FDA Submission

Your Name: Swastik Nath

Name of your Device: Pneumonia Chest X-Ray Detection Assistance System

Algorithm Description

1. General Information

Intended Use Statement: For assisting a radiologist in detection of pneumonia in x-ray images.

Indications for Use: For assisting a radiologist in detection of pneumonia in x-ray images.

Indications for Use:

Screening of x-ray images.

Patient population: * Both men and women * Age: 2 to 90

X-Ray image properties: * Body part: Chest * Position: AP (Anterior/Posterior) or PA (Posterior/Anterior) * Modality: DX (Digital Radiography)

Device Limitations: Consolidation * Edema * Effusion * Hernia

This can be explained by the fact that X-Ray images containing four conditions mentioned above have similar but significantly different distribution that that of Pneumonia X-Rays:

In terms of computer hardware requirements, the algorithm's performance was measured on Intel(R) CORE i5(R) @ 1.60GHz CPU:

- Average image pre-processing time: 34 ms, max: 149 ms
- Average CNN inference time: 915 milliseconds, max: 1434 ms

So, total inference time does not exceed 1434 milliseconds on Intel Core i5 CPU. Similar (or higher performance) CPU is recommended.

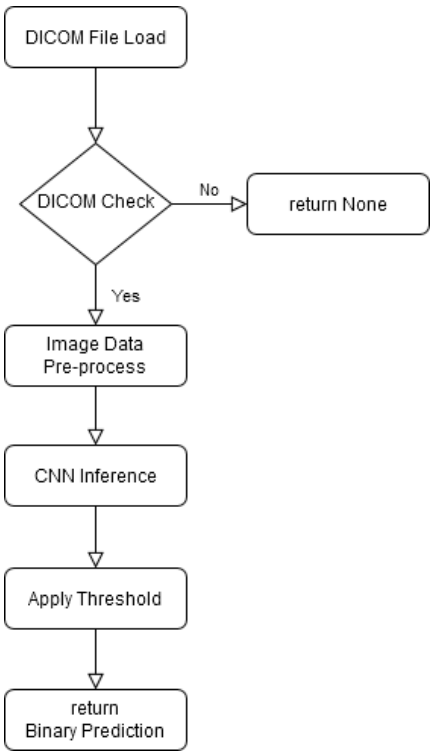
Clinical Impact of Performance: In terms of predictive value:

- If the model predicts negative, it is correct with 90.5% probability
- If the model predicts positive, it is correct with 26.8% probability

Therefore, the algorithm is recommended for assisting a radiologist screening those images that most probably do not contain Pneumonia, to prioritize his/her time and attention to those that potentially do. This should lead to those that require medical attention getting it much faster.

It is also worth mentioning that when algorithm predicts negative is can still be wrong with 9.5% probability. So, those cases predicted negative should still be reviewed by the radiologist.

2. Algorithm Design and Function



The algorithm performs the following checks on the DICOM image:

- Check Patient Age is between 2 and 90 (inclusive)
- Check Examined Body Part is 'CHEST'
- Check Patient Position is either 'PA' (Posterior/Anterior) or 'AP' (Anterior/Posterior)
- Check Modality is 'DX' (Digital Radiography)

Preprocessing Steps

The algorithm performs the following preprocessing steps on an image data:

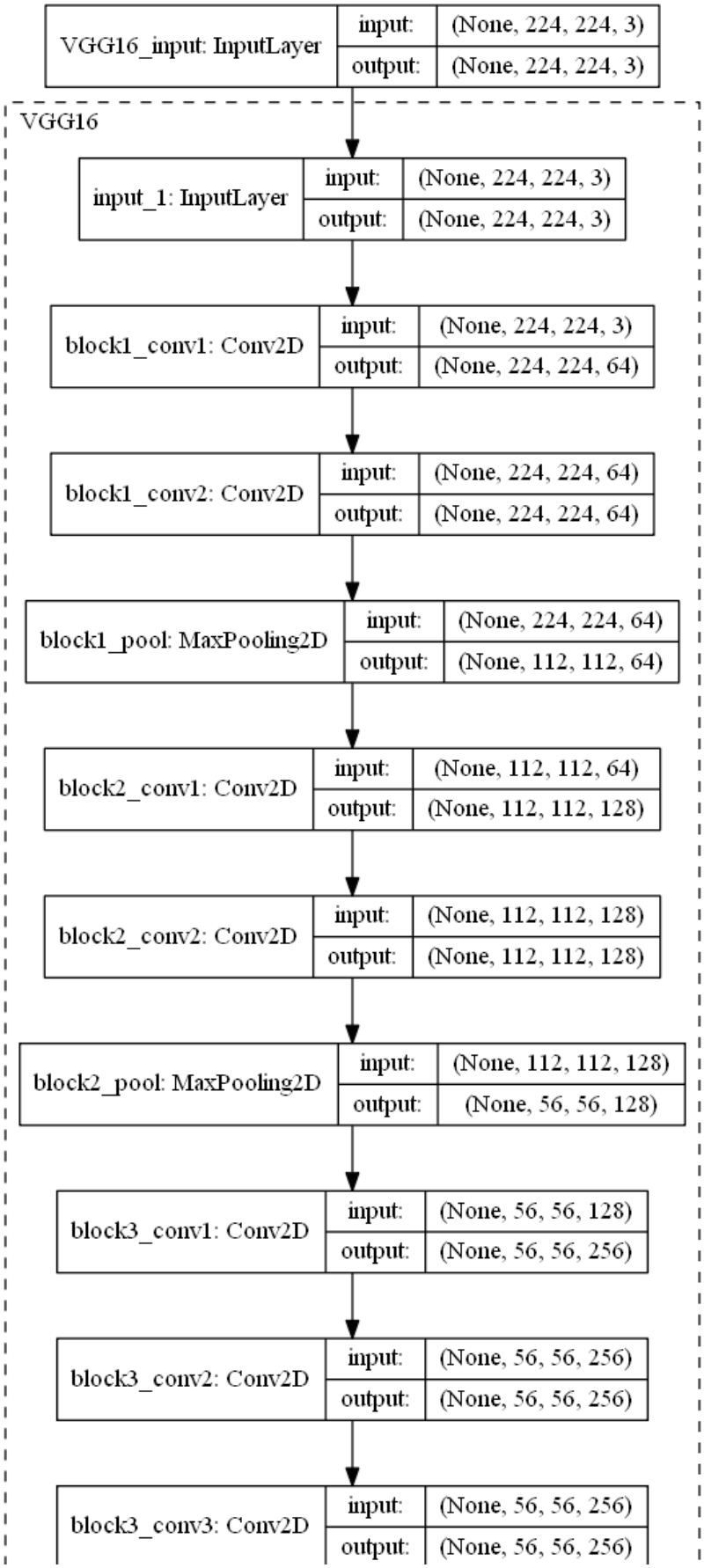
- Converts RGB to Grayscale (if needed)
- Re-sizes the image to 244 x 244 (as required by the CNN)
- Normalizes the intensity to be between 0 and 1 (from original range of 0 to 255)

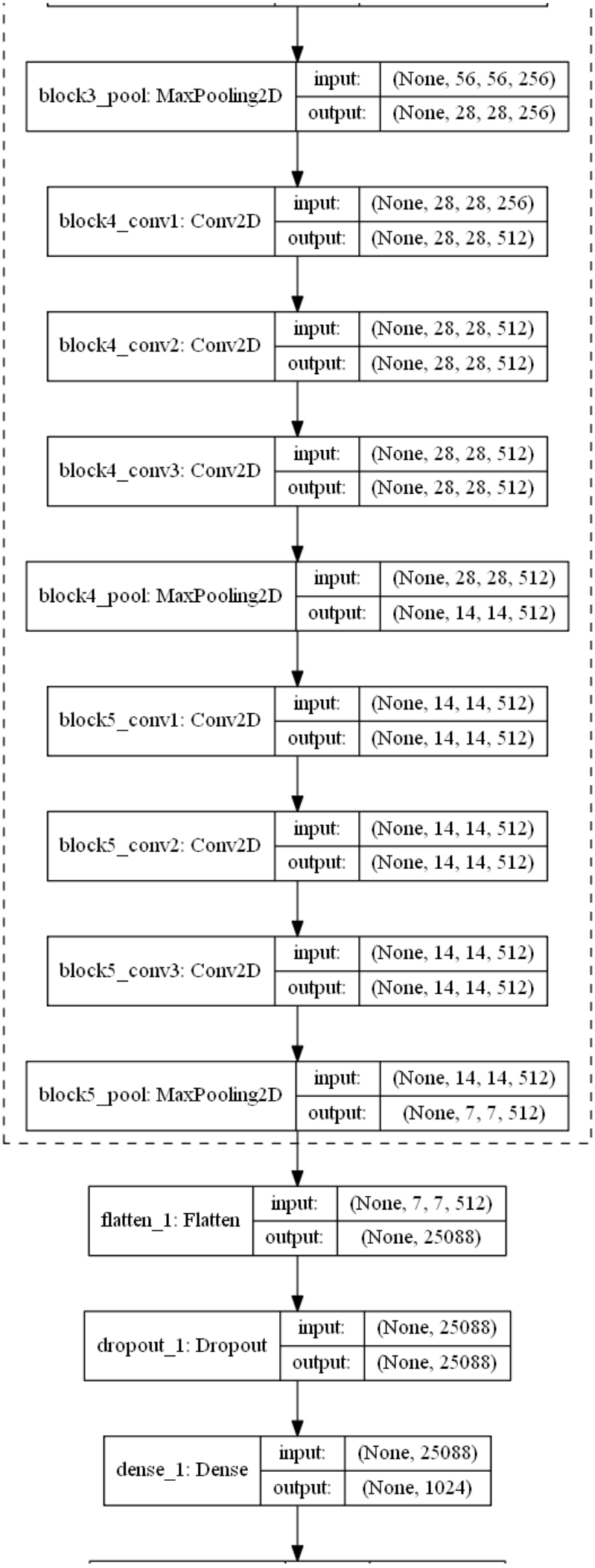
CNN Architecture

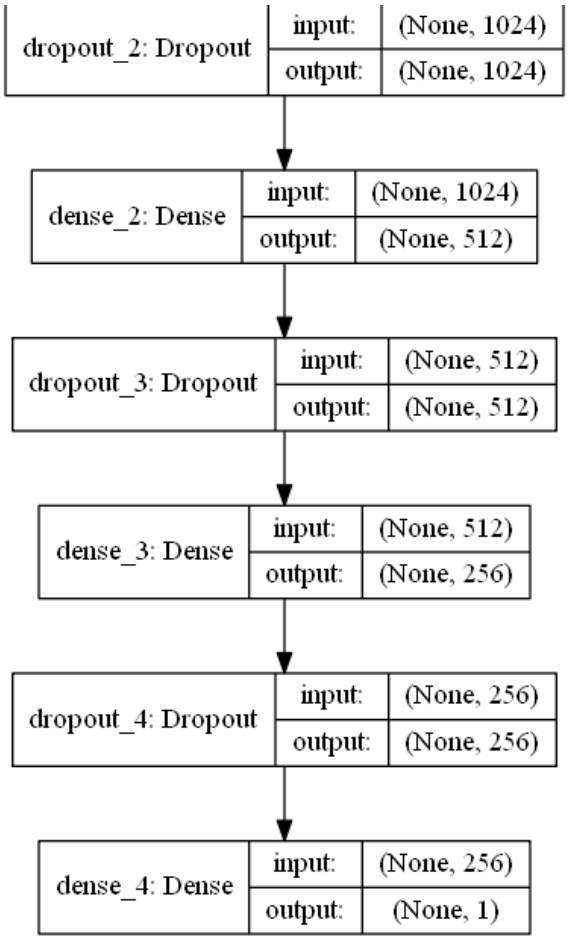
The algorithm uses pre-trained VGG16 Neural Network (except the last block of Convolution + Pooling layers that was re-trained), with additional 4 blocks of 'Fully Connected + Dropout' layers.

The network output is a single probability value for binary classification.

Below is the CNN architecture graph:

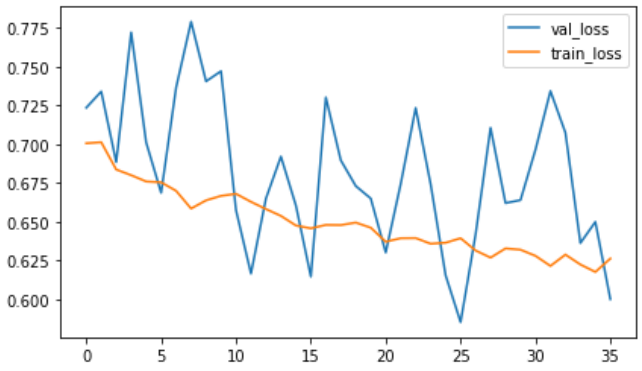


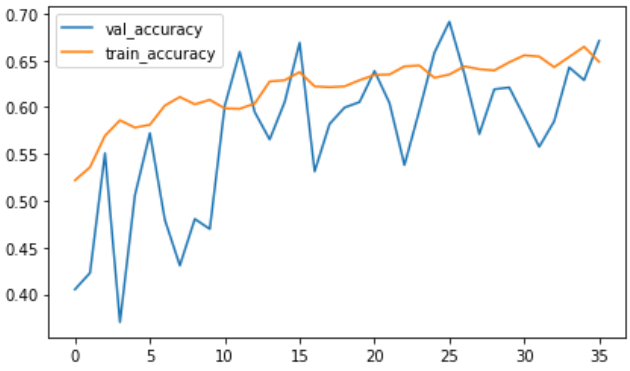

block3_pool: MaxPooling2D
input: (None, 56, 56, 256)
output: (None, 28, 28, 256)
block4_conv1: Conv2D
input: (None, 28, 28, 256)
output: (None, 28, 28, 512)
block4_conv2: Conv2D
input: (None, 28, 28, 512)
output: (None, 28, 28, 512)
block4_conv3: Conv2D
input: (None, 28, 28, 512)
output: (None, 28, 28, 512)
block4_pool: MaxPooling2D
input: (None, 28, 28, 512)
output: (None, 14, 14, 512)
block5_conv1: Conv2D
input: (None, 14, 14, 512)
output: (None, 14, 14, 512)
block5_conv2: Conv2D
input: (None, 14, 14, 512)
output: (None, 14, 14, 512)
block5_conv3: Conv2D
input: (None, 14, 14, 512)
output: (None, 14, 14, 512)
block5_pool: MaxPooling2D
input: (None, 14, 14, 512)
output: (None, 7, 7, 512)
flatten_1: Flatten
input: (None, 7, 7, 512)
output: (None, 25088)
dropout_1: Dropout
input: (None, 25088)
output: (None, 25088)
dense_1: Dense
input: (None, 25088)
output: (None, 1024)



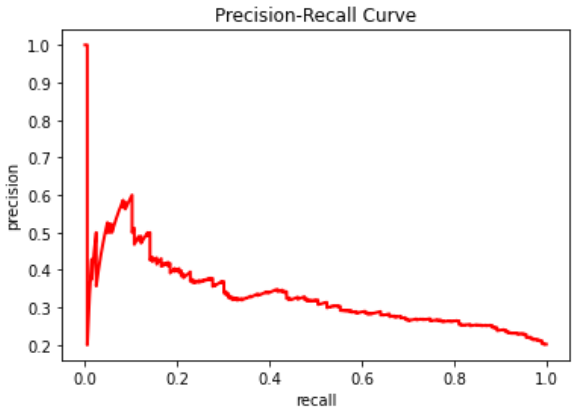
3. Algorithm Training

Parameters: * Types of augmentation used during training: * horizontal flip * height shift: 0.1 * width shift: 0.1 * rotation angle range: 0 to 20 degrees * shear: 0.1 * zoom: 0.1 * Batch size: 32 * Optimizer learning rate: 1e-5 * Layers of pre-existing architecture that were frozen * All except last convolution + pooling block * Layers of pre-existing architecture that were fine-tuned * The last 2 layers of VGG16 network: block5_conv3 + block5_pool * Layers added to pre-existing architecture * flatten_1 (Flatten)
* dropout_1 (Dropout)
* dense_1 (Dense, 1024)
* dropout_2 (Dropout, 0.2)
* dense_2 (Dense, 512)
* dropout_3 (Dropout, 0.2) * dense_3 (Dense, 256)
* dropout_4 (Dropout, 0.2) * dense_4 (Dense) 1





P-R Curve :



Final Threshold and Explanation: The maximum F1 score for the model is 0.427 and it is achieved with threshold value of 0.428. Below is the comparison of F1 score with those given in CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rayswith Deep Learning:

Person or Device	F1	95% CI 2sigma	68% CI 1sigma
Radiologist 1	0.383	(0.309, 0.453)	(0.345, 0.417)
Radiologist 2	0.356	(0.282, 0.428)	(0.319, 0.392)
Radiologist 3	0.365	(0.291, 0.435)	(0.327, 0.399)
Radiologist 4	0.442	(0.390, 0.492)	(0.416, 0.467)
Radiologist Avg.	0.387	(0.330, 0.442)	(0.358, 0.414)
CheXNet	0.435	(0.387, 0.481)	(0.411, 0.458)
Pneumonia Chest X-Ray Detection Assistance System	0.428		

We do not calculate here the models 95% confidence interval for simplicity, and compare models statistical significance by assuming normal distribution and simply comparing the models F1 score to 1-sigma CIs calculated from 2-sigma ones. This model's score is higher and outside of 68% (1-sigma) CI for two radiologists out of four, which points to 'some statistical significance' of the given model.

Comparing the F1 scores themselves, this model achieves higher maximum F1 score than an average radiologist in the study. State of the art neural network, as well as one radiologist from the study, do achieve higher F1 score, but the model's performance is comparable and in many cases exceeds the performance of human radiologists (in terms of F1 score).

Furthermore, since the model does not have a high precision with any meaningful recall value, its usefulness tends to lie in its recall (and negative predictive value). Therefore, it makes sense to maximize recall and NPV even at the cost of small loss in precision. A good threshold value that achieves that is 0.377:

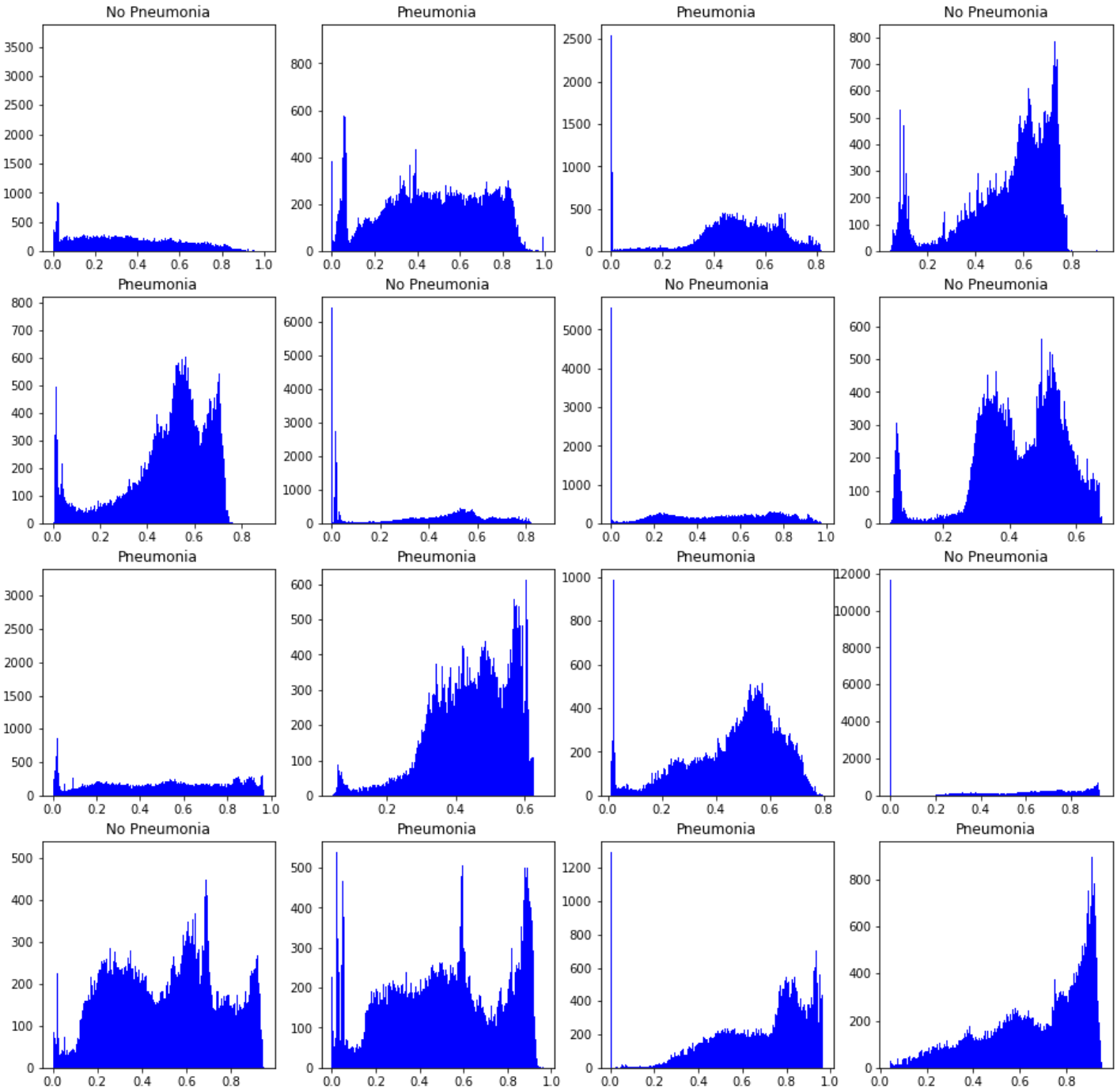
Device	F1	Precision	Sensitivity/Recall
Pneumonia Chest X-Ray Detection Assistance System(Max F1)	0.427	0.319	0.644

If the model predicts negative, it is correct with 90.5% probability. If the model predicts positive, it is correct with 26.8% probability.

Out of all negative cases the model correctly classifies 44.1%, out of all positive cases it correctly classifies 81.8%.

4. Databases

Description of Training Dataset: Training dataset consisted of 2290 chest xray images, with a 50/50 split between positive and negative cases.



Description of Validation Dataset: Validation dataset consisted of 1430 chest xray images, with 20/80 split between positive and negative cases, which more reflects the occurence of pneumonia in the real world.

5. Ground Truth

The data is taken from a larger xray dataset, with disease labels created using Natural Language Processing (NLP) mining the associated radiological reports. The labels include 14 common thoracic pathologies (Pneumonia being one of them): - Atelectasis - Consolidation - Infiltration - Pneumothorax - Edema - Emphysema - Fibrosis - Effusion - Pneumonia - Pleural thickening - Cardiomegaly - Nodule - Mass - Hernia

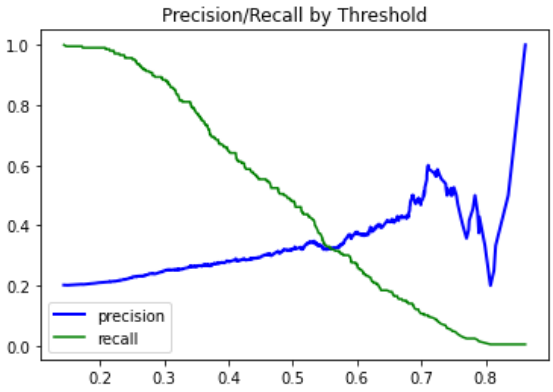
The biggest limitation of this dataset is that image labels were NLP-extracted so there could be some erroneous labels but the NLP labeling accuracy is estimated to be >90%.

The original radiology reports are not publicly available but more details on the labeling process can be found here.

6. FDA Validation Plan

Patient Population Description for FDA Validation Dataset:

The following population subset is to be used for the FDA Validation Dataset: * Both men and women * Age 2 to 90 * Without known comorbid thoracic pathologies listed above



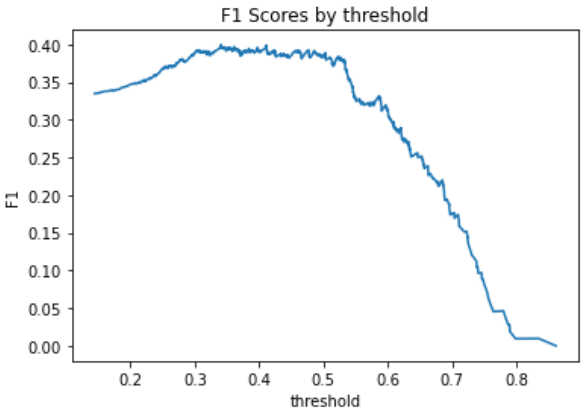
Ground Truth Acquisition Methodology:

The golden standard for obtaining ground truth would be to perform one of these tests (see this Mayo Clinic Link): * Sputum test * Pleural fluid culture

Yet, those tests are quite expensive, and in most cases diagnosis is concluded by the physician based on radiologist's analysis/description. Since the purpose of this device is assisting the radiologist (not replacing him), the ground truth for the FDA Validation Dataset can be obtained as an average of three practicing radiologists (as a widely used 'silver standard'). The same method is used in the mentioned paper.

Algorithm Performance Standard:

In terms of Clinical performance, the algorithm's performance can be measured by calculating F1 score against 'silver standard' ground truth as described above. The algorithm's F1 score should exceed **0.387** which is an average F1 score taken over three human radiologists, as given in CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rayswith Deep Learning, where a similar method is used to compare device's F1 score to average F1 score over three radiologists.



A 95% confidence interval given in the paper for average F1 score of 0.387 is (0.330, 0.442), so algorithm's 2.5% and 97.5% precentiles should also be calculated to get 95% confidence interval. This interval, when subtracted the interval above for the average, should not contain 0, which will indicate statistical significance of its improvement of the average F1 score. The same method for assessing statistical significance is presented in the above paper.