FDA Submission

Name: Harshit
Device Name: Pneumonia Detector

General Information

Intended Use Statement:

This algorithm is intended for use on Pneumonia patients from the ages of 1-100 (as there are outliers in the dataset) who have been administered a screening chest X-ray and have never before demonstrated an abnormal Chest X-ray study.

Indication of Use:

The algorithm can be used for the screening of the chest X-ray which can be helpful in the early detection of Pneumonia.

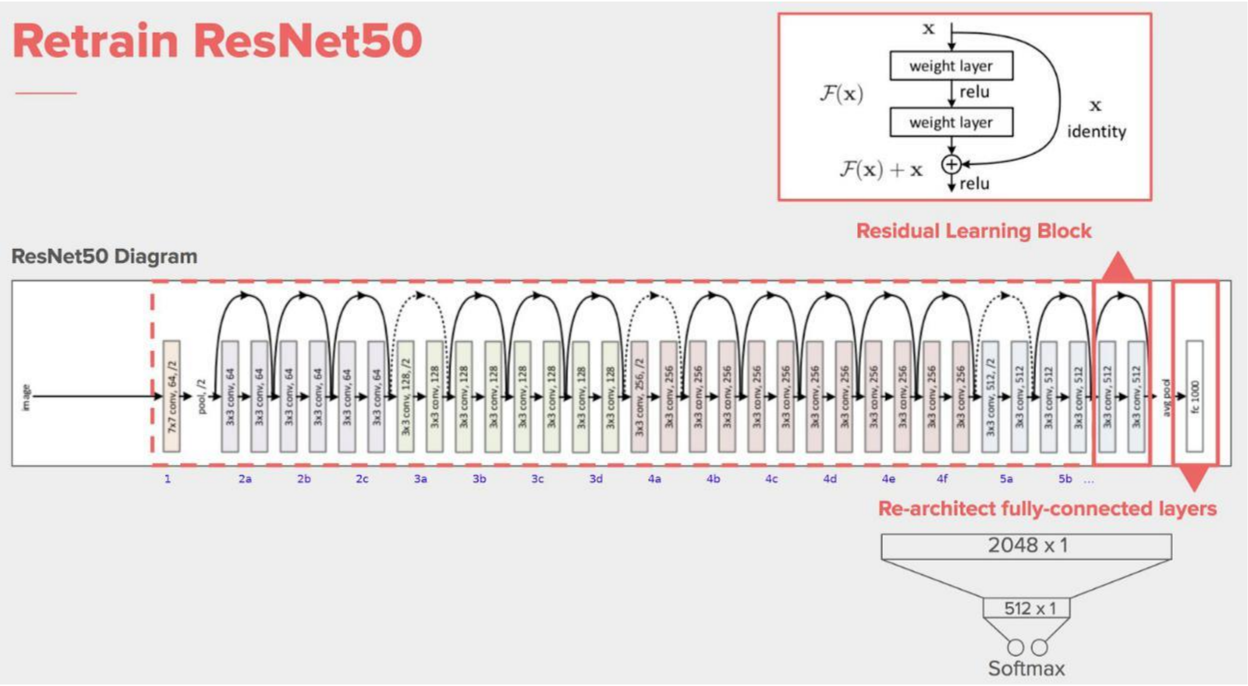
Device Limitations:

The results of the algorithm indicate that the presence of Atelectasis, Cardiomegaly, Consolidation, Edema, Effusion, Emphysema, Fibrosis, Hernia, Infiltration, Mass, Nodule, Pleural Thickening, or Pneumothorax in a chest x-ray may lead to false-positive pneumonia classifications as specificity is 1, however, the presence of Pneumonia can be accurately detected from this algorithm.
The model use is trained on frontal X-rays. So, other angles X-rays not may perform accurate result.

Clinical Impact of Performance:

The algorithm has a very good result in detecting Pneumonia which can be very beneficial in automating the process of detecting Pneumonia cases directly from the chest X-ray which not saves a lot of time in diagnosing the patient and also preparing the patient mentally as well.

Algorithm Design and Function



The algorithm is designed based on the pre-trained resnet50 model. After the last layer of the resnet50 model. Following layers are added:

```
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #   
=================================================================
Resnet (Model)               (None, None, None, 2048)  23564800  
_________________________________________________________________
batch_normalization_1 (Batch (None, None, None, 2048)  8192      
_________________________________________________________________
conv2d_1 (Conv2D)            (None, None, None, 1024)  2098176   
_________________________________________________________________
dropout_1 (Dropout)          (None, None, None, 1024)  0         
_________________________________________________________________
batch_normalization_2 (Batch (None, None, None, 1024)  4096      
_________________________________________________________________
conv2d_2 (Conv2D)            (None, None, None, 256)   262400    
_________________________________________________________________
dropout_2 (Dropout)          (None, None, None, 256)   0         
_________________________________________________________________
average_pooling2d_1 (Average (None, None, None, 256)   0         
_________________________________________________________________
batch_normalization_3 (Batch (None, None, None, 256)   1024      
_________________________________________________________________
conv2d_3 (Conv2D)            (None, None, None, 1)     257       
_________________________________________________________________
reshape_1 (Reshape)          (None, None)              0         
=================================================================
Total params: 25,938,945
Trainable params: 2,367,489
Non-trainable params: 23,571,456
_________________________________________________________________
```

The model takes the input of X-ray images with (224, 224, 3) and gives output probability of Pneumonia / Non-pneumonia.

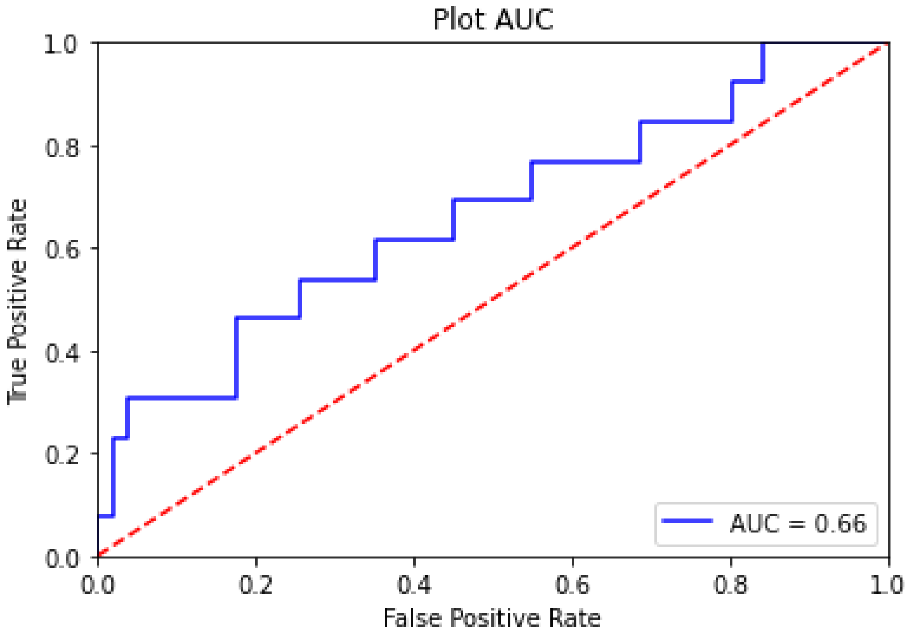
- Resnet50 model with non-trainable layers with pre-trained ImageNet weights.
- Then 3 Conv2D layers with different filters works as a Dense layer
- To fine tune the weights dropout layers are used after the convolution layer with 0.5 probability.
- Average pooling and batch normalization used to normalize the inputs from before layer.
- Last convolution layer gives output in range 0-1.

Dicom checks:

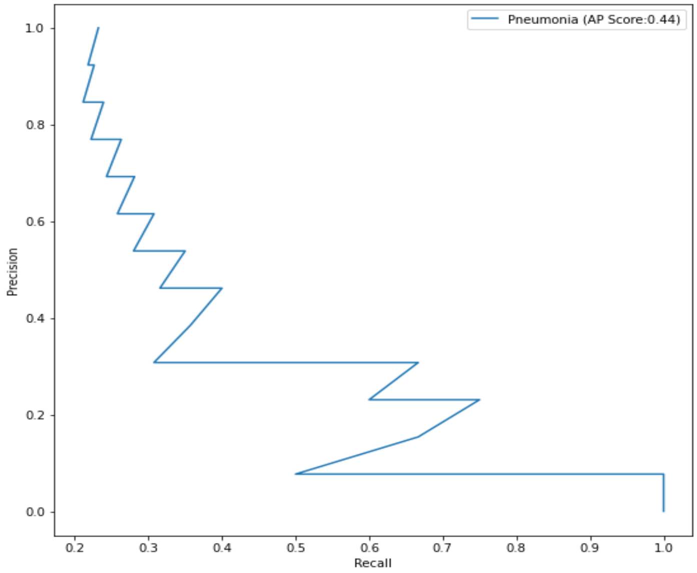
- Input X-ray image must be of ‘CHEST’.
- Modality must be ‘DX’
- Patients position must be ‘Anteroposterior’ or ‘Poster-anterior.’

For training data augmentation, the image is rescaled to 1/255, horizontal flip, height shift range of 0.05, width shift range of 0.1, rotation range of 10, shear range of 0.1, the zoom range of 0.05, so that model will consider those image as well which are slightly rotated or zoomed. For the training image generator, a batch size of 64 images is used.

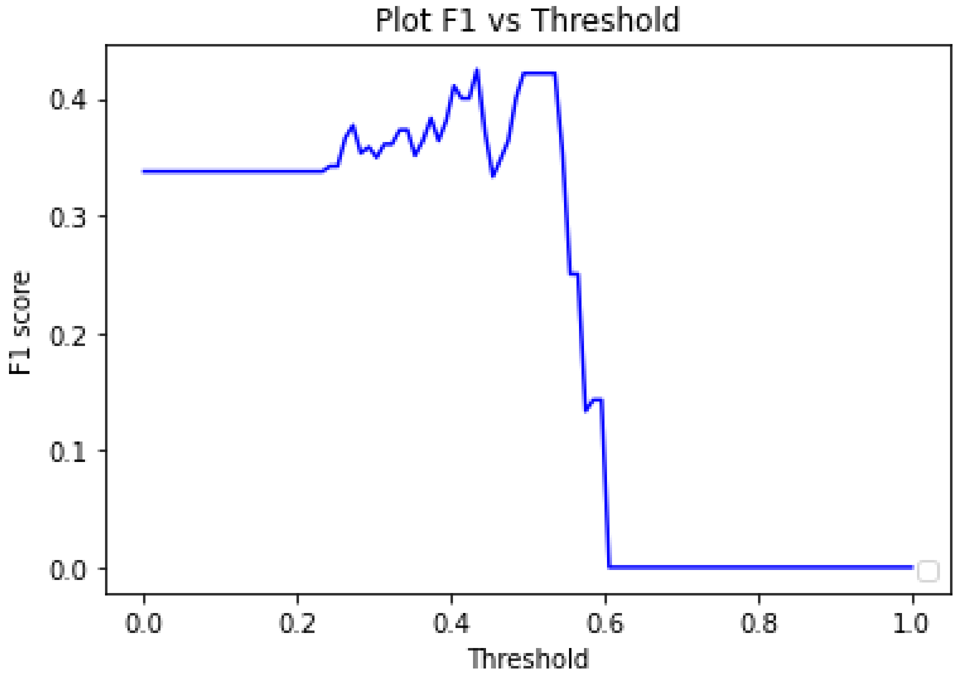
- The algorithm is trained based on RMSprop optimizer and a learning rate of 10^-4. Then the model remaining is fine-tuned for this case for about more than 5 epochs until the model loss is stabilized.
- The current model was trained up to 0.6087 training accuracy and training loss of 0.7355 which has a final validation accuracy of 0.7500.



- The algorithm has an area under the curve for True positive rate and false positive rate of 0.66.



- Precision-Recall curves summarize the trade-off between the true positive rate and the positive predictive value for a predictive model using probability thresholds.
- In medical, better precision-recall needed to reduce False positives. Curve helps to identify better threshold for algorithm.



- The calculated threshold is 0.50 with the F1 score of 0.4210 which has the best accuracy of 89.90%.

At the last overall performance of the model is:

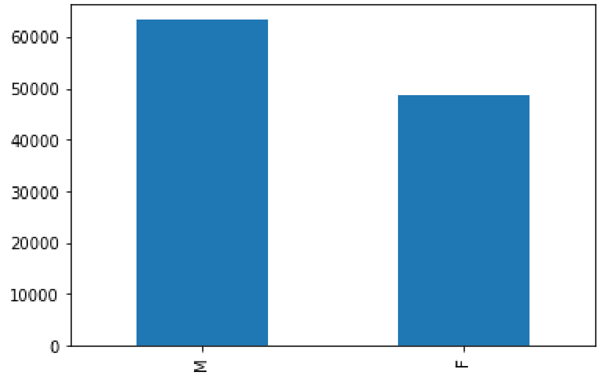
```
CONFUSION MATRIX **************************
[[4     9]
[ 2     49]]

TEST METRICS ******************************
Accuracy: 82.8125%
Precision: 84.48275862068965%
Recall: 96.07843137254902%
F1-score: 89.90825688073394
```

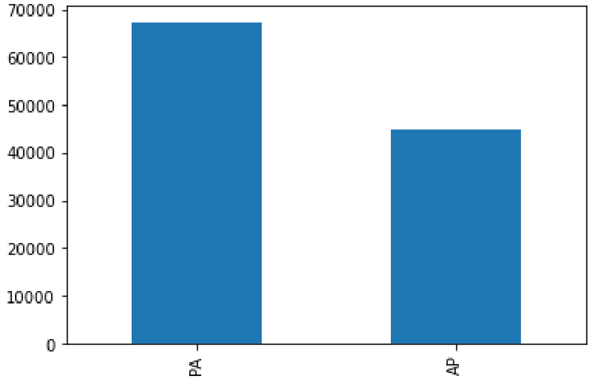
Databases

The dataset was curated by the NIH. There are 112,120 X-ray images with disease labels from 30,805 unique patients in this dataset. The disease labels were created using Natural Language Processing (NLP) to mine the associated radiological reports. The labels include 14 common thoracic pathologies:

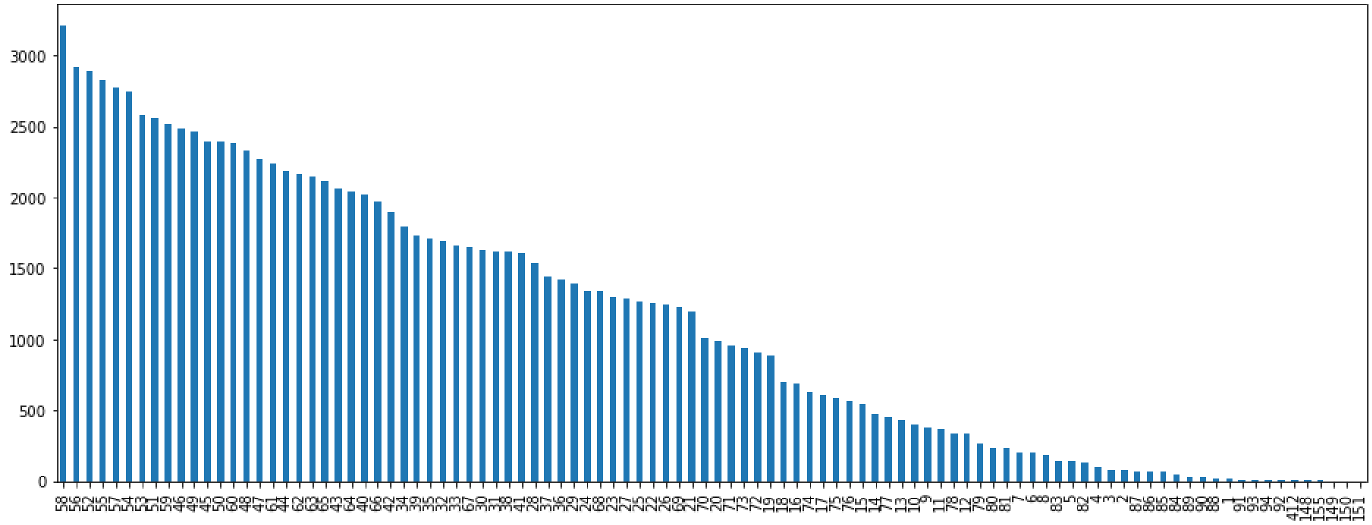
- Atelectasis
- Consolidation
- Infiltration
- Pneumothorax
- Edema
- Emphysema
- Fibrosis
- Effusion
- Pneumonia
- Pleural thickening
- Cardiomegaly
- Nodule
- Mass
- Hernia

The dataset consists of 63340 male participants and 48780 female participants. The positioning of the participant during the chest X-ray is AP in 44810 cases and PA during 67310 cases.

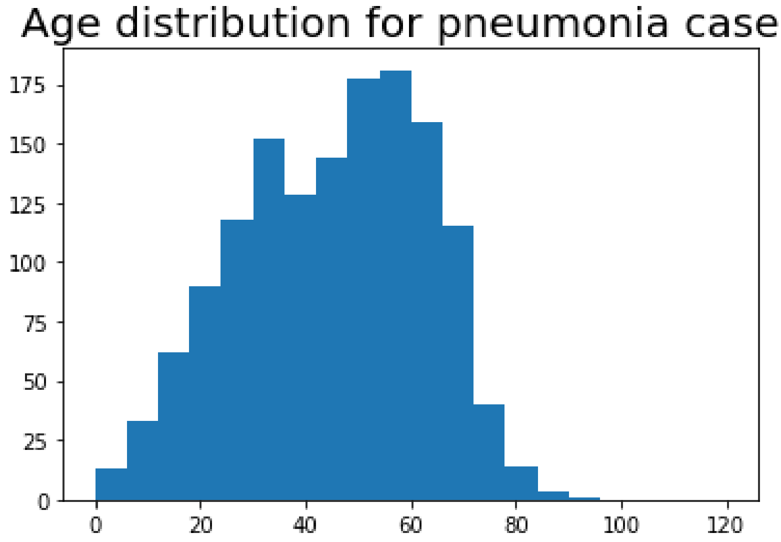


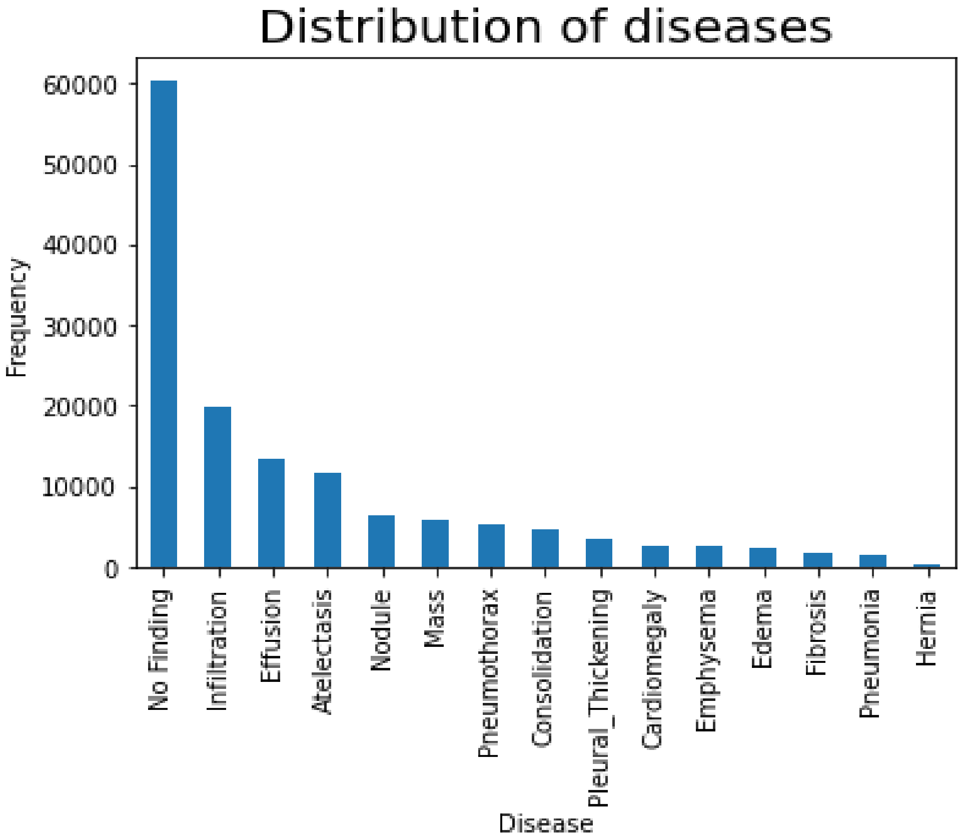


The participants in the dataset range from 1-100 with some outliers.

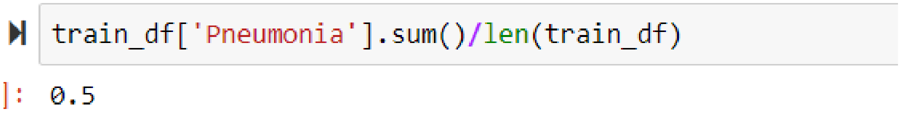


There are 1413 pneumonia cases in the given dataset which can be in the presence of different 15 images. Age distribution shows that there are more chances of Pneumonia in age 20-70.





For training the model, I have split the dataset into 80% training data and a 20% dataset. Then the training data is balanced for 50% pneumonia cases and 50% non-pneumonia cases.

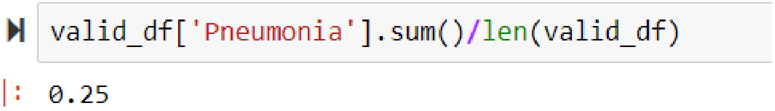


```
train_df['Pneumonia'].sum()/len(train_df)
```

```
]: 0.5
```

The training dataset is augmented as described in the previous section then the training data frame is generated with the target size of 224 x 224 and a batch size of 64.

Similarly, the validation dataset is generated as the validation dataset with 25% pneumonia cases and 50% non-pneumonia cases. But with image augmentation just rescaling to 1/255.



```
valid_df['Pneumonia'].sum()/len(valid_df)
```

```
]: 0.25
```

Ground Truth

The biggest limitation of this dataset is that image labels were NLP-extracted so there could be some erroneous labels but the NLP labeling accuracy is estimated to be >90%.

The original radiology reports are not publicly available but details about the labeling processcan be found here.

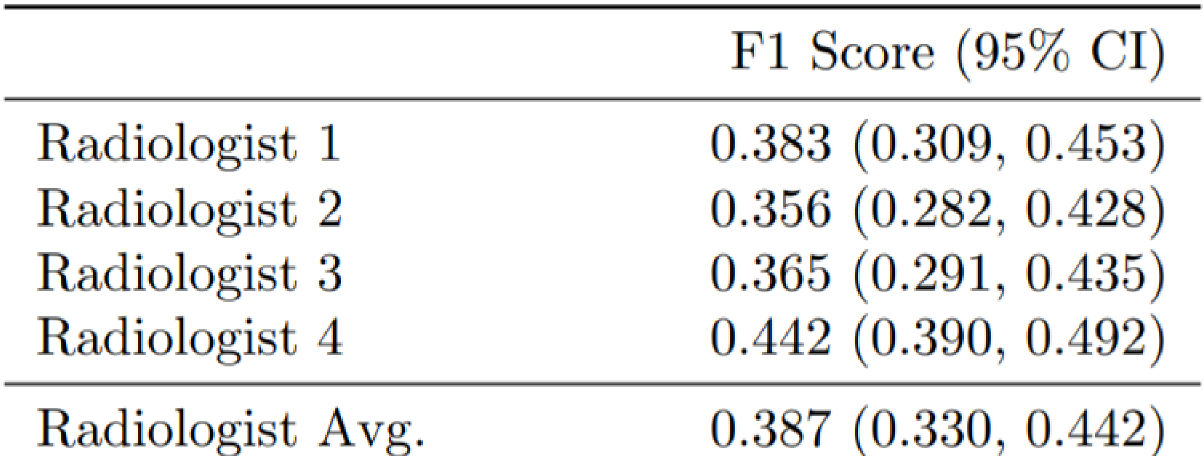
FDA Validation Plan

Patient Population Description for FDA Validation Dataset:
For validation of the algorithm, the collected dataset should be made up of chest X-rays between the ages 1-100 for both male and female. However, it should be made sure that the validation set did not contain a patient with prior history of Pneumonia and the patient with other diseases like Atelectasis, Cardiomegaly, Consolidation, Edema, Effusion, Emphysema, Fibrosis, Hernia, Infiltration, Mass, Nodule, Pleural Thickening, or Pneumothorax should be checked again with Radiologists for False Positive.

Ground Truth Acquisition Methodology:
For the acquisition of the chest X-ray, the patient can be in any position anteroposterior or poster-anterior. The patient should also be checked other diseases mentioned above, as it may result in False positive.

In other case for labeling, try to obtain biopsy data or having radiologists creating labels that should then be weighted across the radiologists to create a single label (a.k.a. “Silver standard”) for each image in the set. This process is better than NLP labeling but also have some conflicts. So, weights should be given properly.

Algorithm Performance Standard:



	F1 Score (95% CI)
Radiologist 1	0.383 (0.309, 0.453)
Radiologist 2	0.356 (0.282, 0.428)
Radiologist 3	0.365 (0.291, 0.435)
Radiologist 4	0.442 (0.390, 0.492)
Radiologist Avg.	0.387 (0.330, 0.442)

My given approach ends up with 0.4210 F1 score at 0.5 threshold value.
As compare to average score of radiologists this model is pretty accurate on detection.