FDA Submission

Your Name: Sebastian Yerovi

Name of your Device: PneumoGuess

Algorithm Description

1. General Information

Intended Use Statement:

Identification of Pneumonia on patient's chest X-ray imaging.

Indications for Use:

- Screening Chest X-ray exams with a 'PA' body position.
- Men and women between of all ages who present less than 4 respiratory pathologies at the time of X-ray imaging including Infiltration, Edema, Atelectasis, and Effusion.

Device Limitations:

- We recommend that physicians take our model's prediction as one data point as well as other data points when coming up with a diagnosis for a patient.
- The software is not intended for identifying any other disease that is not pneumonia.
- The software does not distinguishes between type, severity, stage, cause or any other characteristic of the pneumonia. Hence, some types of pneumonia might not be optimal for identification via this type of system, specially 'rare' types of pneumonia.
- Performance on men and women who do not fit within the 20-70 years old range might be affected due to lower prevalence of pneumonia cases in this age group in the dataset from which the model was built.
- Performance may vary on patients who also suffer from Cardiomegaly, Edema, or Mass. Those diseases have a 'close' degree of similarity in terms of pixel intensity values. In other words, look somewhat similar to Pneumonia. Hence, the algorithm, just like an expert radiologist, could misinterpret those pathologies as Pneumonia.
- Patients who present more than 4 respiratory pathologies at the time of X-ray imaging.
- Not intended for medical imaging that is not Chest-Xray.
- Not intended for diagnosis on animal pneumonia diagnosis.
- Performance may vary in Chest X-ray images taken from a 'AP' body position.

Clinical Impact of Performance:

- The current model has settled for a threshold of 0.45 at which F1-Score equals 0.44.
- Due to the purpose of screening Pneumonia, the algorithm should carry a bias towards False Positive cases over False Negative cases if a FP/FN trade-off is required. A False Negative case could lead to misdiagnosis, over-confidence, lack of proper treatment and potentially patient's death. While a high False Positive rate is generally not desired, if a patient is misdiagnosed with Pneumonia, physicians could pair these results with other symptoms and clinical information to re-validate such results.

2. Algorithm Design and Function

DICOM Checking Steps: In order to integrate into workflow the algorithm uses the `pydicom` library based off on the DICOM Standard.

1. The algorithm reads the DICOM information and displays relevant information such as the Image Type, Body Part, Body Position used to evaluate whether the algorithm would be a good fit for usage.

2. Algorithm reads the image path

3. Displays the images
4. Returns the image

Preprocessing Steps:

1. The image is normalized by rescaling (just like the training data)
2. The normalized image is resized to fit the input size requirements of the model, in this case: `(1,224,224,3)`.

CNN Architecture:

The model takes VGG-16 Convolutional Neural Net architecture with "ImageNet" pre-trained weights as the start up until the `block5_pool` layer. This is a Convolutional Neural Network that has been pre-trained up to a really high standard. The idea is to use those pre-trained weights via transfer learning for speeding up the process of learning lower level features (corners, edges, elements) and just focusing on training higher level (body parts, diseases in X-ray images).

```
 Model: "model_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #   
=================================================================
input_1 (InputLayer)         (None, 224, 224, 3)       0         
_________________________________________________________________
block1_conv1 (Conv2D)        (None, 224, 224, 64)      1792      
_________________________________________________________________
block1_conv2 (Conv2D)        (None, 224, 224, 64)      36928     
_________________________________________________________________
block1_pool (MaxPooling2D)   (None, 112, 112, 64)      0         
_________________________________________________________________
block2_conv1 (Conv2D)        (None, 112, 112, 128)     73856     
_________________________________________________________________
block2_conv2 (Conv2D)        (None, 112, 112, 128)     147584    
_________________________________________________________________
block2_pool (MaxPooling2D)   (None, 56, 56, 128)       0         
_________________________________________________________________
block3_conv1 (Conv2D)        (None, 56, 56, 256)       295168    
_________________________________________________________________
block3_conv2 (Conv2D)        (None, 56, 56, 256)       590080    
_________________________________________________________________
block3_conv3 (Conv2D)        (None, 56, 56, 256)       590080    
_________________________________________________________________
block3_pool (MaxPooling2D)   (None, 28, 28, 256)       0         
_________________________________________________________________
block4_conv1 (Conv2D)        (None, 28, 28, 512)       1180160   
_________________________________________________________________
block4_conv2 (Conv2D)        (None, 28, 28, 512)       2359808   
_________________________________________________________________
block4_conv3 (Conv2D)        (None, 28, 28, 512)       2359808   
_________________________________________________________________
block4_pool (MaxPooling2D)   (None, 14, 14, 512)       0         
_________________________________________________________________
block5_conv1 (Conv2D)        (None, 14, 14, 512)       2359808   
_________________________________________________________________
block5_conv2 (Conv2D)        (None, 14, 14, 512)       2359808   
_________________________________________________________________
block5_conv3 (Conv2D)        (None, 14, 14, 512)       2359808   
_________________________________________________________________
block5_pool (MaxPooling2D)   (None, 7, 7, 512)         0         
=================================================================
Total params: 14,714,688
Trainable params: 14,714,688
Non-trainable params: 0
```

After, which several Dense layers were added alongside Dropout layers to avoid overfitting.

The model culminates with a Dense binary classification output layer.

```
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #   
=================================================================
model_1 (Model)              (None, 7, 7, 512)         14714688  
_________________________________________________________________
flatten_1 (Flatten)          (None, 25088)             0         
_________________________________________________________________
dropout_1 (Dropout)          (None, 25088)             0         
_________________________________________________________________
dense_1 (Dense)              (None, 1024)              25691136  
_________________________________________________________________
dropout_2 (Dropout)          (None, 1024)              0         
_________________________________________________________________
dense_2 (Dense)              (None, 512)               524800    
_________________________________________________________________
dropout_3 (Dropout)          (None, 512)               0         
_________________________________________________________________
dense_3 (Dense)              (None, 256)               131328    
_________________________________________________________________
dense_4 (Dense)              (None, 1)                 257       
=================================================================
Total params: 41,062,209
Trainable params: 26,347,521
Non-trainable params: 14,714,688
```

3. Algorithm Training

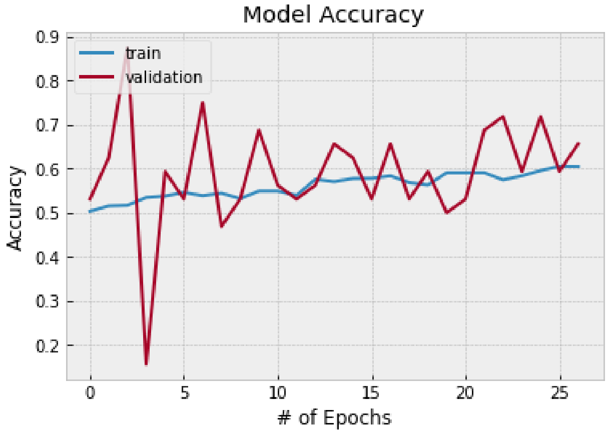
Parameters: * Types of augmentation used during training:

- Horizontal flip, since its representative of human body.
- `height_shift_range= 0.1` to represent tall, skinny people
- `width_shift_range=0.1` to represent shorter, fat people.
- `rotation_range=20` to account for different positions.
- `shear_range = 0.1,` o account for measurement techniques or image distortions.

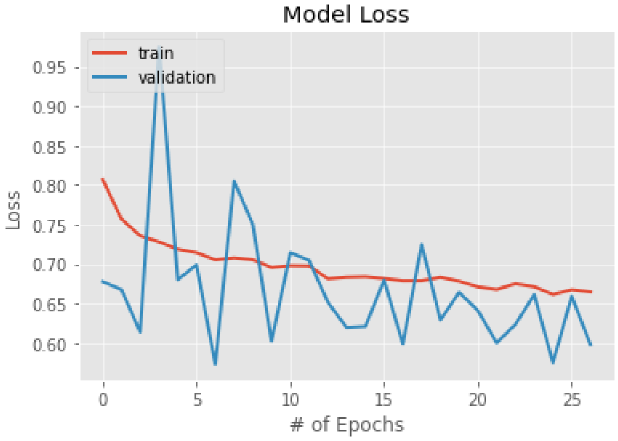
- `zoom_range=0.1` to account for measurement techniques
    - Batch size = 32
    - Optimizer learning rate: `Adam Optimizer` with `learning_rate = 1e-4`
    - Layers of pre-existing architecture that were frozen: All VGG-16 imported layers except `block5_pool` layer.
    - Layers of pre-existing architecture that were fine-tuned: VGG-16's `block5_pool` layer plus all added layers.
    - Layers added to pre-existing architecture:

```
flatten_1 (Flatten)          (None, 25088)             0         
_________________________________________________________________
dropout_1 (Dropout)          (None, 25088)             0         
_________________________________________________________________
dense_1 (Dense)              (None, 1024)              25691136  
_________________________________________________________________
dropout_2 (Dropout)          (None, 1024)              0         
_________________________________________________________________
dense_2 (Dense)              (None, 512)               524800    
_________________________________________________________________
dropout_3 (Dropout)          (None, 512)               0         
_________________________________________________________________
dense_3 (Dense)              (None, 256)               131328    
_________________________________________________________________
dense_4 (Dense)
```

Accuracy



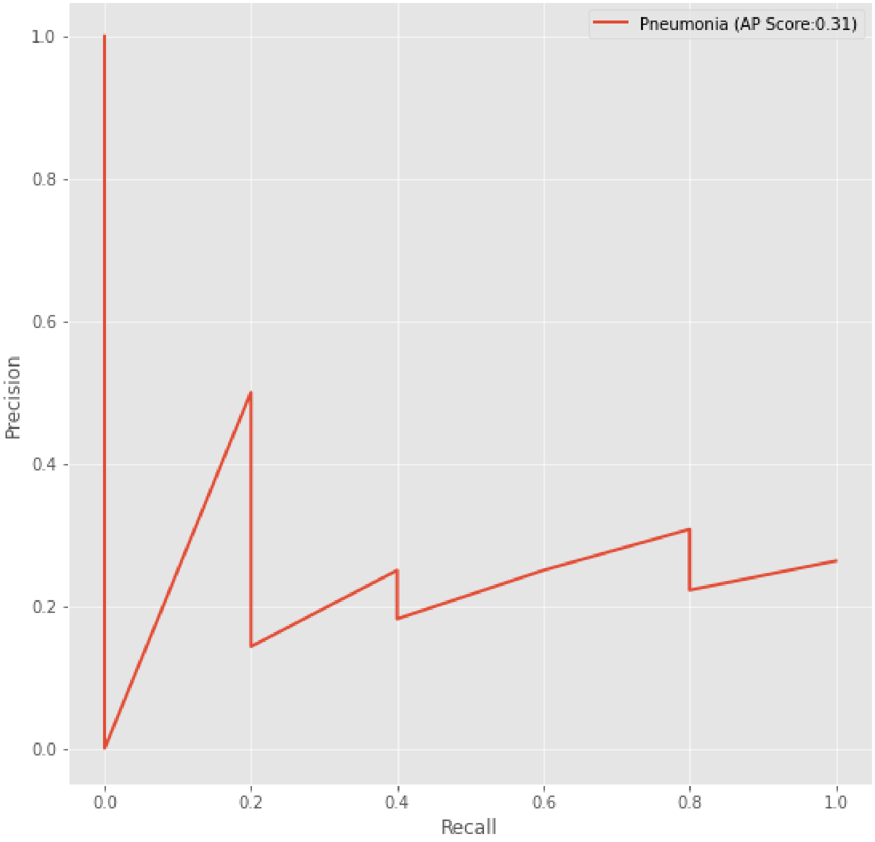
Loss:

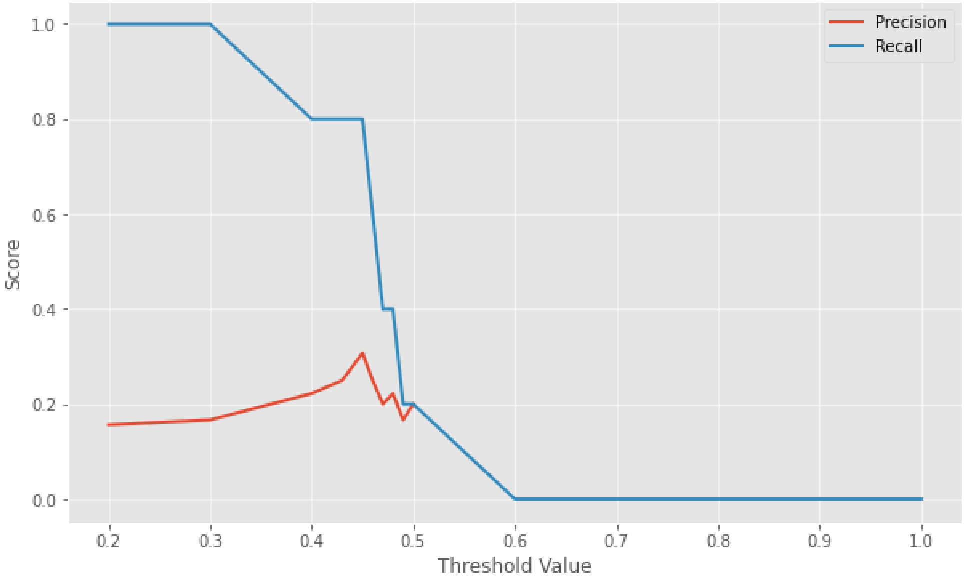


Precission-Recall curve:

The precision-recall curve and trade-off is a really important one to consider. In general, for our type of problem a bias towards recall its a better approach. Higher recall means that we are going to get right more often the ones that are positive cases, this very likely would also mean a low precision. So, from the ones that we categorized as positive, a chunk of it are actually False Positives. While this is in fact a problem the consequence are 'lighter'. Yes, patients might get scare and the healthcare provider would incur into more expenses but patient's life is at a lesser risk than with a high rate of False Negatives.

So, low recall, would translate into higher False Negatives a problem that might put in danger patients lives.





Final Threshold and Explanation:

The best threshold to choose based on this model is 0.45 given that at such point we can achieve the highest F1-Score (0.44). At such threshold as well we are able to achieve the highest precision (0.3), without compromising so much on recall (0.8) which is a more relevant metric than precision. The best combination so far. At such point:

FN vs FP

When t=0.2 the number of False Positives is really high. This makes sense as with the Threshold being so low, the model classifies everything as positives. As we keep increasing our T, the number of FP decreases.

Also, after hitting t=0.5 the trends reverse. At that point apparently the model starts to present a bias towards categorizing as negative cases. So, the number of False Negatives start rising.

Sensitivity vs Specificity

Specificity starts really high at 1. It is rational that this occurs as our algorithm at this point classifies everything as positive which means we won't have any negative cases hence no False Negatives. But this alone is not a good metric since the number of False positives is really high - which would mean that we are going to tell every patient that comes in that has pneumonia (completely useless).

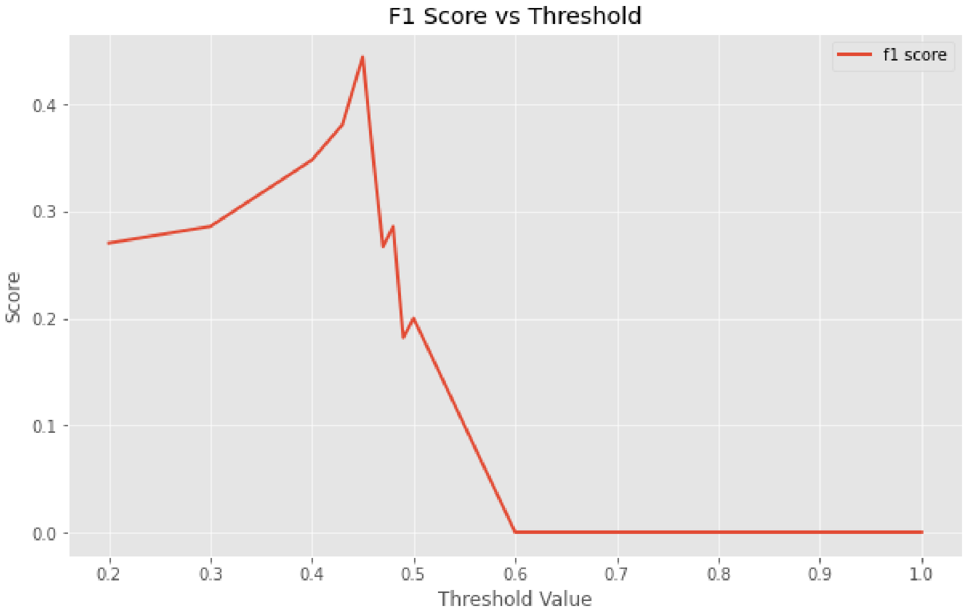
We see that Specificity increases as we increase t until hitting again 0.45. Which means that predictions are getting better at classifying the negative cases as well.

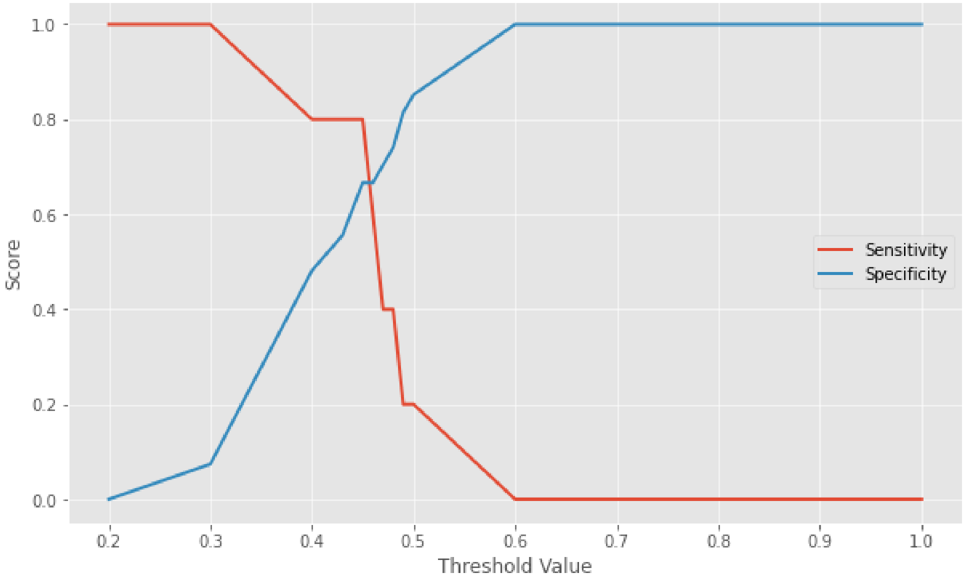
F1

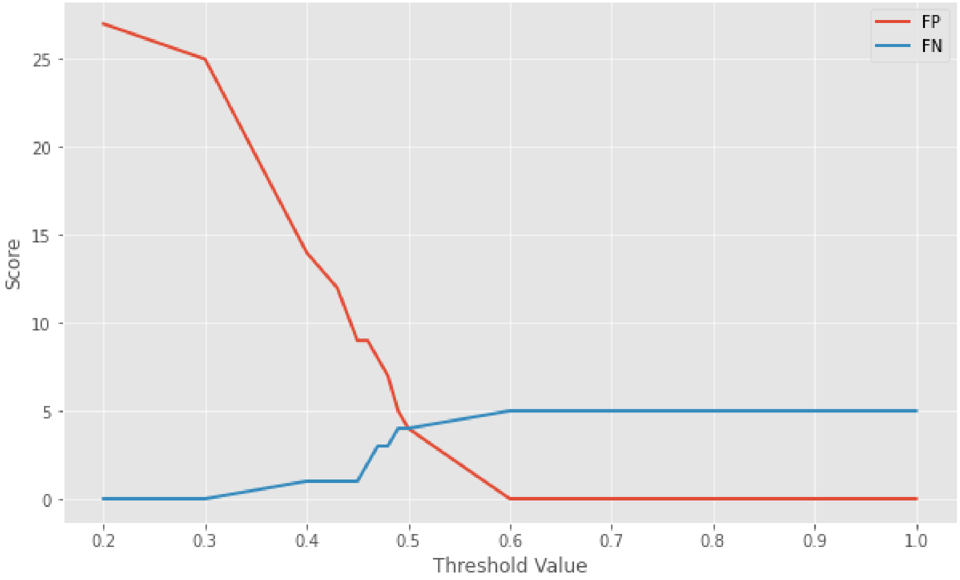
At t=0.45, F1 Score hits 0.44. The nature of the F1 curve shows that the F1 Score keeps increasing exponentially from t=0.2 until 1 at 0.45. This occurs as the number of false positives decreases.

Precision - Recall

Considering our intended bias towards recall and according to the other metrics t=0.45 is a good metric for minimizing potential misdiagnosis errors that would lead to missing positive cases (FN) as a consequence of low recall.







4. Databases

Description of Training Dataset:

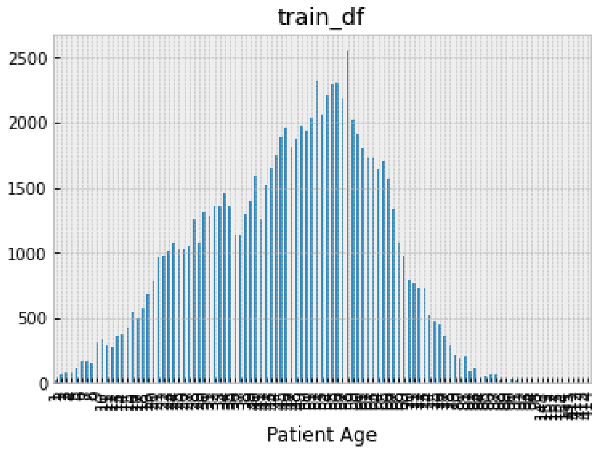
- Training set was created by randomly splitting the full dataset receiving 80% of registries.
- Original pneumonia presence was 1.2%, representative of the original prevalence.
- However, training dataset was rebalanced to 50% pneumonia vs 50% non-pneumonia cases.

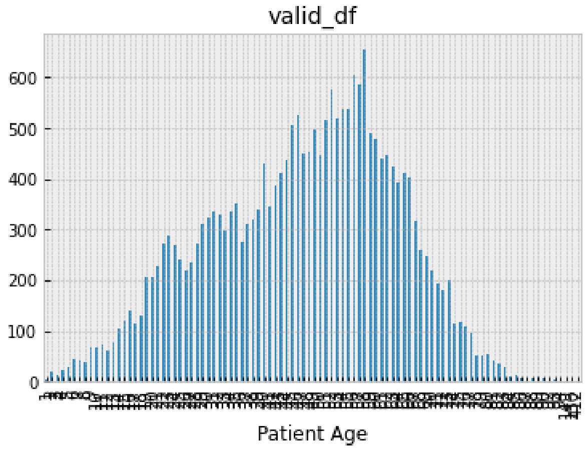
Description of Validation Dataset:

- Validation set was created by randomly splitting the full dataset receiving 20% of registries.
- Original pneumonia presence was 1.2%, representative of the original prevalence.
- Validation dataset was not rebalanced to 50/50 configuration.
- Instead, it was reconfigured to a 20% pneumonia prevalence, assuming 20% as the representation of real-world.

Both, datasets shared demographic data consistency upon split.

For example with age:





However, demographic data consistency was minimally affected after rebalancing.

A cautionary note that NLP-derived labels are suboptimal and could impact your algorithm's clinical performance.

5. Ground Truth

We use expert Radiologists' performance metrics as the Ground Truth reference.

The following metrics are taken from the paper CheXNet in which 4 expert radiologists with different number of years of experience classify 420 X-ray images for Pneumonia.

```
                    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)
CheXNet            0.435     (0.387, 0.481)
```

6. FDA Validation Plan

Patient Population Description for FDA Validation Dataset:

The FDA-Partner validation dataset should have the following characteristics:

- Patient population of both Male and Female of all ages with a higher proportion of people between the ages of 20-70 years old.
- Pneumonia prevalence in validation set should be 20%
- The majority of the images should be Chest X-Rays taken from a 'PA' body position. It should also include 'AP' type of images as well. The ideal ratio should be 3:2 (PA:AP).

Ground Truth Acquisition Methodology:

For ground truth we suggest to use the silver standard of Radiologist Avg. mentioned in the CheXNet paper. It takes into consideration 4 different expert radiologists.

Algorithm Performance Standard:

Considering the Ground Truth of Radiologist Avg, our PneumoGuess algorithm should hit at least a 0.387 f1-score in the validation set.