- **PGP AI ML – Capstone Interim Submission**

Group A CV Batch 1

Automating Pneumonia screening in chest radiographs and providing affected area details through powerful AI techniques can assist physicians to make better clinical decisions

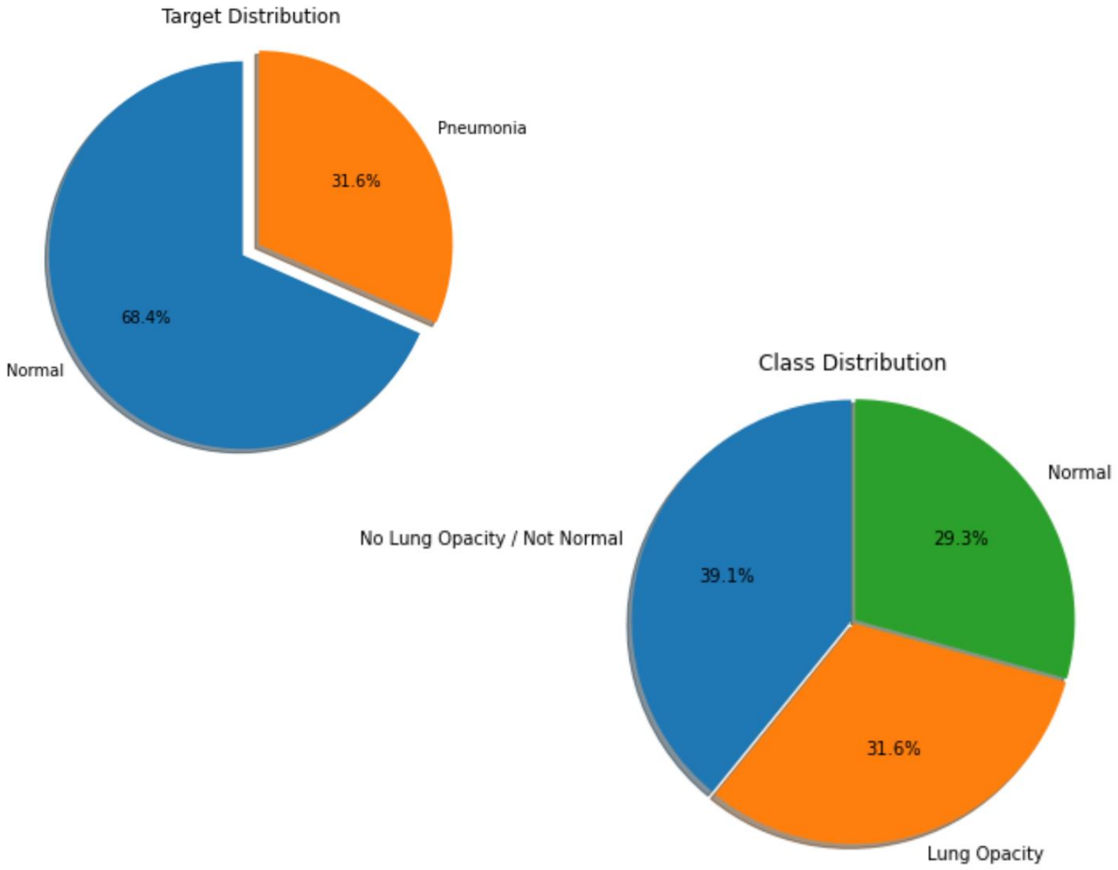
- **PROBLEM STATEMENT:** Pneumonia accounts for over 15% of all deaths of children under 5 years old internationally. Accurately diagnosing pneumonia requires review of a chest radiograph (CXR) by highly trained specialists and confirmation through clinical history, vital signs and laboratory exams. **Pneumonia usually manifests as an area or areas of increased opacity on CXR**. However, the **diagnosis of pneumonia on CXR is complicated** because of a number of other conditions in the lungs such as fluid overload (pulmonary edema), bleeding, volume loss (atelectasis or collapse), lung cancer, or post-radiation or surgical changes. A number of factors such as positioning of the patient and depth of inspiration can alter the appearance of the CXR, complicating interpretation further. Tissues with sparse material, such as lungs which are full of air, do not absorb the X-rays and appear black in the image. Dense tissues such as bones absorb X-rays and appear white in the image

- **OBJECTIVE:** We need to ***build a Pneumonia detection model*** to detect a visual signal for pneumonia in medical images. Specifically, the algorithm needs to automatically locate lung opacities on chest radiographs, providing affected area details through bounding box

- **DATA DESCRIPTION**: **Data provided is Dicom original images**: - Medical images are stored in a special format called DICOM files (*.dcm). They contain a combination of header metadata as well as underlying raw image arrays for pixel data. While we are theoretically detecting “lung opacities”, there are lung opacities that are not pneumonia related. In the data, some of these are labeled “Not Normal No Lung Opacity”. This extra third class indicates that while pneumonia was determined not to be present, there was nonetheless some type of abnormality on the image and oftentimes this finding may mimic the appearance of true pneumonia. Lung Opacity class refers to Pneumonia cases

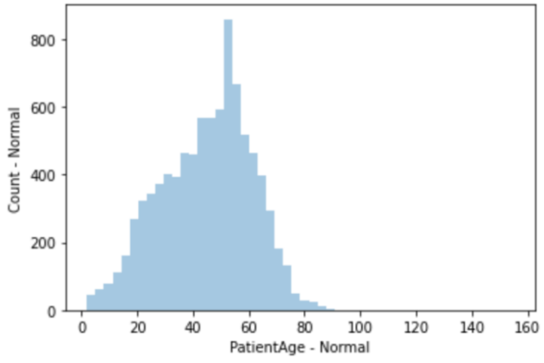
Exploratory Data Analysis

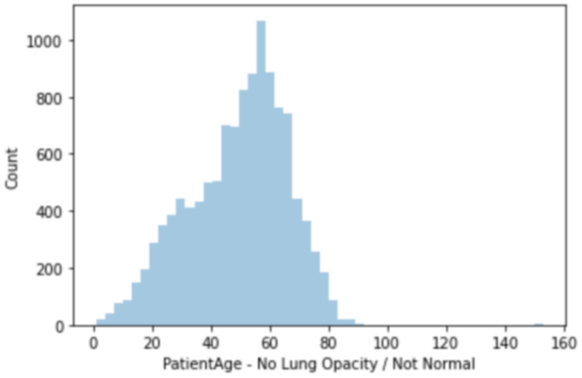
- Used pydicom library to read Dicom images
- 26684 unique patientIDs
- Few patients have multiple bounding boxes
  - 3266 patients have 2 bounding boxes defined
  - 119 patients have 3 bounding boxes defined
  - 13 patients have 4 bounding boxes defined
- We will run a binary classification to predict patients with pneumonia
- Target – 1 indicates patients with pneumonia and target – 0 indicates normal patients or patients having lung opacity but not pneumonia
- 8,851 (29.3%) patients are healthy/Normal
- 9,555 (31.6%) patients have Lung Opacity
- 11,821 (39.1%) patients have No Lung Opacity but may have other lung abnormality
- Clearly, dataset is bit imbalanced with Target class being only 31.6% of the whole dataset

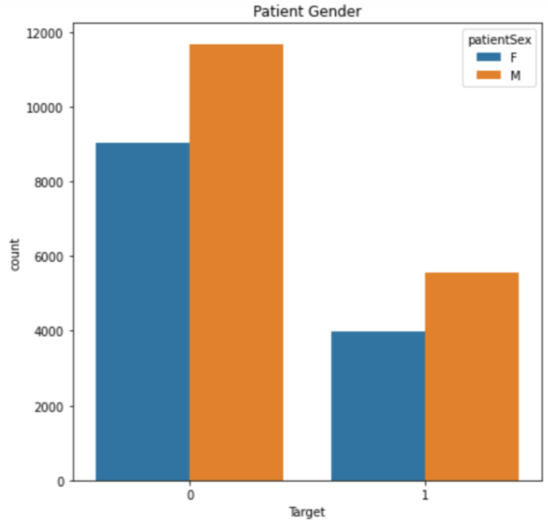


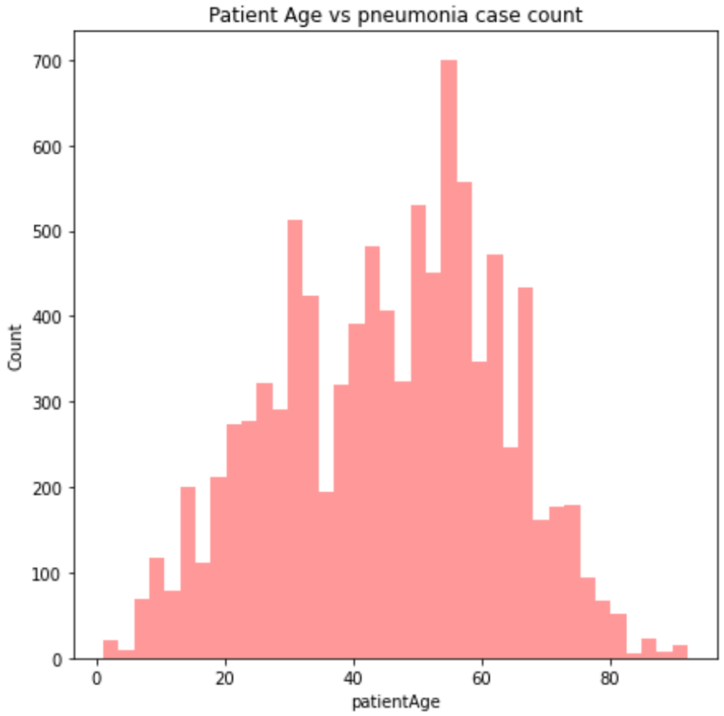
	class	Target	Patient Count
0	Lung Opacity	1	9555
1	No Lung Opacity / Not Normal	0	11821
2	Normal	0	8851

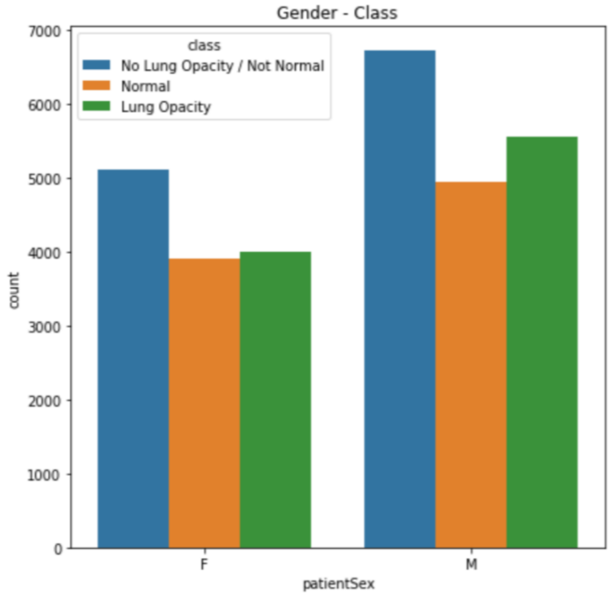
- Patient distribution shows most pneumonia patients or patients are found between age 40-60
- More records for male is observed in the data
- Proportion of pneumonia and normal patients re equal for female patients
- However, for males, proportion of pneumonia patients are more than normal

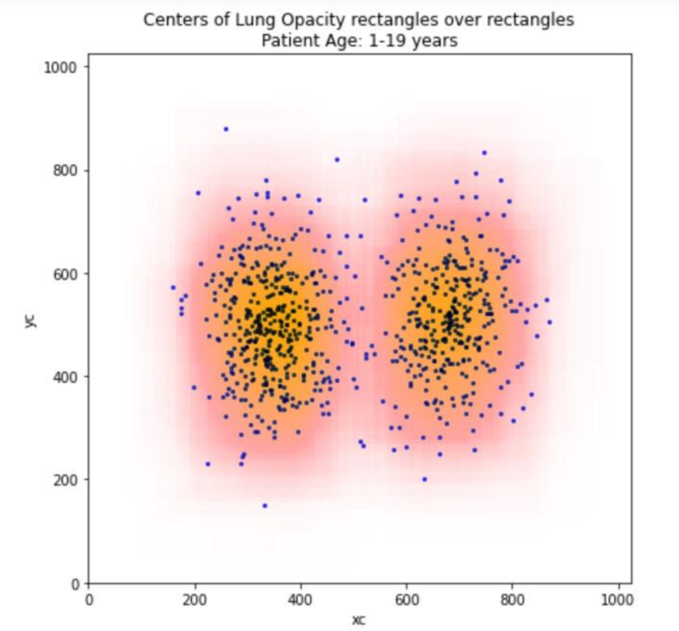


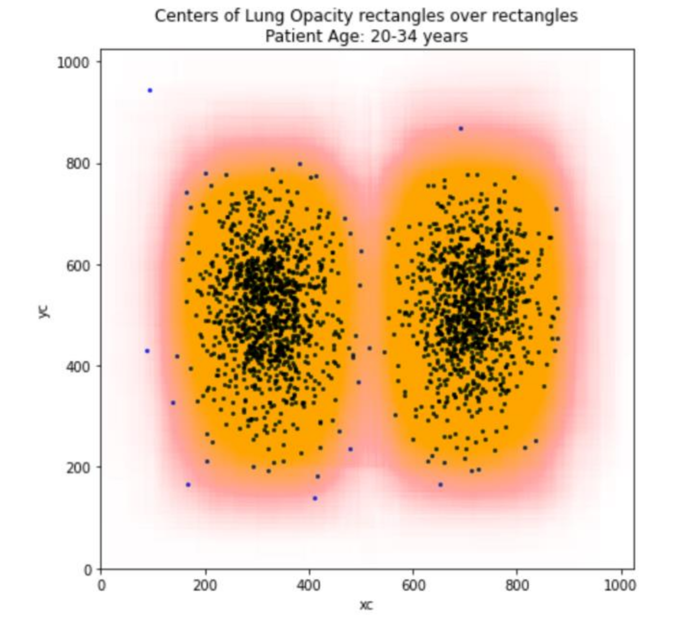


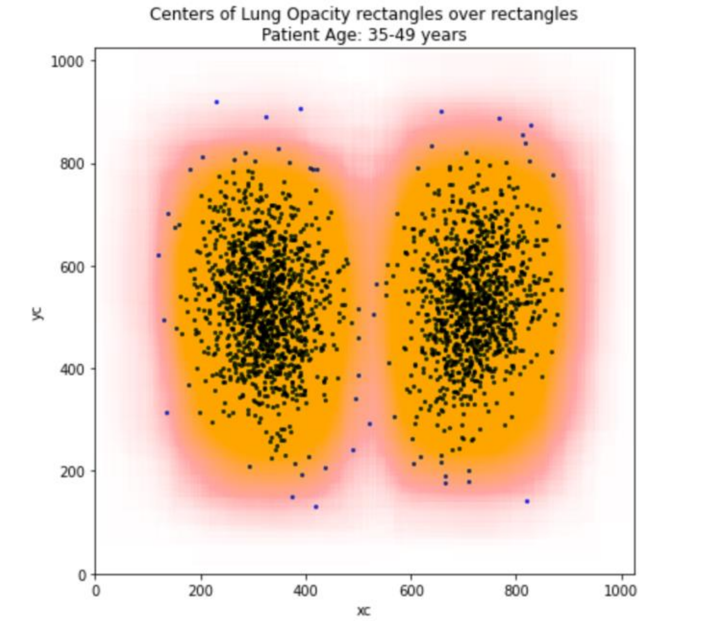


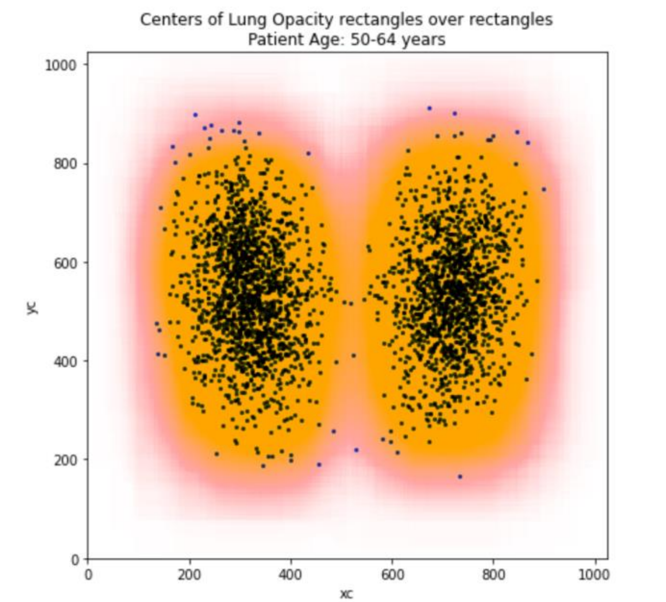


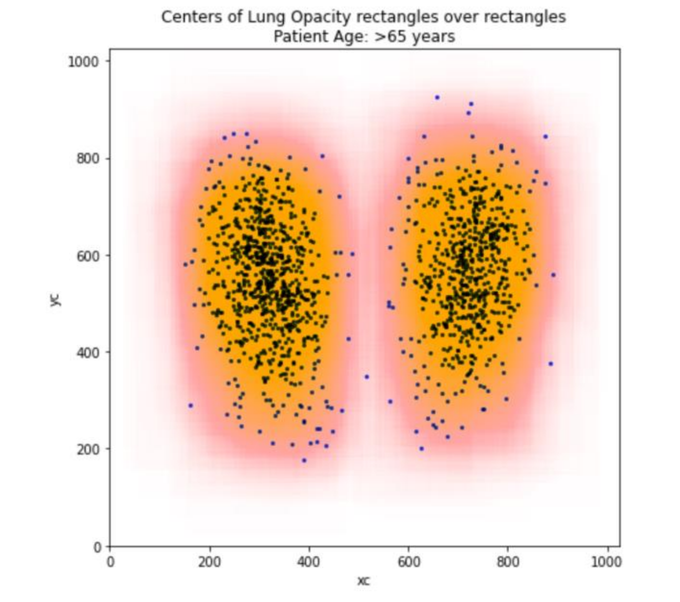






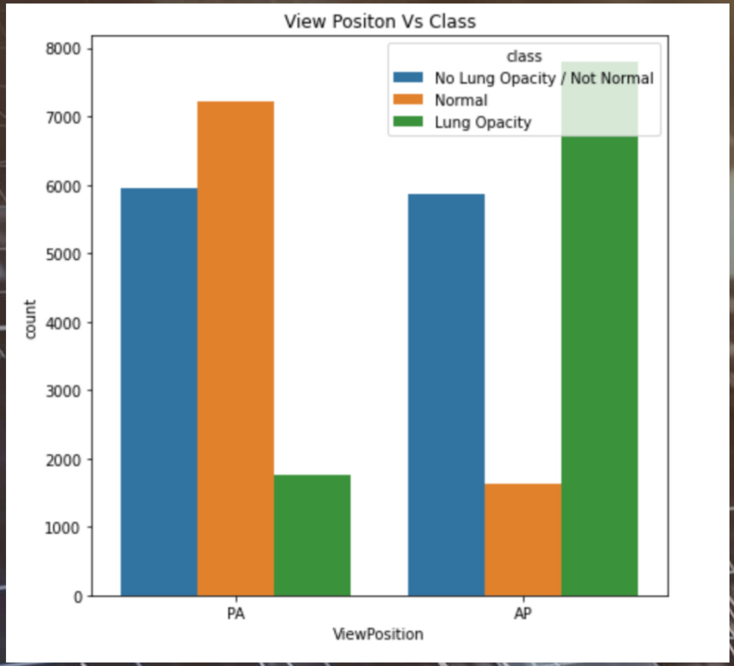


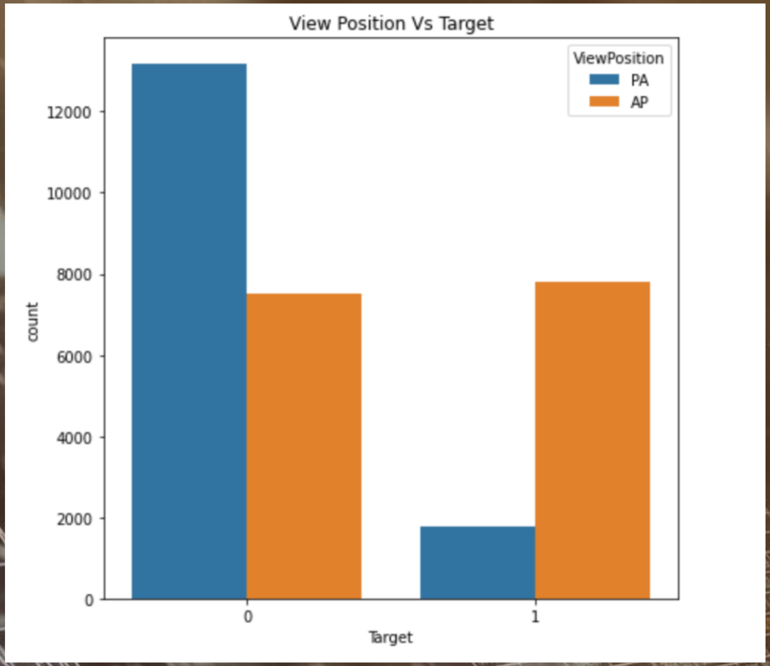


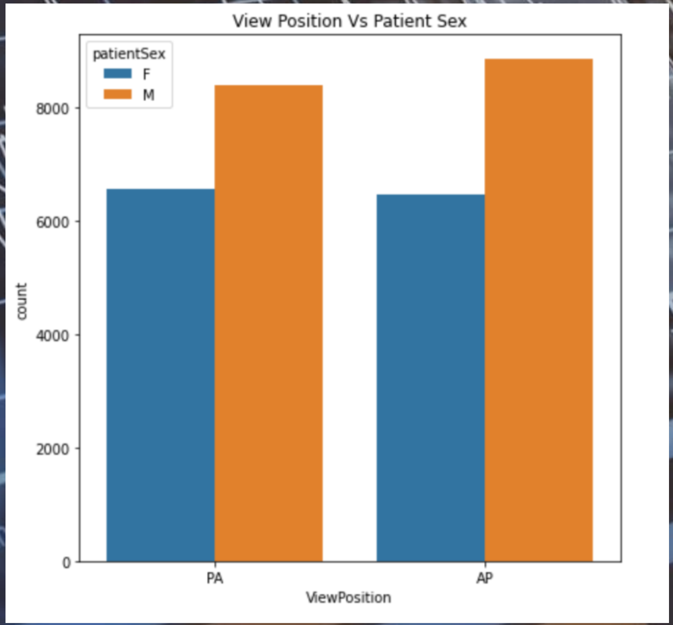


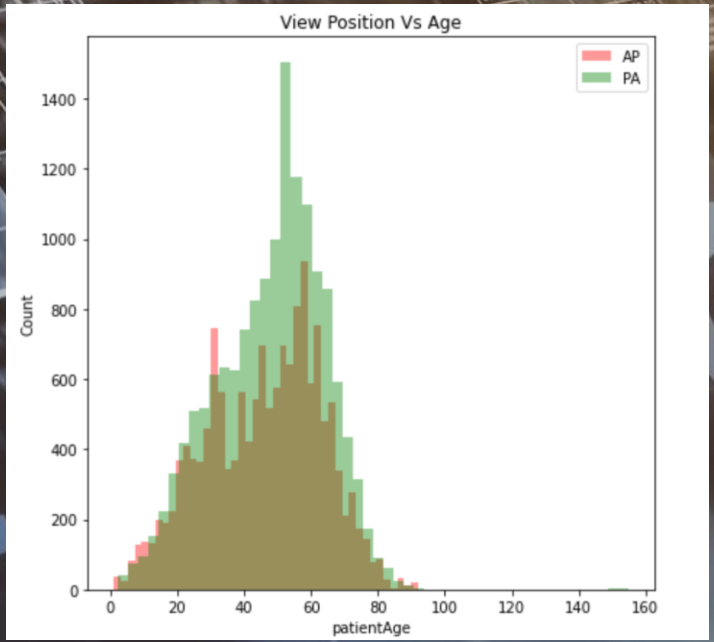
Lung opacity is seen maximum in age 50-65 followed by age group 20-50

- Chest X-ray can be done in **Posterior-Anterior (PA)** projection and **Anterior-Posterior (AP)** Projection
- Lung opacity seems to be more visible in AP position

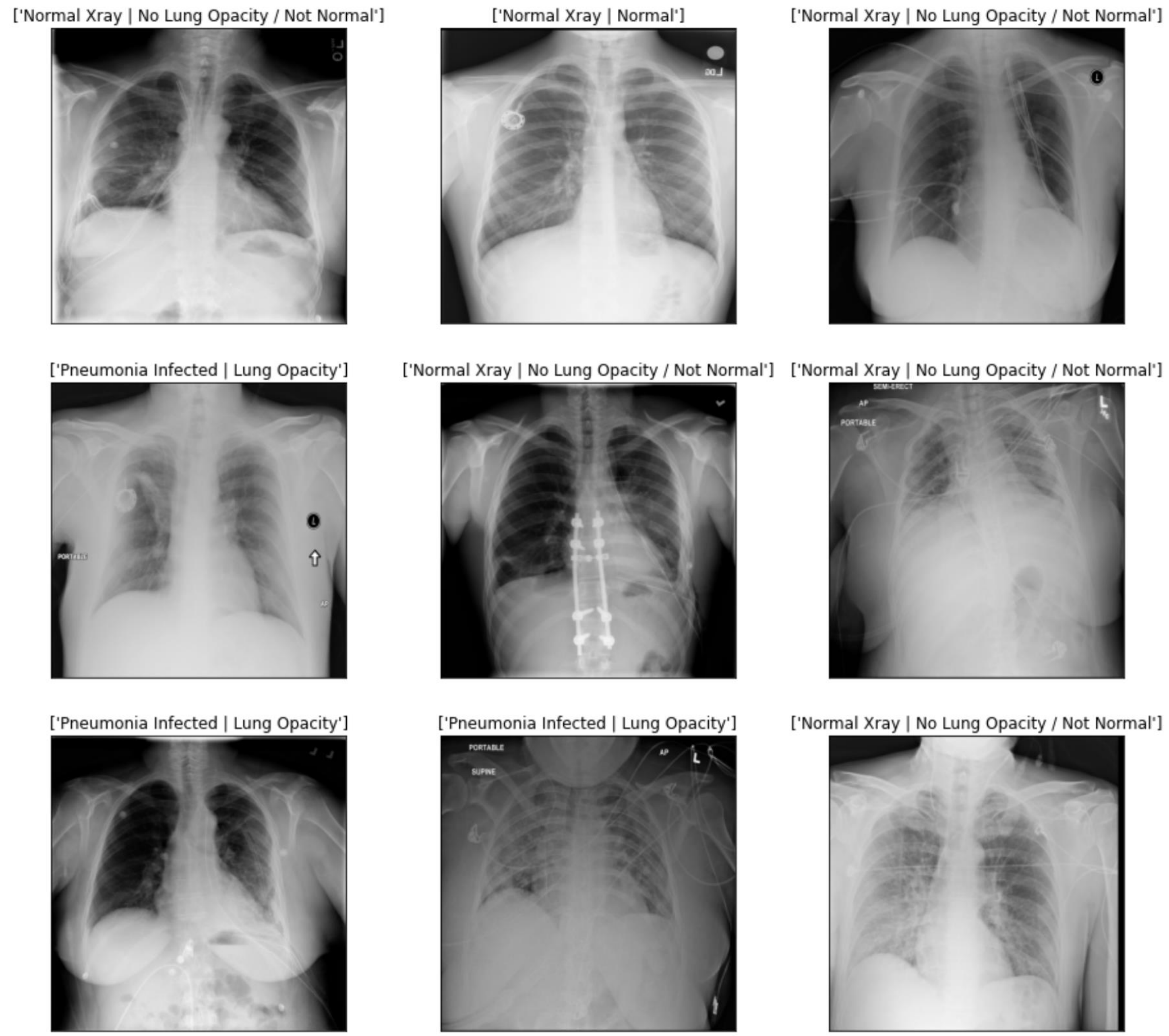




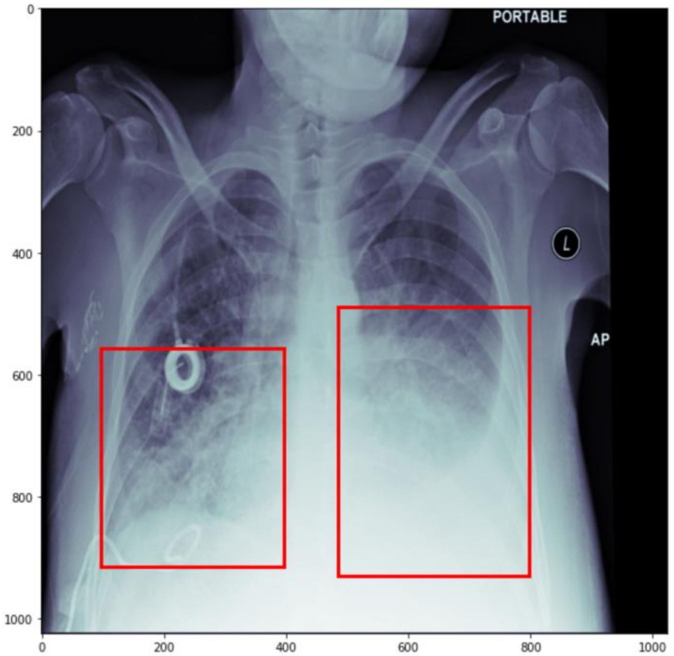


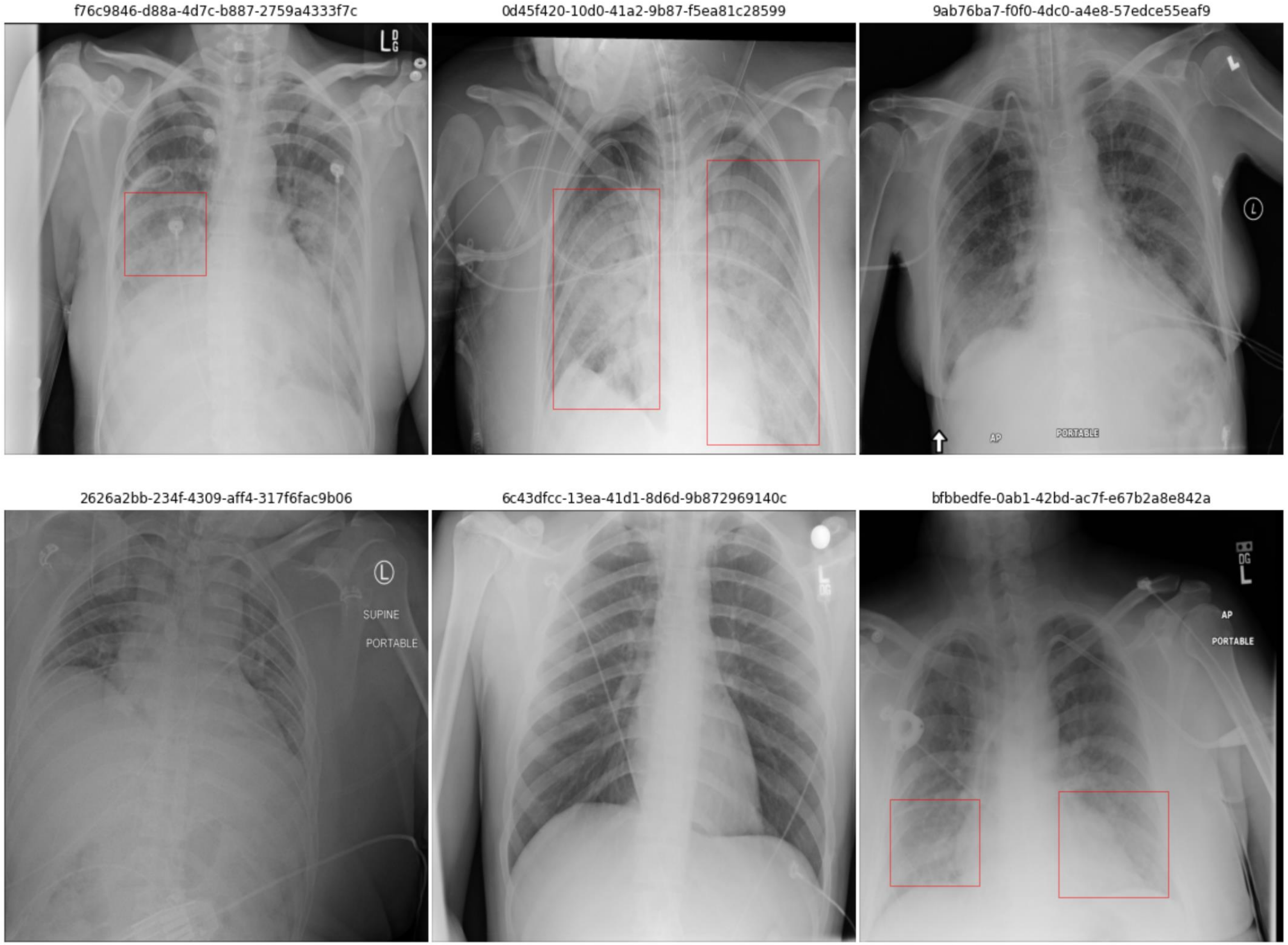


Visualization of sample X-ray images for each category

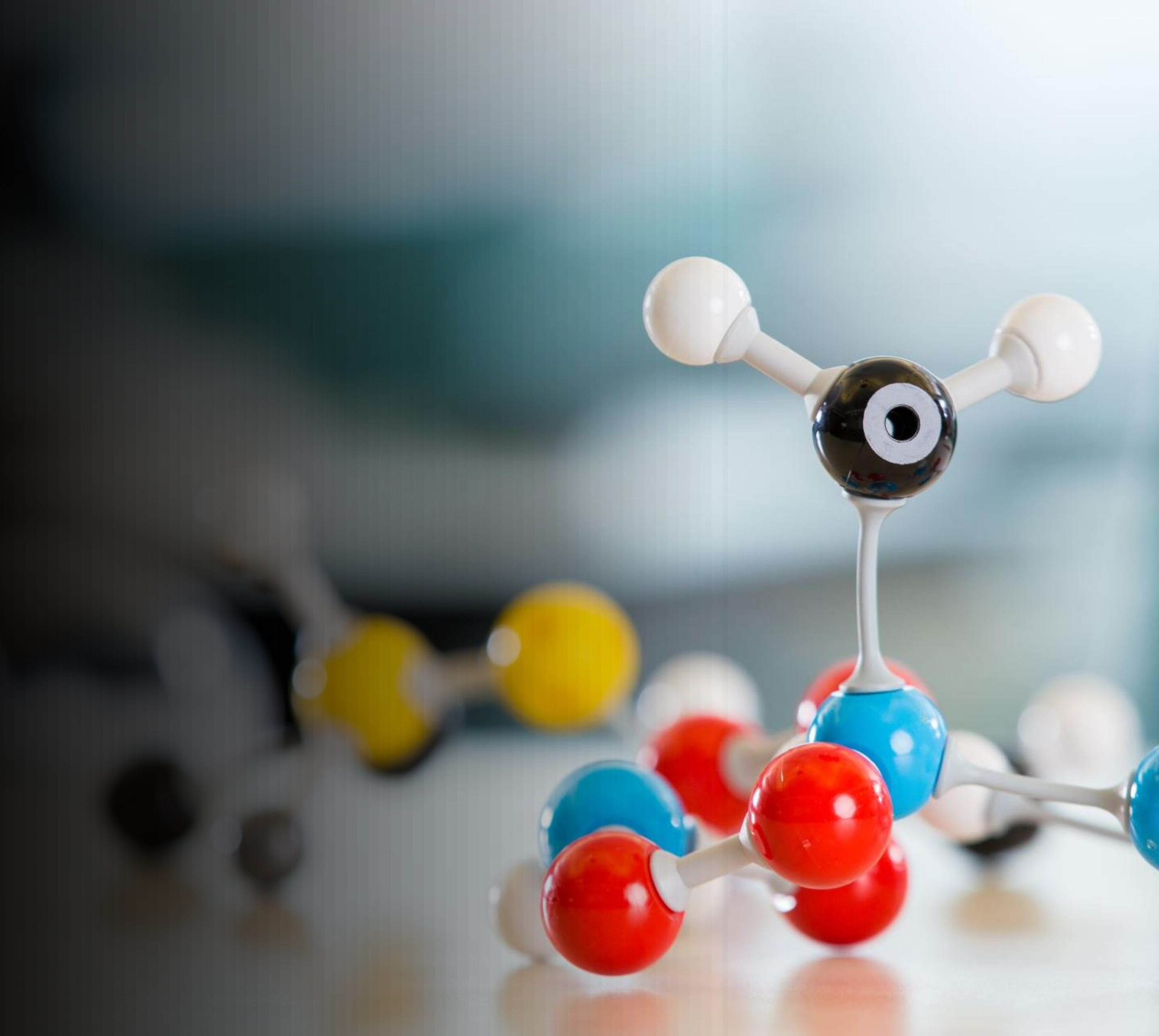


Visualization of sample X-ray images with bounding boxes depicting lung opacity





Model Building



Architectures	Accuracy	Precision	Recall	AUC-ROC
UNet	23%	23%	100%	
Faster RCNN	61%	36%	83%	
VGG19	81%	55%	93%	91.7%
VGG16	80%	50%	81%	90.7%
ResNet50	77%	55%	80%	83.2%
InceptionNet v3	86%	55%	81%	86.8%

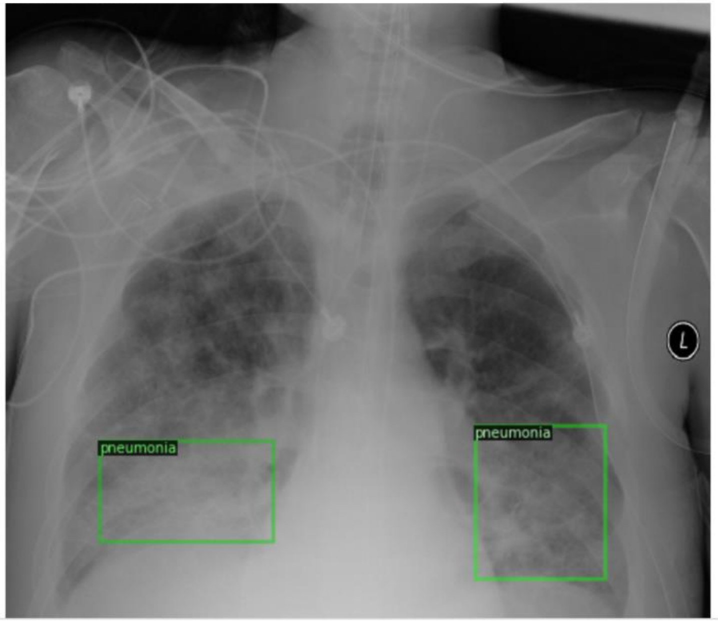
Since the problem statement requires us to read images, we use CNN and various state of the art deep learning architectures listed here to identify lung opacity with higher recall and accuracy

We have run these architectures on full data set except UNet which was processed on subset of data

From the consolidated results, **VGG19** seems to be the best model with highest recall

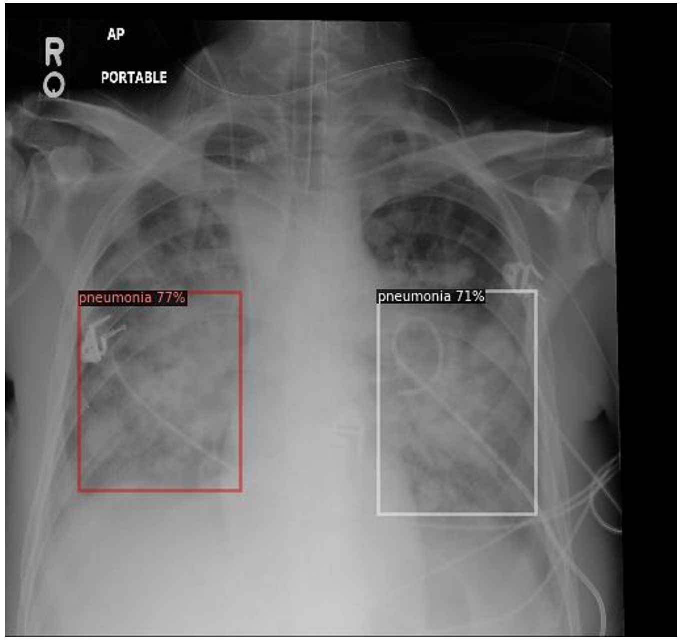
Since the aim is to detect Pneumonia cases with bounding boxes, we will proceed with Faster RCNN

GROUND TRUTH



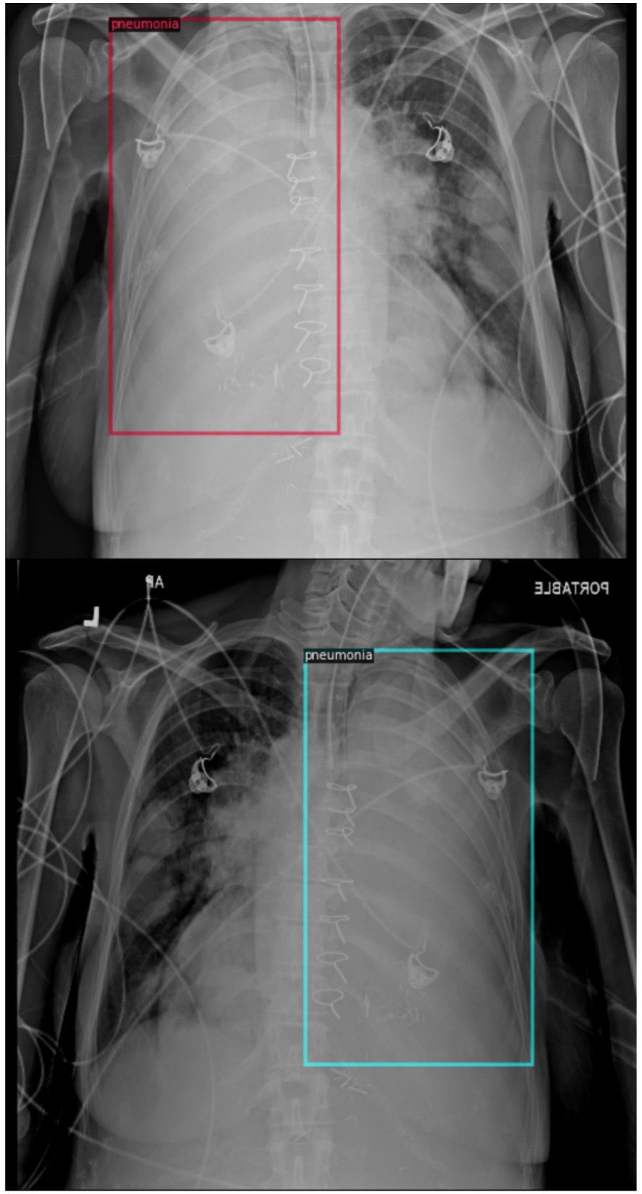
PREDICTION (Against ground truth)

PREDICTION FOR ANOTHER PATIENT



Above are few cases of “Faster RCNN” model predicting bounding boxes of pneumonia patients with confidence intervals

Data Augmentation on Faster RCNN



Faster RCNN model was developed using Detectron2 on pytorch. All dicom images are converted to JPG as Detectron2 can't directly read dicom images. We prepare the dataset into Json format and then feed into the Detectron2 model for training

Standard augmentation function for images with bounding boxes was not feasible and hence we developed a pipeline using Albumentations library from PyPi. Augmentation was performed only on 10% of dataset sampled randomly.

The X-ray picture shows that augmentation is working with **horizontal flip of image** with bounding box

Below is the outcome of our model with augmentation. **However, since recall of Faster RCNN before augmentation is better, we will proceed with it.**

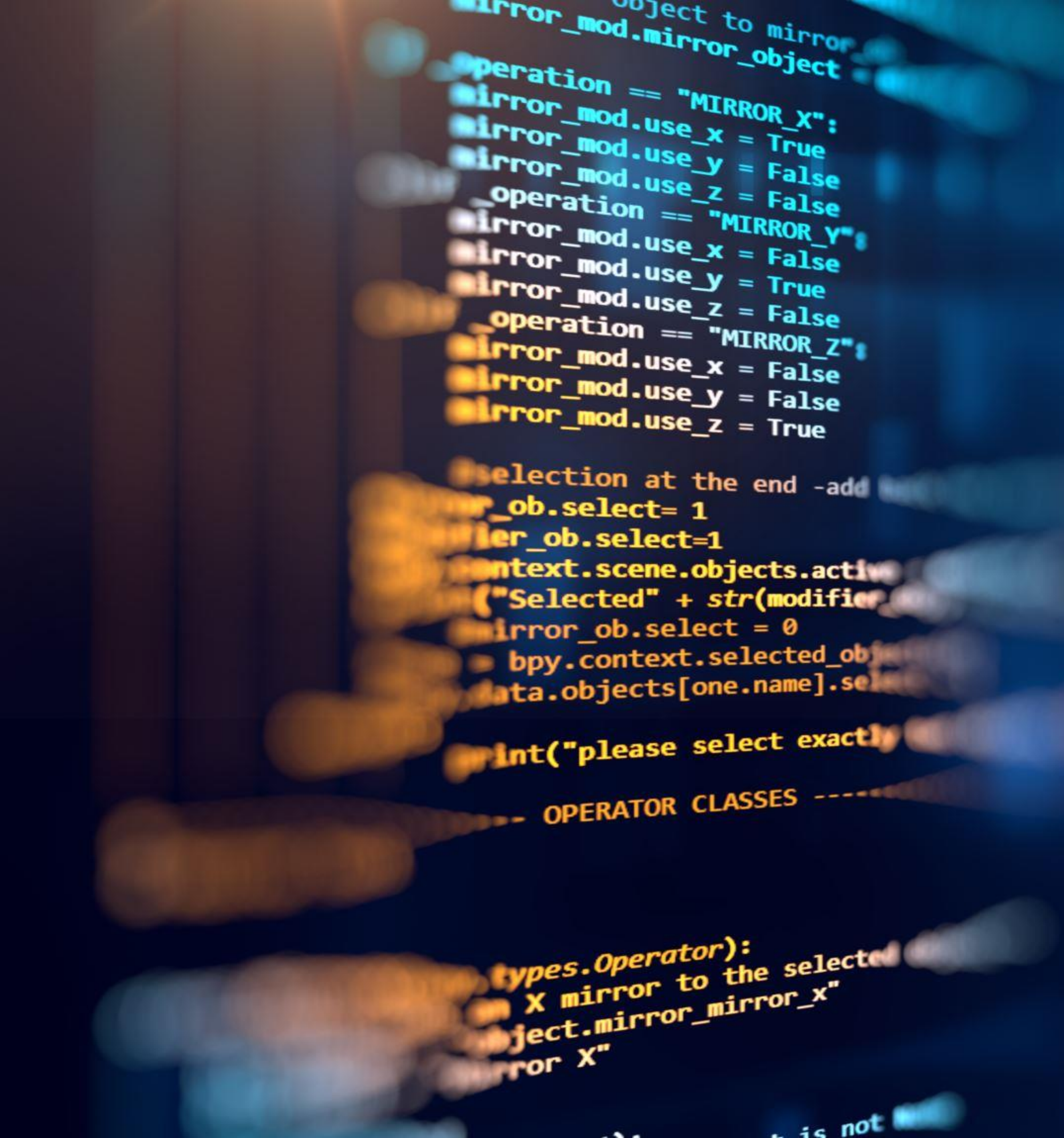
```
[71] print(classification_report(df_aug['ground_truth'], df_aug['predicted']))
     accuracy = accuracy_score(df_aug['ground_truth'], df_aug['predicted'])
     print('Model accuracy is: ', accuracy)

                   precision    recall  f1-score   support

               0       0.86      0.64      0.73      2068
               1       0.43      0.72      0.54       777

        accuracy                           0.66      2845
       macro avg       0.65      0.68      0.64      2845
    weighted avg       0.74      0.66      0.68      2845

    Model accuracy is:  0.6636203866432338
```



References – Code Base

Model/ Analysis	HTML file
EDA, VGG19, VGG16, ResNet, InceptionNet v3	Pneumonia_EDA_VGG19_VGG16_ResNet50_InceptionNet_Complete.html
UNET	Pneumonia_UNET.html
Faster RCNN using Detectron	Pnuemonia_pytorch_complete.html