

## Title

Federated Learning for Radiology Diagnosis

## Authors

Dabbara Keshava Chowdari (B19CSE027)

Nunna Radhasyam (B19CSE060)

## Mentor

Dr. Angshuman Paul

## Abstract

Chest radiography is one of the most common medical diagnosing practices in day to day life. So the detection of abnormality must be fast and accurate. In this work, we have developed models based on densenet121<sup>[5]</sup> architecture as backbone on the NIH-Chest-Xray dataset<sup>[1]</sup> and Chexpert(Stanford chest-XRay) dataset<sup>[2]</sup> for normal vs abnormal classification. And then we merged both of the models using simple averaging<sup>[3]</sup>. We have achieved an AUC of 0.88 on the CheXpert dataset and an AUC of 0.71 on the NIH dataset. Using the averaged model we have achieved an AUC of 0.71 on the NIH dataset and an AUC of 0.85 on the CheXpert dataset.

## Introduction

In the current world there are several institutions working on a similar problem, let it be the Chest radiography. Each institution holds some data and it should be kept private. If we want to increase accuracy we need to combine the data and train the model. The institutions may not agree to this as the data should be kept private. But even if they are ready to break the privacy protocol, it will be difficult to merge datasets as each institution will have its own radiologist and each radiologist will label the X-ray differently. So to solve these issues Federated learning comes into the picture.

In FL, each client's model is trained independently. To put it another way, the model training process is customised for each client. Only learnt model parameters are delivered to a trusted centre, where they are combined and fed into the main model. Then the trusted centre returns the aggregated model to the individual clients and this process is repeated as the new data is entered. The following can be visualized in [Fig.1](#). In this context we have a simple implementation with NIH and Stanford Chest X-Ray datasets.

## Design problem formulation

For improving the accuracy of a model, centralising the data compromise privacy and also centralising the data is costly and difficult to merge the data. To overcome this problem, we aim to design a machine learning model using decentralised data.

## Possible methods to solve the problem

The tricky part in federated learning is the work going on at the global server. We aggregate the local models and make the global model which will be then sent to the local servers for use. There are several methods for this aggregation .

In them one is simple model averaging, here we directly average the weights of the local models and generate the model. This is the simplest method for implementing. But this method is not the best one as it is giving equal priority to all the local models, but in general all models are not similar some models are trained with images with good variance and some may not be trained in that way. Another method for aggregating is weighted averaging, this is similar to simple averaging but here all local models are not treated the same. We give more priority to the model that is trained with more data than the model trained with less data.

And also selecting a good backbone for training local models is also a very important factor. There are many state-of-the-art backbones such as Imagenet, Resnet, Densenet, Alexnet, Inception-v3, VGG.<sup>[4]</sup>

## Methodology adopted for the project

In this work, we have adopted densenet121 as backbone for training individual models and simple averaging for averaging the trained models.

## Work Done

### Section-1: Introduction

- We have implemented two models trained with NIH chest X-Ray dataset and CheXpert (Stanford) dataset using densenet121 architecture as the backbone in pytorch.
- In CheXpert, each image is assigned to fourteen observations out of which eleven are pathologies. If the image is found to have any of the eleven pathologies it is considered an abnormal image, otherwise it is a normal image.
- We have tried to avoid the bias by maintaining normal(31,278 images) and abnormal(32810 images) images almost equal in number. Also maintained an equal number of images for training(28,500 images) and validation(28,500 images).
- In the NIH chest X-Ray dataset, each image is assigned with fourteen pathologies and presence of any pathology classifies the image as an abnormal image. Otherwise it is considered a normal image.
- The training(43,225 images) and validation(43,275 images) dataset consists of 50,496 normal images and 36,004 abnormal images.
- The testing data(25,595) consists of 25,336 normal images and 17,939 abnormal images.
- The densenet architecture can be visualized with the following equation.

$$X_i = H_i([X_0, X_1, \dots, X_l - 1]) \text{ ----Eq.1}$$

Here  $X_i$  is the output of each layer and  $X_0$  is the input image. The final output is the concatenation of feature maps produced by each layer.  $H_i$  is the non-linear transformation function

- Along with the densenet model, we have used the Stochastic gradient descent Optimizer function, BCEWithLogitsLoss loss function and trained for 100 epochs.
- The working is shown in [Fig.1](#).

### Section-2: Federated Averaging

- We have implemented a new model by Simple averaging the individual models trained with NIH and Stanford datasets.
- For this purpose, we have used a simple averaging technique i.e. we have averaged the output weights from each model.

$$W = \frac{1}{n} \left( \sum_{k=1}^n W_k \right) \text{ ----Eq.2}$$

Here n is the total number of local models, in our case it is 2.  $W_k$  is the weight of the kth model and W is the updated weight.

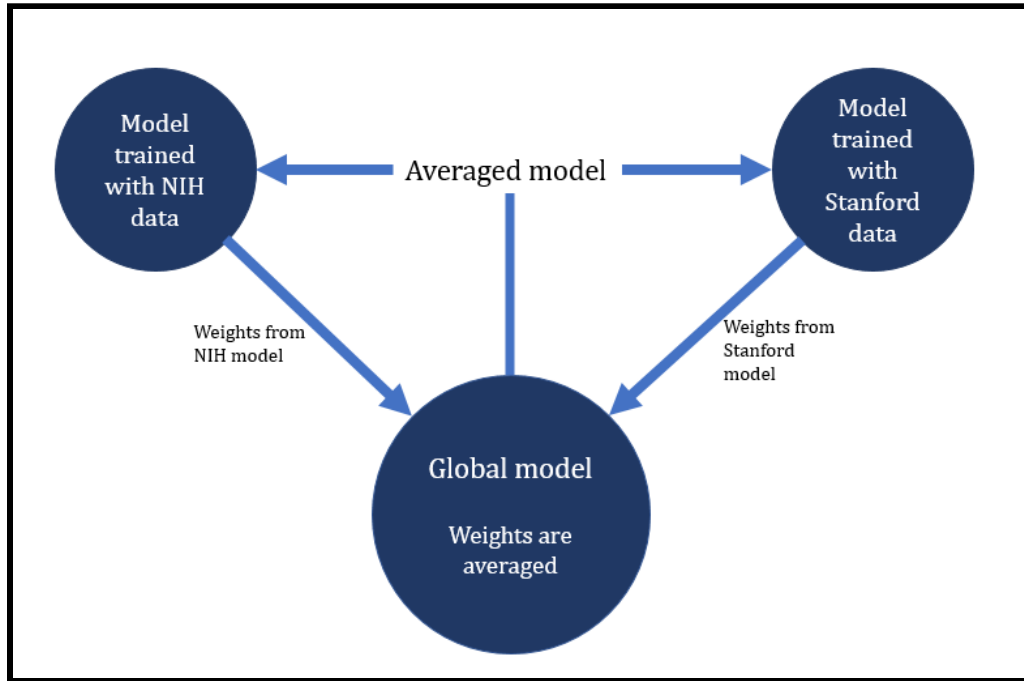


Fig. 1 Block diagram of Federated averaging

- The global server aggregates the received models from the local servers using simple model averaging and sends that model to the local servers as shown in Fig.1.

## Results and Analysis

We have tested all the three models i.e model\_nih(model trained on NIH chest X-Ray), model\_stanford(model trained on CheXpert), and model\_global(model made by simple model averaging). Table.1 shows the performance of each model when tested on NIH data(25,595 images) and when tested on CheXpert test data(7,088 images) .

Table. 1. Classification performance metrics of the three models							
Test data	Models	AUC	Sensitivity	Specificity	Precision	F1 score	Accuracy
nih test data	model_nih	0.71	0.82	0.50	0.72	0.77	0.70
	model_stanford	0.69	0.70	0.60	0.74	0.72	0.66
	model_global	0.71	0.78	0.55	0.73	0.75	0.69
stanford test data	model_nih	0.77	0.97	0.20	0.35	0.51	0.44
	model_stanford	0.88	0.77	0.83	0.67	0.72	0.81
	model_global	0.85	0.90	0.61	0.50	0.64	0.69

From the table.1 we can observe model\_nih has high accuracy when tested on NIH test data and model\_stanford has high accuracy when tested on CheXpert test data. This is because they are trained with similar data and they performed well when tested on the similar data. model\_global performed well in both the cases, it is not the best but it performed well. We can use model\_global instead of model\_nih and model\_stanford for practical purposes as it includes both the models and can perform well. We can observe the same from the below roc curves.

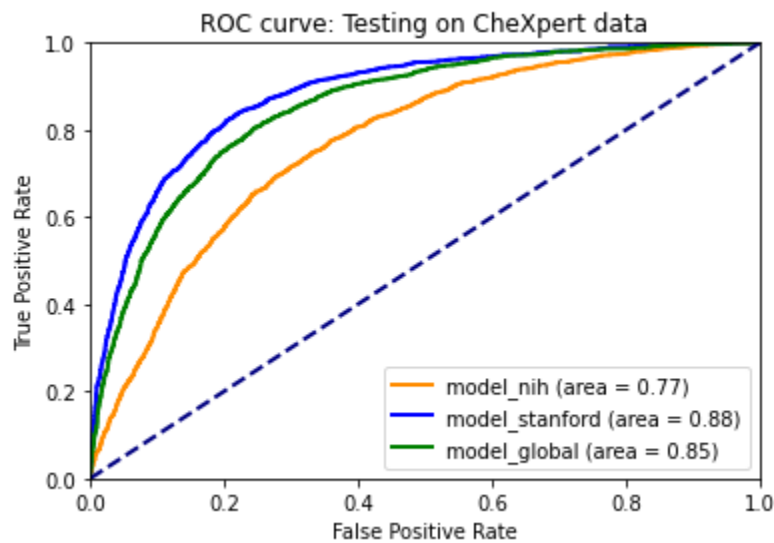


Fig.2 ROC curve for three models testing on CheXpert(Stanford) data.

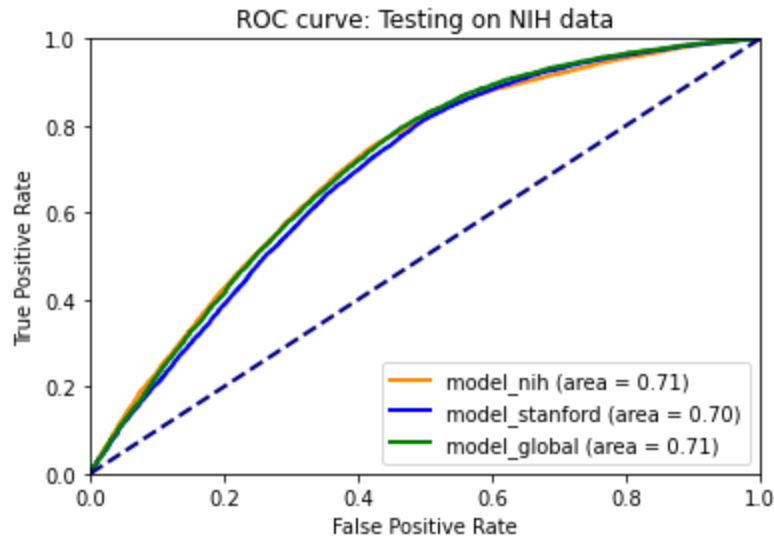


Fig.3 ROC curve for three models testing on NIH data.

On testing on Chexpert data we can see that the performance of models is in the order of  $\text{model\_stanford} > \text{model\_global} > \text{model\_nih}$ . We can observe that there is a quiet difference in the performance of the models from the roux curves shown in Fig.2.

On testing on Nih data we can see that the performance of the models is in the order of  $\text{model\_nih} > \text{model\_global} > \text{model\_stanford}$ . But from Fig.3 we can see that there is not much difference in the performance.

## Conclusion

The study describes that federated averaging improves the performance in a cost efficient way. We implemented simple model averaging and it showed us that it is better than the individual models. The averaged model has more variance than the individual models. The method of simple model averaging is not the best but it showed us some good results. Making an innovative method for aggregating the models improves the performances easily as it is cost efficient, protects the privacy and many more benefits.

## References

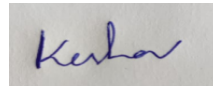
- [1] Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M. and Summers, R.M., 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2097-2106).
- [2] Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghighi, B., Ball, R., Shpanskaya, K. and Seekins, J., 2019, July. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 33, No. 01, pp. 590-597).
- [3] McMahan, B., Moore, E., Ramage, D., Hampson, S. and y Arcas, B.A., 2017, April. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics* (pp. 1273-1282). PMLR.

[4] Tang, Y.X., Tang, Y.B., Peng, Y., Yan, K., Bagheri, M., Redd, B.A., Brandon, C.J., Lu, Z., Han, M., Xiao, J. and Summers, R.M., 2020. Automated abnormality classification of chest radiographs using deep convolutional neural networks. *NPJ digital medicine*, 3(1), pp.1-8.

[5] Zhu, Y. and Newsam, S., 2017, September. Densenet for dense flow. In *2017 IEEE international conference on image processing (ICIP)* (pp. 790-794). IEEE.

We hereby declare that no part of this report has been copied from any other source and all the references are properly cited.

D. Radhakrishnan

A small, rectangular image showing a handwritten signature in blue ink. The signature appears to be 'Karan'.

Signature of the Student (B19CSE060)

Signature of the Student (B19CSE027)

Angshuman Paul

Signature of the Mentor