FDA Submission

Your Name: Yi Wang

Name of your Device: Pneumonia detector

Algorithm Description

1. General Information

Intended Use Statement: This algorithm is designed to assist radiologists to detect pneumonia in X-ray images faster.

Indications for Use:

This algorithm is a classifier assisting radiologists to detect pneumonia in X-ray images with both male and female patients in the age of 1 to 95 years old. And besides, the following conditions must be observed:

1. Body part of x-ray must be chest.
2. Viewed position must be AP or PA
3. Modality must be DX.

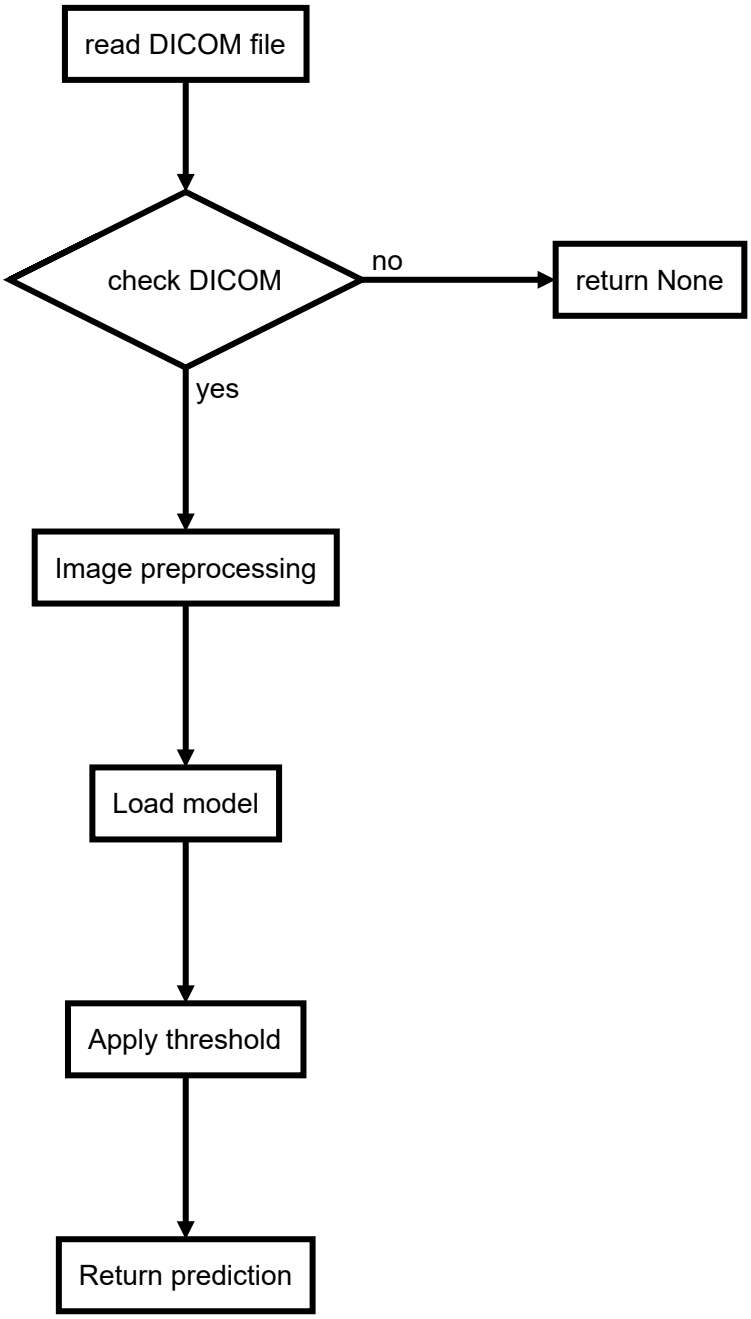
Device Limitations:
According to the EDA file, the X-ray images of patients with these diseases:

- Infiltration
- Atelectasis
- Effusion
- consolidation
  has a similar but different intensity of pneumonia, which might confuse our model prediction. So it's recommended not to use this algorithm with these diseases.

Clinical Impact of Performance:
Upon the choose of differed threshold the accuracy of positive detection and negative detection is different. In clinical, if an images is classified as negative, it might be slow down the urgency of clinicians.

This is a risk for patients who has pneumonia but is detected to health. But if a patient without pneumonia is diagnosed with pneumonia, it would not risk his/her life. So it seems that decreasing the rate of false negative cases is more important. Because of that, this algorithm choose threshold as 0.3, which can get a high recall. This means the rate of false negative is very low. So this algorithm can get a better accuracy of detecting negative cases.

2. Algorithm Design and Function



DICOM Checking Steps:

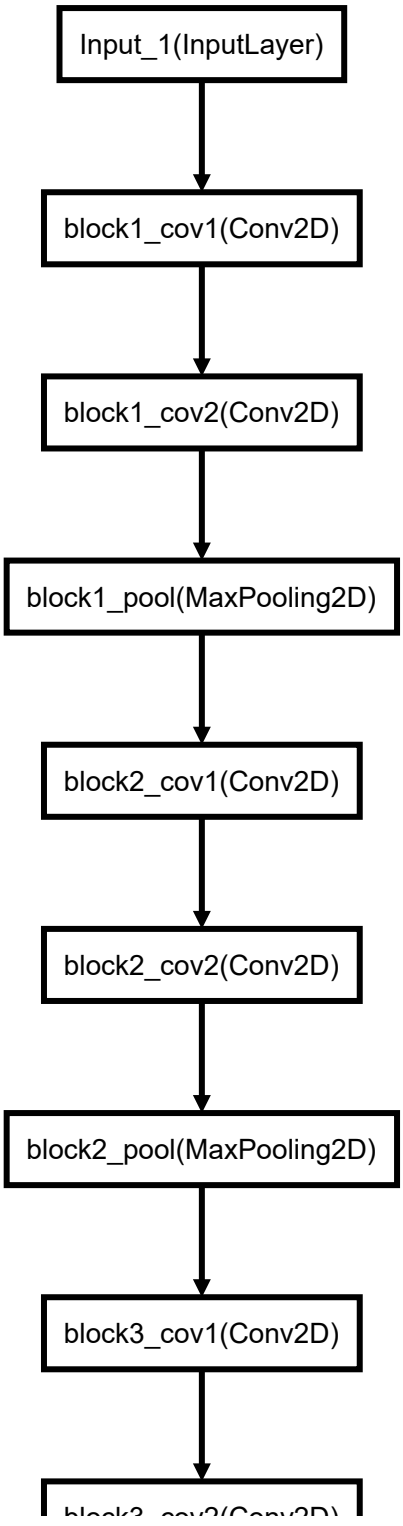
1. check if the age of patients is between 1 and 95 years old
2. check if the patient's position is AP or PA

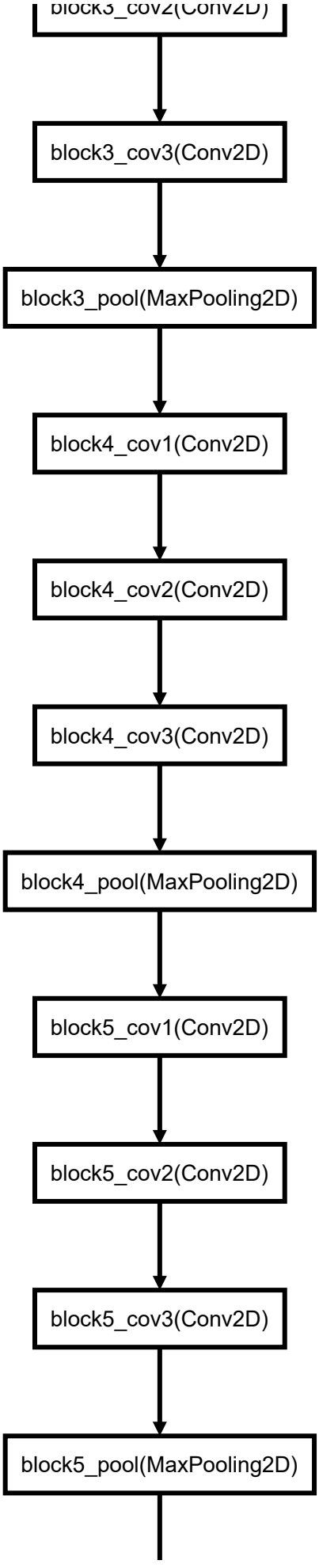
3. check if the body part is chest
4. check if modality is DX

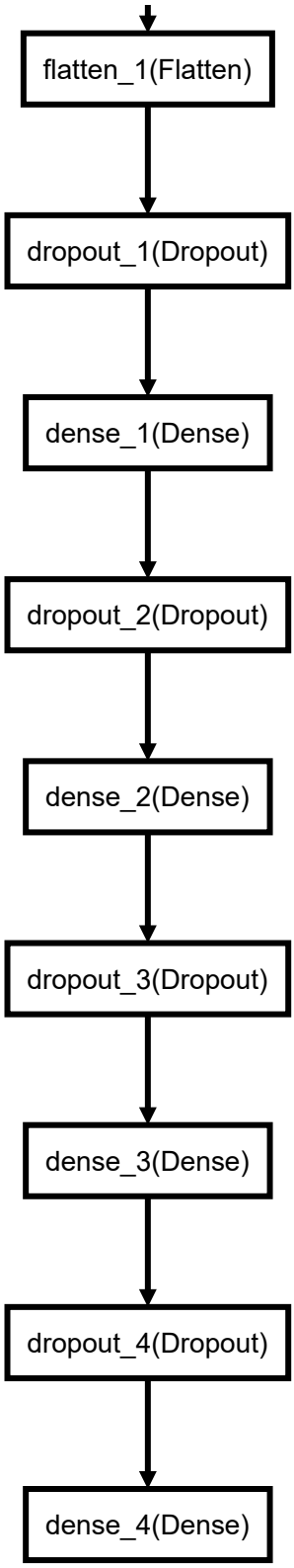
Preprocessing Steps:

The image to be process is from Check_DICOM. Because the use of VGG16 model, the input image should have image size of [1,224,224,3]. So we need to resize this image first and the normalize the intensity to be between zero and one.

CNN Architecture:




block3_cov2(Conv2D)
block3_cov3(Conv2D)
block3_pool(MaxPooling2D)
block4_cov1(Conv2D)
block4_cov2(Conv2D)
block4_cov3(Conv2D)
block4_pool(MaxPooling2D)
block5_cov1(Conv2D)
block5_cov2(Conv2D)
block5_cov3(Conv2D)
block5_pool(MaxPooling2D)



The base CNN network(from Inputlayer to block5_pool) is VGG16 pre-trained on ImageNet dataset.

3. Algorithm Training

Parameters:

- Types of augmentation used during training
  - horizontal_flip = True

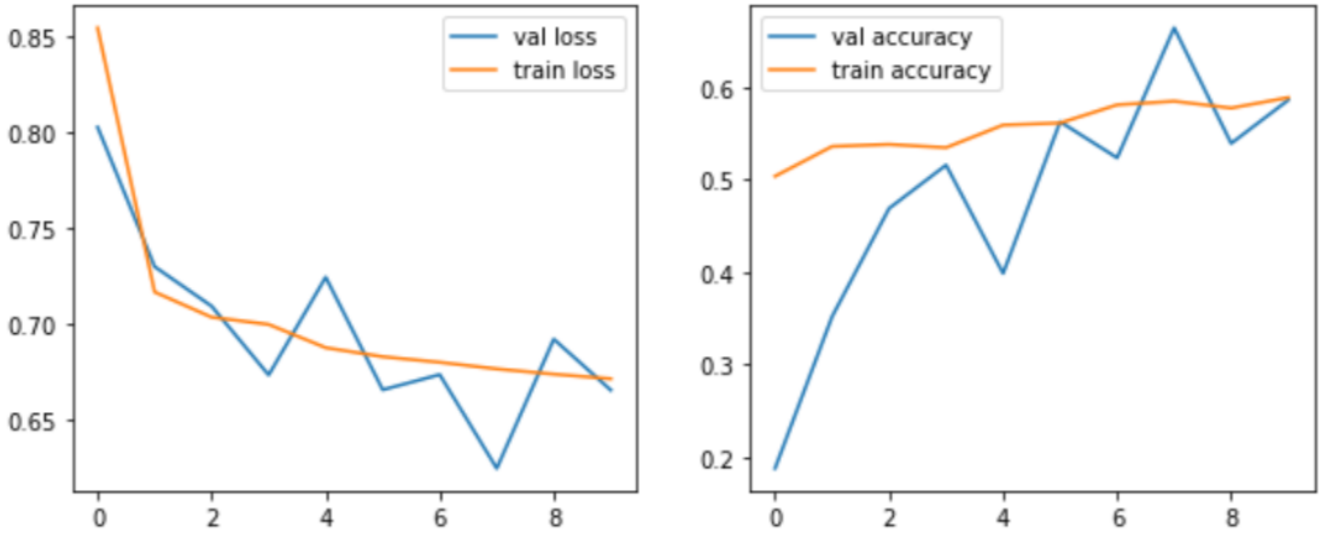
    - vertical_flip = False
    - height_shift_range = 0.1,
    - width_shift_range = 0.1,
    - rotation_range = 20,
    - shear_range = 0.1,
    - zoom_range= 0.2
- Batch size
    - training: 16
    - validation: 128
- Optimizer learning rate
    - 1e-4
- Layers of pre-existing architecture that were frozen
    - First 17 layers of VGG16

```
input_1 False
block1_conv1 False
block1_conv2 False
block1_pool False
block2_conv1 False
block2_conv2 False
block2_pool False
block3_conv1 False
block3_conv2 False
block3_conv3 False
block3_pool False
block4_conv1 False
block4_conv2 False
block4_conv3 False
block4_pool False
block5_conv1 False
block5_conv2 False
Model: "sequential_1"
```

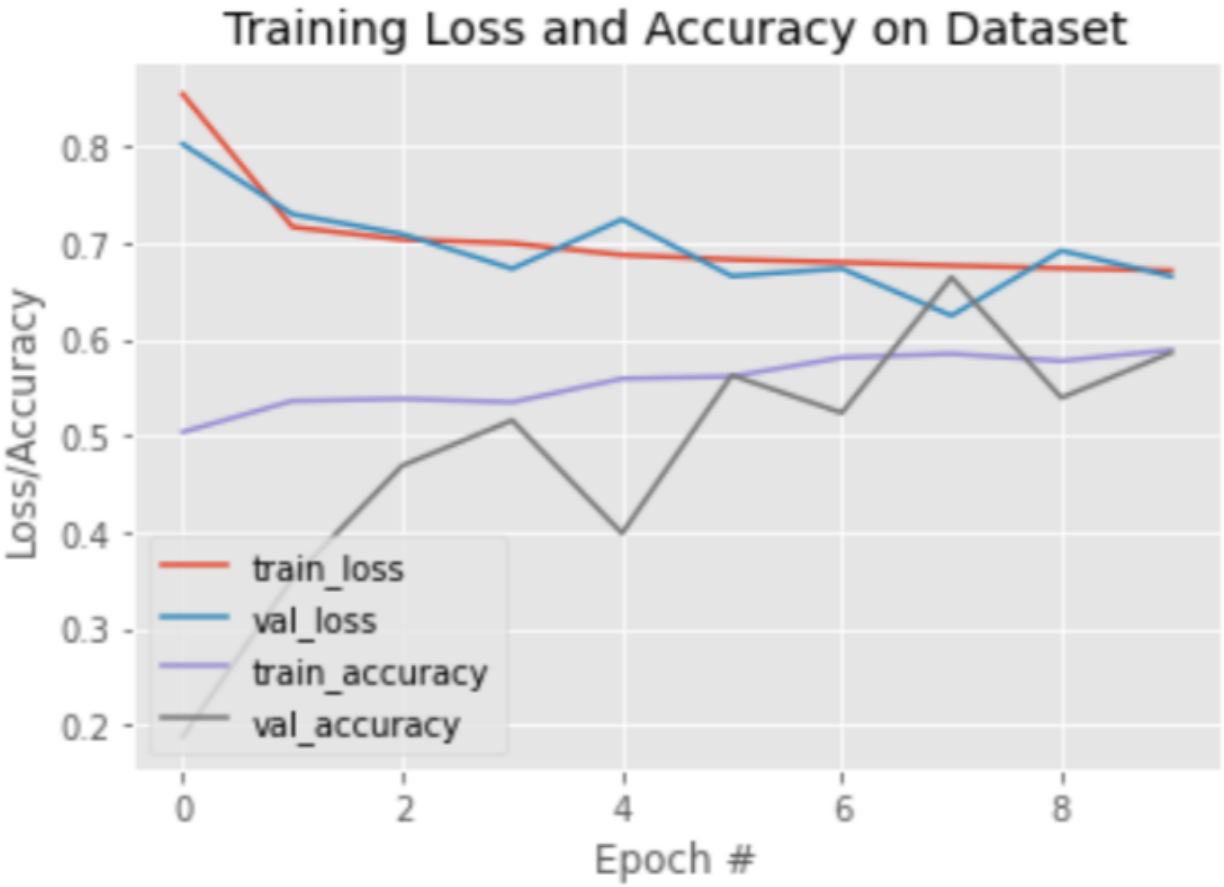
- Layers of pre-existing architecture that were fine-tuned
    - fully connected layers
- Layers added to pre-existing architecture
    - Flatten()
    - add(Dropout(0.5)
    - Dense(1024, activation='relu')
    - Dropout(0.5)
    - Dense(512, activation='relu')
    - Dropout(0.5)
    - Dense(256, activation='relu')

- Dropout(0.5)
- Dense(1, activation='sigmoid')

Algorithm training performance visualization



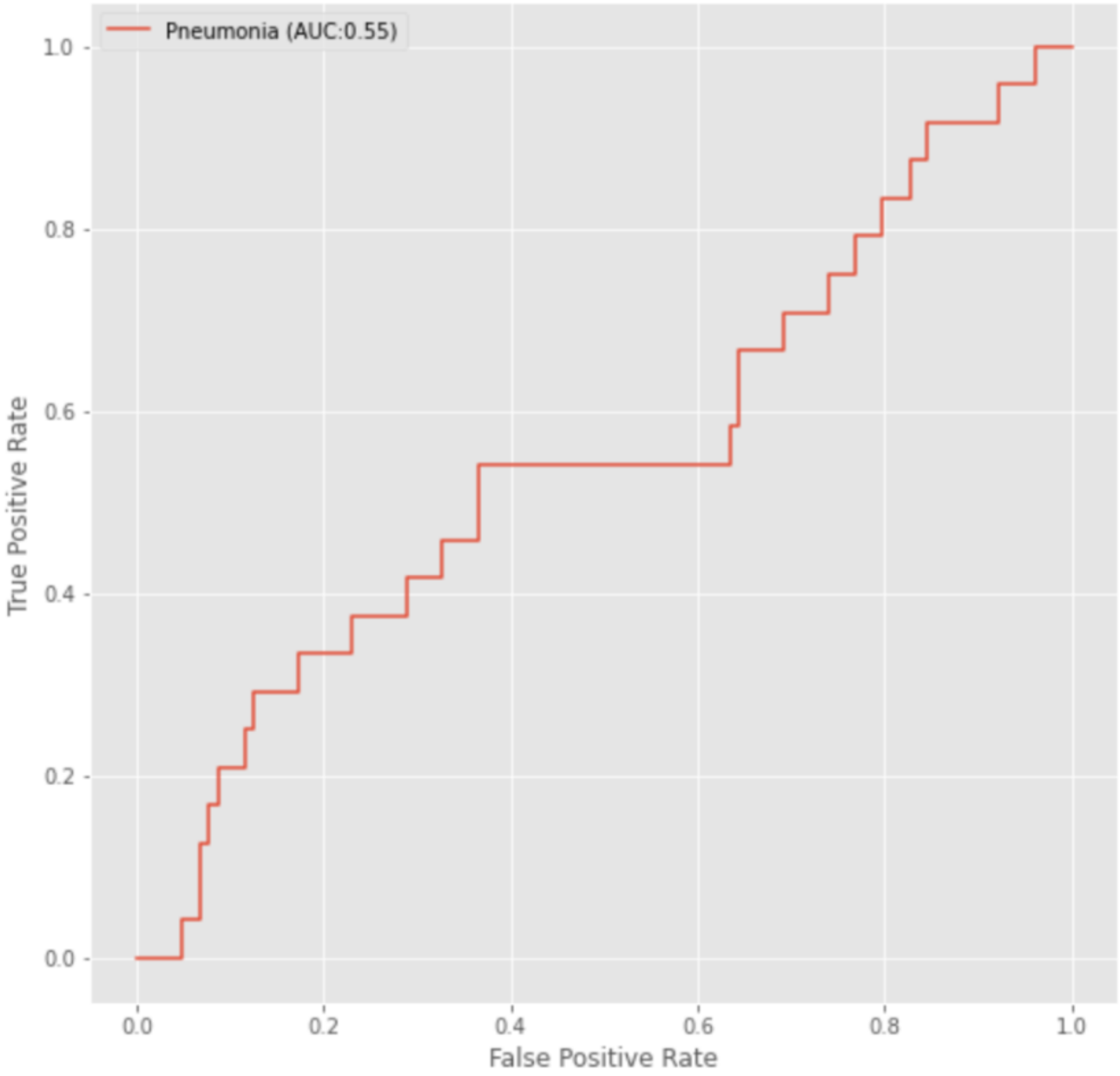
This diagram shows the development of loss and accuracy over timw for training of model. Overall, it can be seen that Loss has decreased while Accuracy increased.

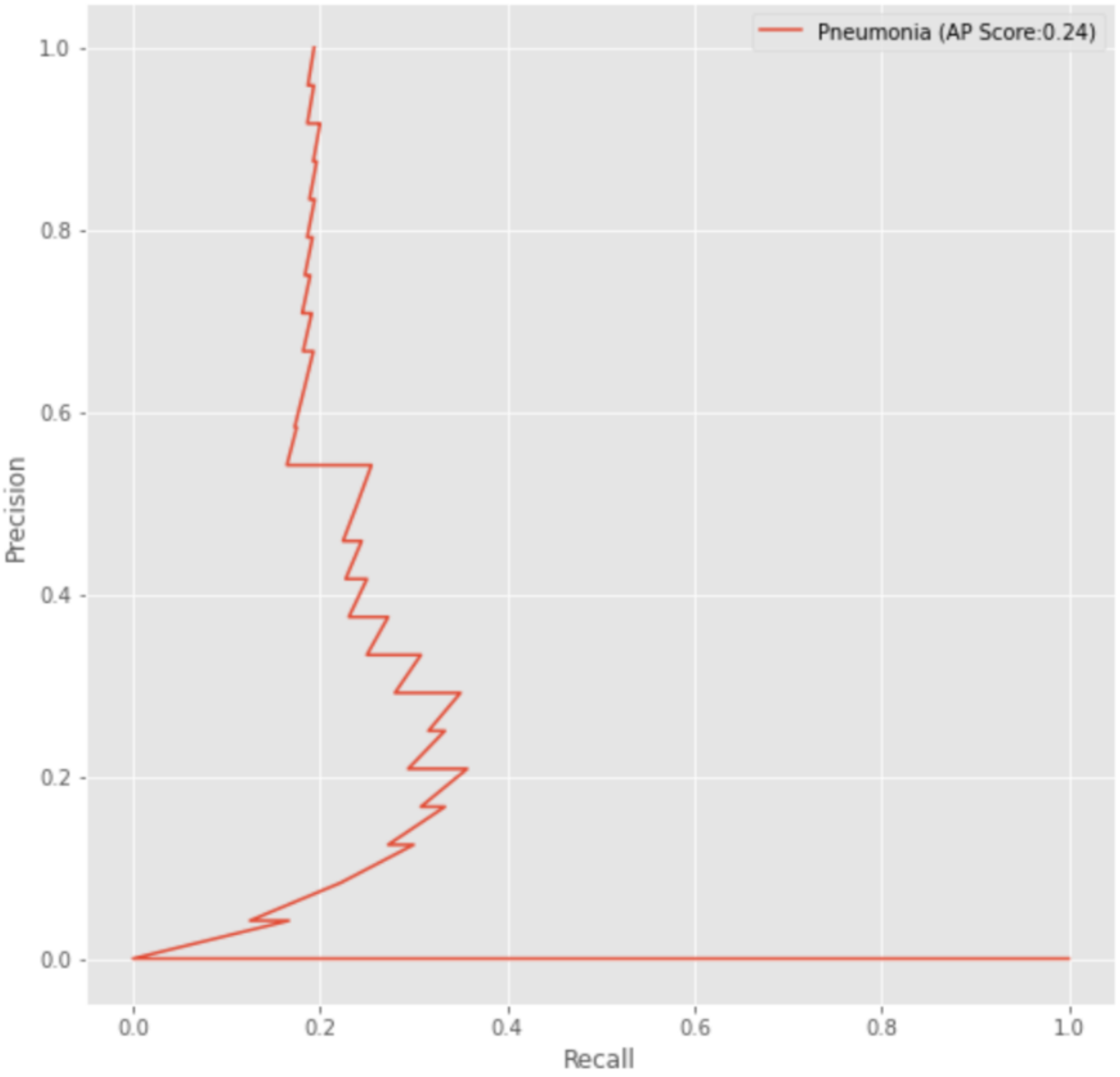


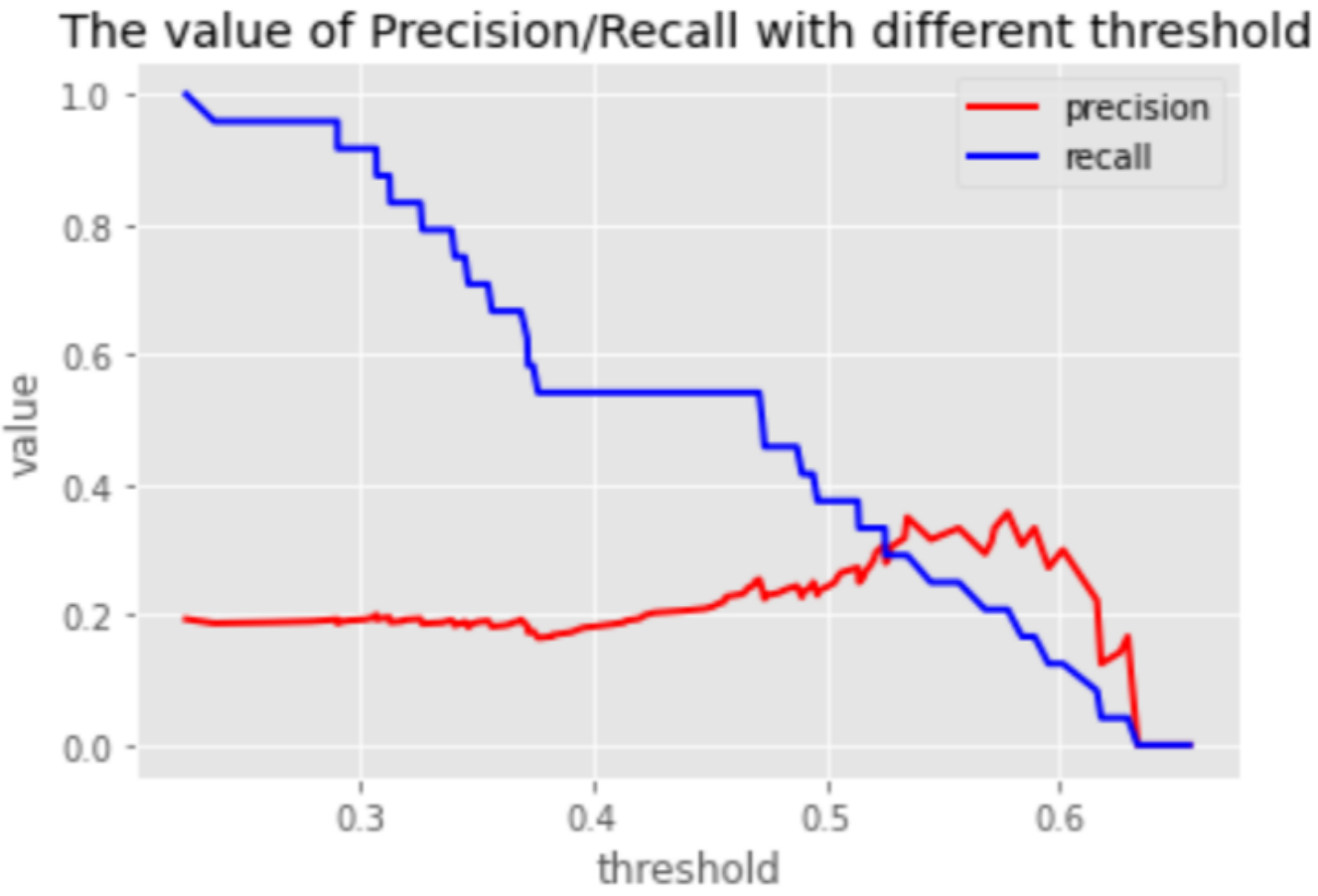
As can be seen from the above figure, although Loss is decreasing and Accuracy is rising, the value of

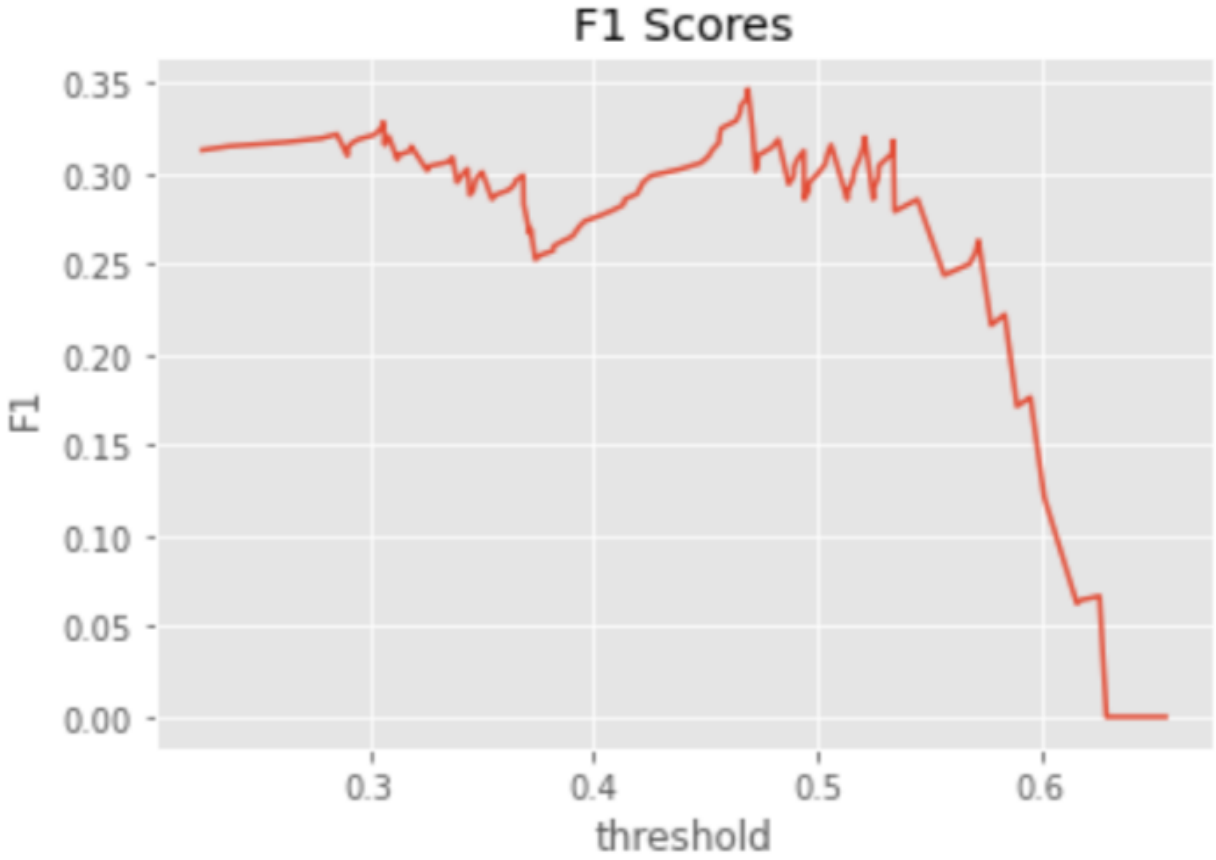
Loss is still very large and the accuracy is not good enough.

P-R curve




Pneumonia (AP Score:0.24)
Precision
Recall
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0


The value of Precision/Recall with different threshold
precision
recall
value
threshold
1.0
0.8
0.6
0.4
0.2
0.0
0.3
0.4
0.5
0.6


F1 Scores
F1
threshold
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0.3
0.4
0.5
0.6

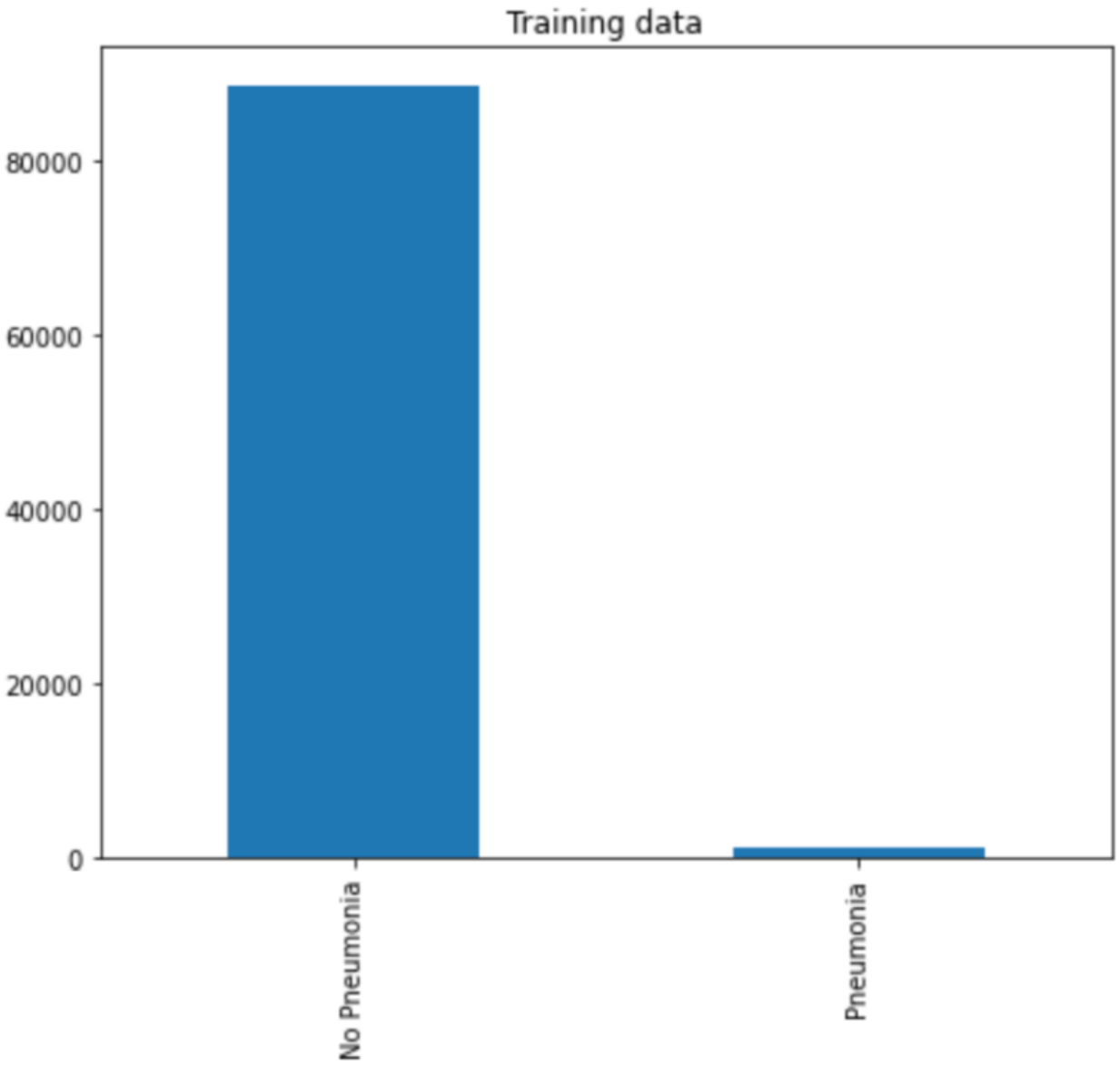
Final Threshold and Explanation:

From figures above it can be seen that the value of threshold 0.47 is optimal for F1 scores. But the precision and recall are both still not so high. As we explained in Clinical Impact of Performance, to ensure a higher accuracy of detecting negative cases has lower risk for patients. So we focus on decreasing the rate of false negative. Because of this, this algorithm choose threshold as 0.3, which can get a high recall. And the rate of false negative can be very low.

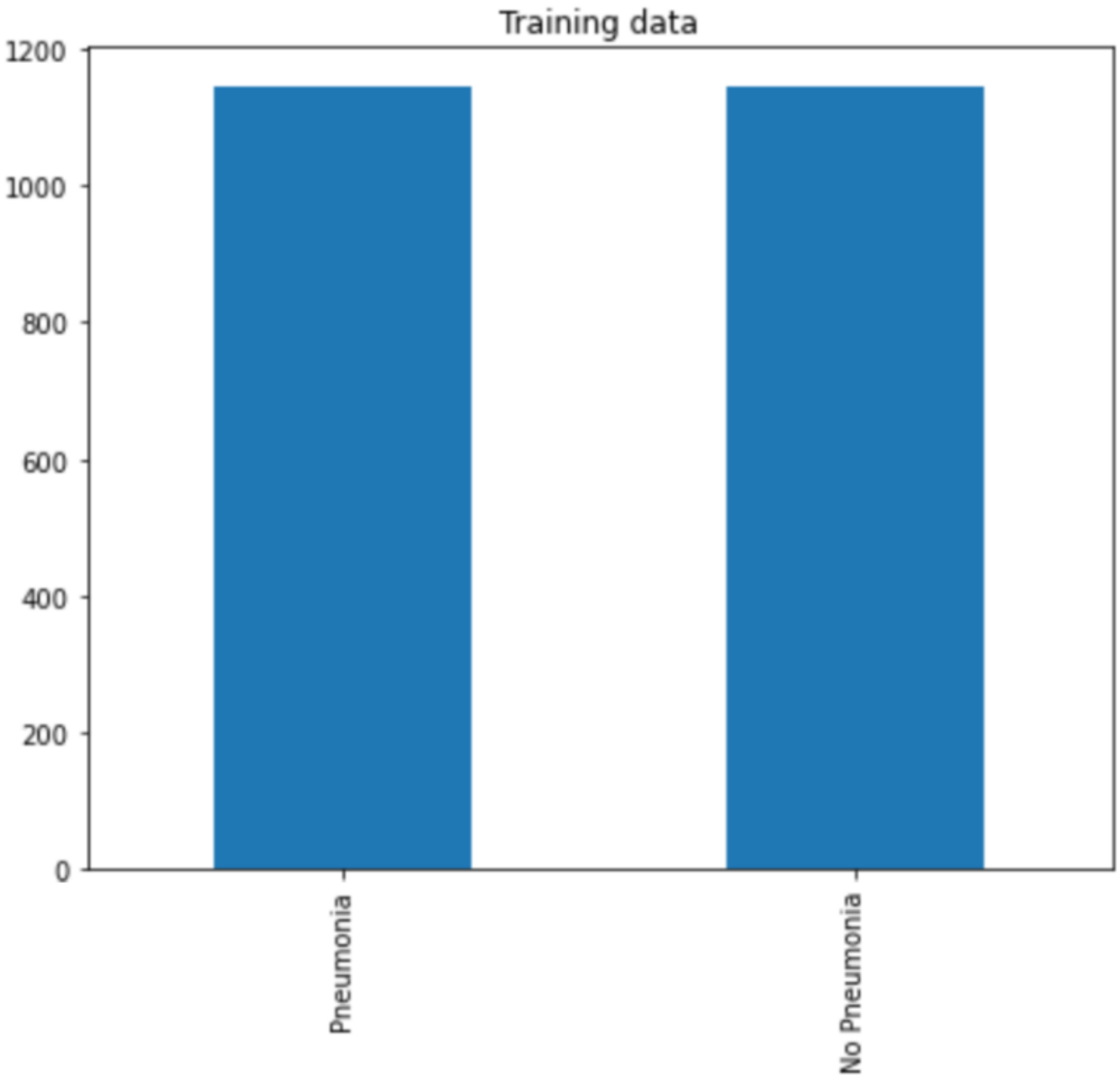
4. Databases

(For the below, include visualizations as they are useful and relevant)

Description of Training Dataset:
As the first step of training, the dataset was divided into training dataset and validation dataset. The raw training dataset can be seen as below. It obviously that a huge imbalance appears in this dataset, which means the label with 'No Pneumonia' is much more than label with 'Pneumonia'. If we use this dataset to train our model, the model will tend to guess the result instead of learning.

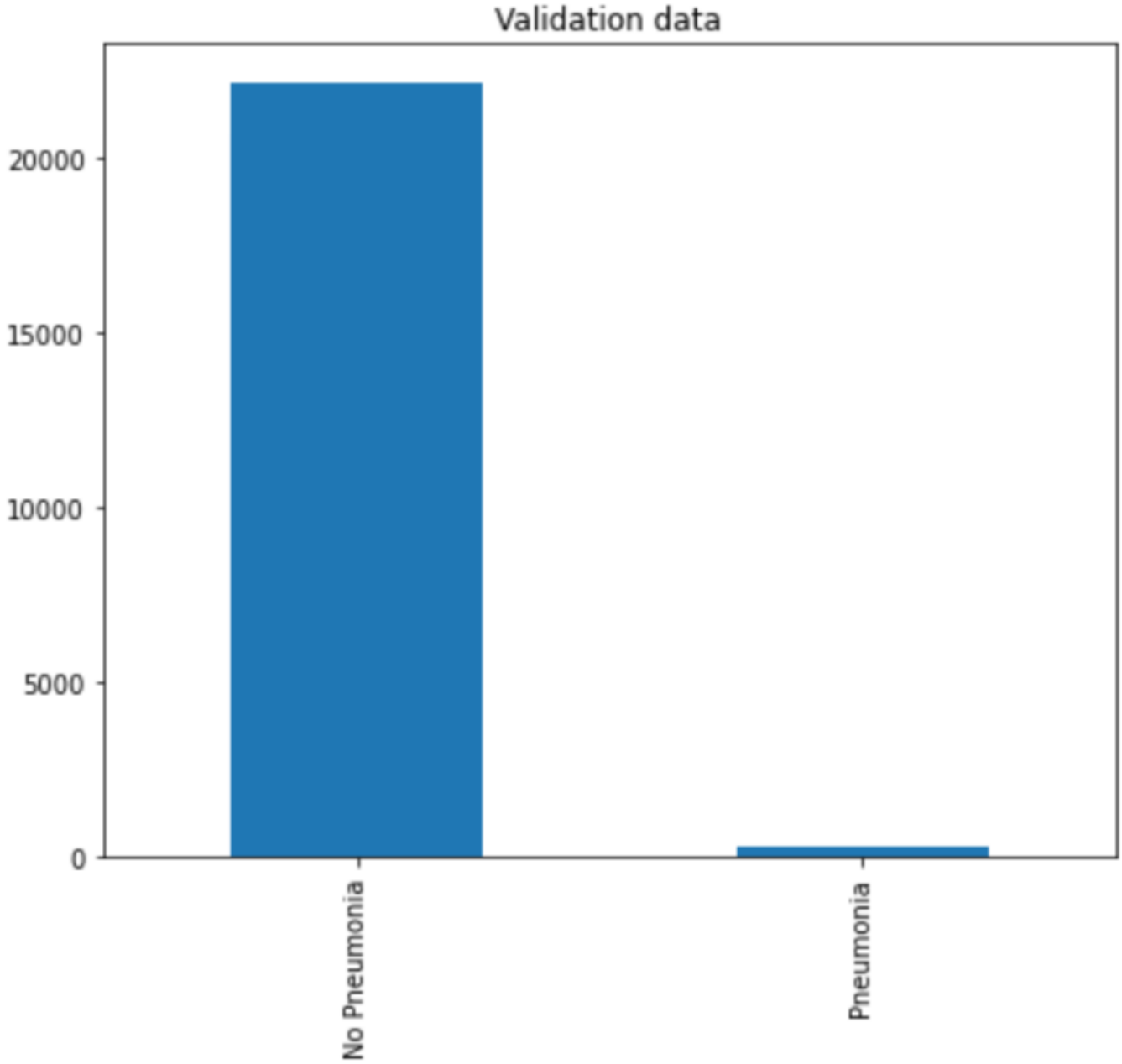


So in order to have equal amount of positive and negative cases of Pneumonia in Training a operation of downsampling is needed.

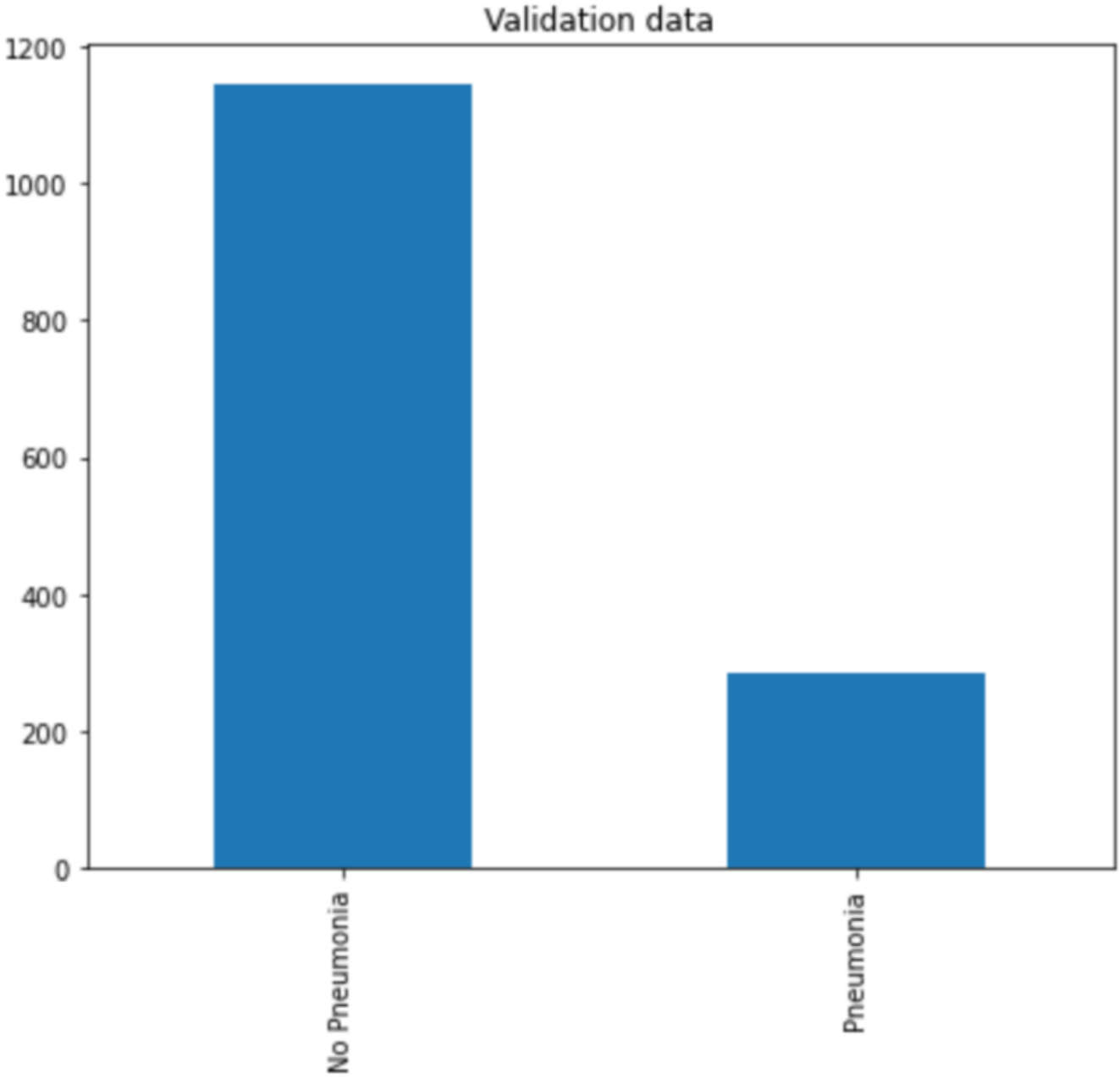


Description of Validation Dataset:

As we explained above, validation dataset also need to be resampling.



As the figure below showed, validation dataset was resampled by a similar operation of training dataset. But random sample of positive data was used to create more positive cases in prder to get 80:20 of negative and positive dataset.



5. Ground Truth

The Ground Truth are obtained from the radiologist reports of NIH Clinical Center. So this might be silver standard ground truth. These ground truth are easier for algorithm developer to get. Due to the experience of the radiologist and other potentional objective reasons, some artificial misjudgment may appear. But the probability of misjudgment can be reduced by using of voting system or weighted system. But the cost will increase as the number of radiologists increases.

6. FDA Validation Plan

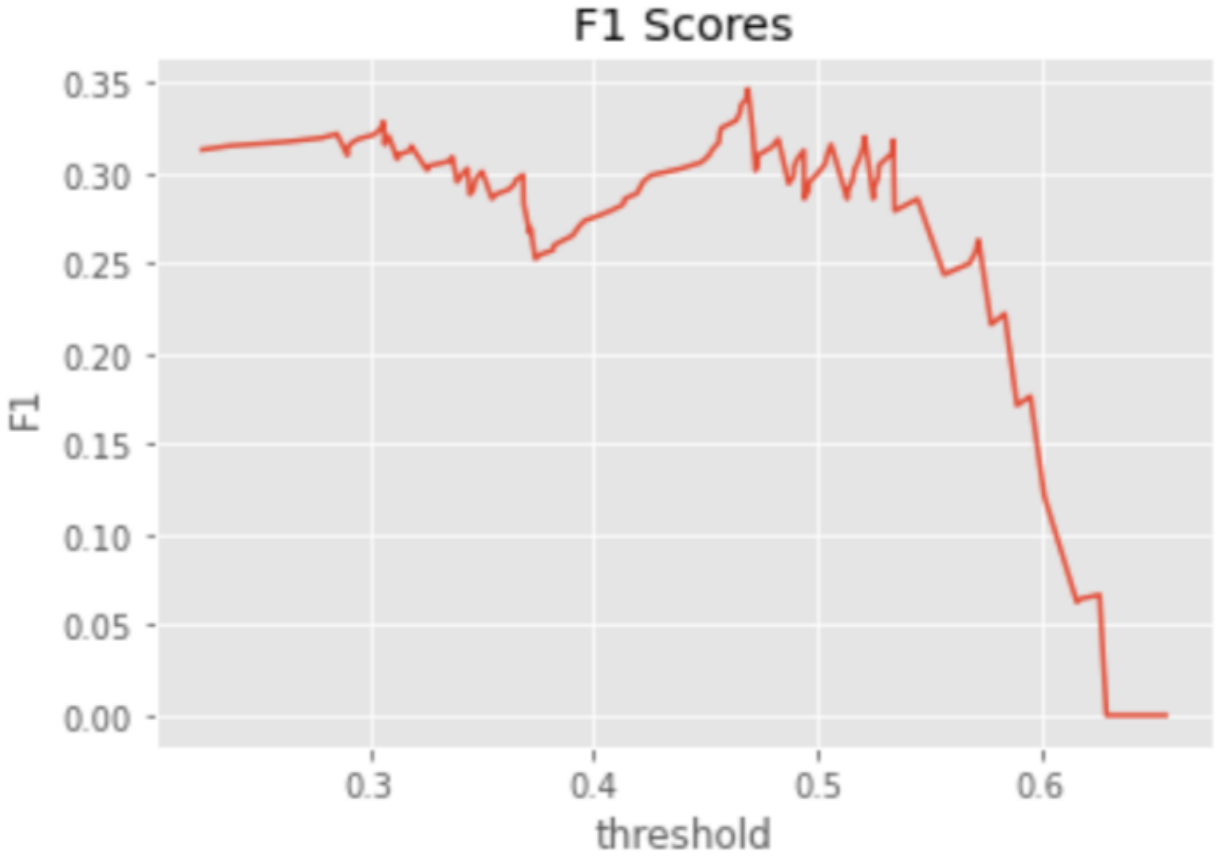
Patient Population Description for FDA Validation Dataset:

- Gender: both man and woman
- Age: Between 1 to 95
- Body part: Chest
- Imaging modality: DX
- Patient position: PA or AP
- Known diseases: Without infiltration, Atelectasis, Effusion and consolidation

Ground Truth Acquisition Methodology:

In generall gold standard ground truth is the best ground truth for training. It can be obtained by using of some physical experiment such as biopsy. It costs a lot of time and very expensive.
In comparison to that silver standard ground truth are also important. The final diagnosis is determined by a voting system across all of the radiologists' labels for each image.

Algorithm Performance Standard:
This algorithm's performances can be measured by calculating F1 scores.



As the F1 scores diagramm above it can be seen our optimal threshold is about 0.47. However, in order to aviod missing positive cases, a recall weighted threshold be selected.