finetuning, with the ConVIRT model in a centralized setting as the baseline, we will be training the same architecture in a federated learning environment by studying the impact of number of clients, different

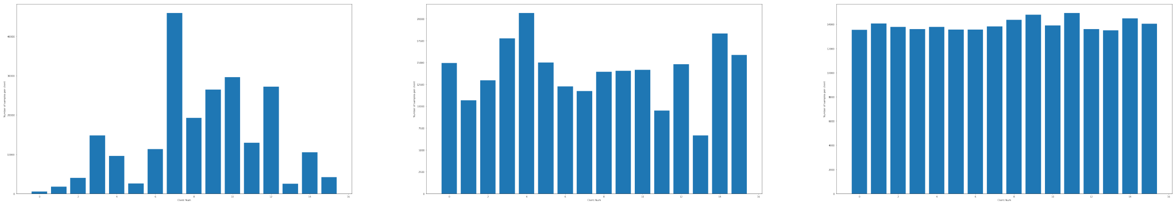

Figure 1: Quantity IID vs non IID Data Distribution. X-axis indicating the client number and Y-axis indicating the number of samples at the client

strategies of splitting data among these clients to simulate IID and Non-IID settings (utilizing 1%, 10% and entire training data) and the FedAvg [13] aggregation algorithm.

3.2.1. Pre-Training

The self supervised model here will consist of an image encoder and a text encoder that will be jointly trained on the MIMIC-CXR [10] dataset. We will be training the image encoder function (Resnet-50 or ViT-Base/32) and for text will be using the embeddings from the BERT language model previously trained on MIMIC clinical notes (ClinicalBERT [1]). During training time we allow the encoder to adapt to our contrastive task by freezing the embeddings and the first 6 transformer layers of this BERT encoder and fine-tuning the last 6 layers. Unlike how ConVIRT selects a random sentence for the complete radiology report for each image, we will be building on idea [17], wherein they only utilize the impression section of the report under the hypothesis that random text selection by ConVIRT causes inconsistencies during training. Similar to the InfoNCE loss, we will be computing the image-text contrastive loss by computing the log softmax scores of the cosine similarities of image-text and text-image feature vector pairs.

3.2.2. Fine Tuning

In the downstream fine tuning task we will be evaluating the pretrained image encoders (Resnet-50) on a multi-label classification task for chest disease detection using the CheXpert dataset. On the similar lines of the ConVIRT paper we will be fine tuning the model, wherein we will train both the backbone and the classification head.

Additionally, we finetune the model in a federated learning environment under different experimental settings. We simulate sub populations of downstream data for chosen number of clients, and finetune the model by sending a copy of the model to each of the clients, finetuning them locally at the clients on the simulated private downstream data and aggregating the parameters on central server.

3.3. Data Distribution Techniques

We setup various experiments with different number of clients and follow the below strategies to split data into IID and Non-IID settings.

Attribute Skew: Attribute skew occurs in scenarios in which the feature distribution across each clients' attributes is different from each other. These data attributes can be non-overlapped (Not a valid case for our dataset), overlapped or even the same. For the fully overlapping attribute skew:

1. Train data was divided into age sets (equal to the number of clients), based on percentiles, and these sets were then distributed among all clients in an equal manner for IID partitioning. For non-IID partitioning, each age set was assigned to a single client
2. Sub populations for clients are created by applying k-mean clustering on numeric features of images generated using Resnet-50. [18]

Label Skew: Label distribution skew occurs in scenarios in which clients' label distributions are different but their conditional feature distributions $P_k(x|y)$ are same.

1. Single Class Per Client: This occurs when the number of classes are more than number of clients. In the scenario, each label is assigned to a client randomly, with each entire label data belonging to a single client. This is to replicate the real scenario where rare diseases can occur in only a few hospitals.
2. Single Client Per Class: This arrangement occurs when the number of clients exceeds the number of classes. Under these circumstances, each client is assigned a single class and will contain only data from that class. It is possible for a single class to be assigned to multiple clients in this scenario

Quantity skew: Quantity skew occurs in scenarios in which client's have different number of training data samples. This can co-occur along with any of the above scenarios. Like in the above case of label distribution imbalance, different number of data samples are distributed to different client sub populations according to Dircichlet distribution, $q \sim Dir_N(\beta)$, where β indicates the amount of imbalance [15]

3.4. Federated Learning Framework

We utilize Flower and Ray to simulate the federated learning environment for model training and experimentation.

Flower is a framework that facilitates the creation of federated learning systems. It is designed to be highly customizable, meaning that it can be tailored to fit a variety of different use cases. It is also designed to be extensible, allowing for the incorporation of new state-of-the-art techniques. Additionally, Flower is compatible with any machine learning framework, and it is written with maintainability in mind, making it easy to read and understand the codebase. Flower's goal is to provide a tool that streamlines the development and deployment of federated learning systems in diverse settings. [2]

Ray is a framework for scaling AI and Python applications from a single device (such as a laptop) to a cluster. It allows you to use the same code for both small-scale and large-scale applications, and it is designed to be able to run any kind of workload efficiently

3.5. Training Setup

We used PyTorch Lightning with Hydra19 for rapid and reproducible ML experimentations. We use the same training parameters as mentioned in ConVIRT [20] to maintain consistency with the baseline. We have changed or added parameter values only when there was missing information in the paper.

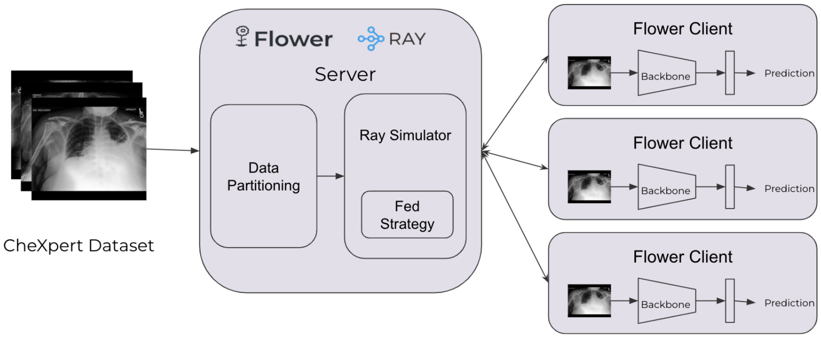

Figure 2: Federated Setup

Pre-Training: To prepare the data for pre-training, all images from the MIMIC-CXR-JPG [10] dataset were resized to 224×224 so that they could fit in memory during batch training. We kept the text model projected dimension size as 512. For the image model, we experimented with 'IMAGENET1K_V2' and 'ConVIRT [20]' weights to initialize the Resnet-50 backbone. We applied a series of transformations (like RandomResizedCrop, RandomAffine, ColorJitter, and Normalize) on the images, and tokenized the 'findings' and 'impressions' section text from the reports to form the input for our multi-modal network. We trained this setup for 105 epochs with a batch size of 128 on a single RTX8000 gpu node keeping the optimizer as Adam initialized with learning rate of 0.0001 and weight_decay as 0.000001. We ran a scheduler (i.e. ReduceLROnPlateau - mode: min, factor: 0.5 and patience: 5) for the learning rate so that it can adjust itself when the evaluation metric stops improving. We implemented a Contrastive loss function based on the details from the paper.

Fine-Tuning: To prepare the data for fine-training, we square padded all images from the CheXpert [8] dataset and resized them to 224×224. While creating the data module for the fine-tuning network, we only applied normalization transformation on CheXpert resized images.

For the centralized setting, we train a classification head on top of the unfreeze image backbone for 120 epochs with a batch size of 256 on a single RTX8000 gpu node keeping the optimizer as Adam (lr: 0.0001 and weight_decay: 0.000001), learning rate scheduler as ReduceLROnPlateau (mode: min, factor: 0.75, and patience: 10) and loss function as BCELoss because the task is multi-label classification.

For the federated setting, we use the Flower and Ray simulator. We ran each experiment for 120 rounds, and during each round local training was performed at the clients for one epoch. To analyze the impact of different parameters on the performance of the system, we varied: Sampling Fraction (i.e. the percentage of the total data to be used for the experiment - 100%, 10%, or 1%); Fit Fraction (i.e. the percentage of clients to be trained at each round - 0.125, 0.25, 0.5, 1); Skew Type (i.e. the type of skew to be applied to the experiment - attribute, label, quantity); Skew-specific parameters (i.e. whether the data is IID or non-IID, number of clients, etc.)

(*a*) Centralized Setting

Method	Baseline (AUC) 100% CheXpert [8] Dataset	Our Implementation (AUC) 100% CheXpert [8] Dataset
ImageNet	87.6	78.9
ConVIRT	88.1	87.07

(*b*) ConVIRT + Federated Fine-Tuning (Volume Distribution)

Distribution Type	Scale	Fine-Tuning - 100% CheXpert [8] Dataset (AUC)		
		4 clients	8 clients	16 clients
IID	-	87.82	87.29	87.38
Non-IID	1	87.95	**87.65**	**88.26**
Non-IID	10	**87.97**	87.18	87.59
Non-IID	100	87.6	87.5	87.72

(*c*) ConVIRT + Federated Fine-Tuning (Attribute Distribution)

Distribution Type	Fine-Tuning - 100% CheXpert [8] Dataset (AUC)		
	4 clients	8 clients	16 clients
Age: IID	87.5	86.93	88.08
Age: Non-IID	87.79	86.9	87.15
KMeans: IID	87.2	87.3	87.0
Kmeans: Non-IID	86.9	86.1	86.3

(*d*) ConVIRT + Federated Fine-Tuning (Label Distribution)

Exclusive	Equal Number Samples	Fine-Tuning - 100% CheXpert [8] Dataset (AUC)		
		2 clients	4 clients	5 clients
False	False	84.14	85.16	85.04
False	True	86.54	85.83	84.03
True	False	85.07	85.14	85.49
True	True	86.81	86.11	84.11

Table 1: Results for fine-tuning 100% CheXpert [8] data in different settings. The test set AUC scores mentioned in the table are the average scores for 5 classification labels.

4. Results

We have summarized and reported the mean AUC scores on the test set in Table-1 for the centralized setting and different data distribution federated setting approaches that we experimented with. In the appendix section A, we have displayed the mean AUC scores for the additional experiments we ran with 1% and 10% of the original fine-tuning dataset [8].

Comparison across total number of samples: We do not observe statistically significant differences between centralized and federated learning setting when full volume is used. However, under our partial volume settings (1% and 10%), we do see significant difference in performance between central and federated learning [Subtables: 0b, 1a, 1b]. We conjecture that this is due to the diverse samples at the client model when there is significant volume, which enables it to train well without relying on the global model.

Comparison across different skews: The most challenging setting is label distribution skew, when all the data of a label belongs to a single class. In contrast, the feature distribution skew and quantity skew setting have little influence. This is clearly seen in the labels vs feature performance under 1% setting [Subtables: 1b, 1d].

Comparison across fraction fit: When we reduce the sampling fraction (i.e. the percentage of total clients sampled and trained in every round) while keeping the overall volume constant, the performance does appear to decrease, particularly in low volume settings [Subtable: 1d]. We speculate that this is because the total number of samples (and therefore the information and diversity) is reduced during the parameter update in that round, leading to a decrease in overall performance (as shown in 1% training data).

Comparison across Label Skew specific parameters: Our results show that when we enforce exclusivity of the sample across clients, the performance decreases due to the reduction in sample diversity at each client [Subtable: 1d]. Similarly, we observe a decrease in accuracy when we enforce an equal number of samples across clients, as the number of samples at each client is reduced to the minimum number of samples present at any client [Subtable: 1d]

5. Discussions and Limitations

The aim of the study was to investigate the performance of using an in-domain image backbone models finetuned for a downstream medical image classification task in a federated learning setup. Despite promising results, there are still several factors that need to be considered in order to optimize local and global models. For example, it would be useful to experiment with different numbers of rounds and epochs to see if there is any trend in terms of auc scores in different settings.

In future work, it would be helpful to explore various federated learning strategies to find the best optimizer and learning rate dynamics for a given setting. It may also be useful to experiment with strategies to decouple the client learning rate from the global learning rate while still communicating between the client and global models to improve the efficiency of the learning process. We could also experiment with more kinds of attribute skew, such as noise-based feature imbalance, where sub-populations for clients are created by dividing the dataset into multiple parts randomly and equally, and adding different levels of Gaussian noise to those subsets to represent different feature distributions

Further research is needed to fully understand the impact of multiple skews on the performance of federated learning in real-world scenarios. In the current study, the effects of individual skews and non-IID distributions were not analyzed by combining the skews.

References

[1] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78,

Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-1909. URL https://www.aclweb.org/anthology/W19-1909.

[2] Theodora Brisimi, Ruidi Chen, Theofanie Mela, Alex Olshevsky, Ioannis Paschalidis, and Wei Shi. Federated learning of predictive models from federated electronic health records. *International Journal of Medical Informatics*, 112, 01 2018. doi: 10.1016/j.ijmedinf.2018.01.007.

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/chen20j.html.

[5] Shiqing Liu Yaochu Jin Hangyu Zhu, Jinjin Xu. Federated learning on non-iid data: A survey, 6 2021. URL https://doi.org/10.48550/arXiv.2106.06843.

[6] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[7] Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3942–3951, 2021.

[8] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn L. Ball, Katie S. Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *CoRR*, abs/1901.07031, 2019. URL http://arxiv.org/abs/1901.07031.

[9] Alistair Johnson, Tom Pollard, Seth Berkowitz, Nathaniel Greenbaum, Matthew Lungren, Chih-ying Deng, Roger Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6: 317, 12 2019. doi: 10.1038/s41597-019-0322-0.

[10] Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. MIMIC-CXR: A large publicly available database of labeled chest radiographs. *CoRR*, abs/1901.07042, 2019. URL http://arxiv.org/abs/1901.07042.

[11] Ce Ju, Dashan Gao, Ravikiran Mane, Ben Tan, Yang Liu, and Cuntai Guan. Federated transfer learning for EEG signal classification. *CoRR*, abs/2004.12321, 2020. URL https://arxiv.org/abs/2004.12321.

[12] Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. Federated learning for vision-and-language grounding problems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11572–11579, 2020.

[13] H. Brendan McMahan, Eider Moore, Daniel Ramage, and Blaise Agüera y Arcas. Federated learning of deep networks using model averaging. *CoRR*, abs/1602.05629, 2016. URL http://arxiv.org/abs/1602.05629.

[14] Adnan Qayyum, Kashif Ahmad, Muhammad Ahtazaz Ahsan, Ala I. Al-Fuqaha, and Junaid Qadir. Collaborative federated learning for healthcare: Multi-modal COVID-19 diagnosis at the edge. *CoRR*, abs/2101.07511, 2021. URL https://arxiv.org/abs/2101.07511.

[15] Quan Chen Bingsheng He Qinbin Li, Yiqun Diao. Federated learning on non-iid data silos: An experimental study, 2 2021. URL https://doi.org/10.48550/arXiv.2102.02079.

[16] Micah Sheller, G. Reina, Brandon Edwards, Jason Martin, and Spyridon Bakas. *Multi-institutional Deep Learning Modeling Without Sharing Patient Data: A Feasibility Study on Brain Tumor Segmentation: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part I*, volume 11383, pages 92–104. 01 2019. ISBN 978-3-030-11722-1. doi: 10.1007/978-3-030-11723-8_9.

[17] Tiu, Talius, and Patel et al. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nat. Biomed. Eng (2022)*, 2022. doi: https://doi.org/10.1038/s41551-022-00936-9.

[18] Peter Hellinckx Wim Casteels. Exploiting non-i.i.d. data towards more robust machine learning algorithms, 7 2020. URL https://doi.org/10.48550/arXiv.2010.03429.

[19] Omry Yadan. Hydra - a framework for elegantly configuring complex applications. Github, 2019. URL https://github.com/facebookresearch/hydra.

[20] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D. Manning, and Curtis P. Langlotz. Contrastive learning of medical visual representations from paired images and text. *CoRR*, abs/2010.00747, 2020. URL https://arxiv.org/abs/2010.00747.

Appendix A. Extended Results

(*a*) ConVIRT + Federated Fine-Tuning (Volume Distribution) (10% Training Data)

Distribution Type	Scale	Fraction Fit - proportion of clients sampled (AUC) 0.125	0.25	0.5	1
IID	-	85.34	85.38	83.93	84.13
Non-IID	1	86.82	86.85	84.78	84.34

(*b*) ConVIRT + Federated Fine-Tuning (Volume Distribution) (1% Training Data)

Distribution Type	Scale	Fraction Fit - proportion of clients sampled (AUC) (16 Clients) 0.125	0.25	0.5	1
IID	-	85.63	85.09	84.30	84.45
Non-IID	1	85.94	83.65	83.61	82.31

(*c*) ConVIRT + Federated Fine-Tuning (Label Distribution) (10% Training Data)

Exclusive	Equal Number of Samples	Fraction Fit - proportion of clients sampled (AUC) (5 Clients) 0.25	0.5
False	False	84.96	83.43
False	True	83.25	81.79
True	False	83.84	81.71
True	True	80.13	83.44

(*d*) ConVIRT + Federated Fine-Tuning (Label Distribution) (1% Training Data)

Exclusive	Equal Number of Samples	Fraction Fit - proportion of clients sampled (AUC) (5 Clients) 0.25	0.5
False	False	84.41	85.82
False	True	81.03	83.46
True	False	80.83	83.00
True	True	77.32	79.33

Table 2: AUC Score on the Chexpert Test Dataset based on training on 1% and 10% Training Data for Label and Volume Distributions and under different values of fraction fit.