FDA Submission

Device name: Pneumonia presence detection algorithm

Algorithm Description

1. General Information

Intended Use Statement: Detection of the presence of pneumonia on a chest x-ray.

Indications for Use: Assists in the detection of pneumonia presence on x-rays, with an indication for use of removal from a radiologist's priority queue, and to flag x-rays with potential pneumonia presence for an earlier review.

Device Limitations: For fast execution of the algorithm, GeForce 1070 or higher required.

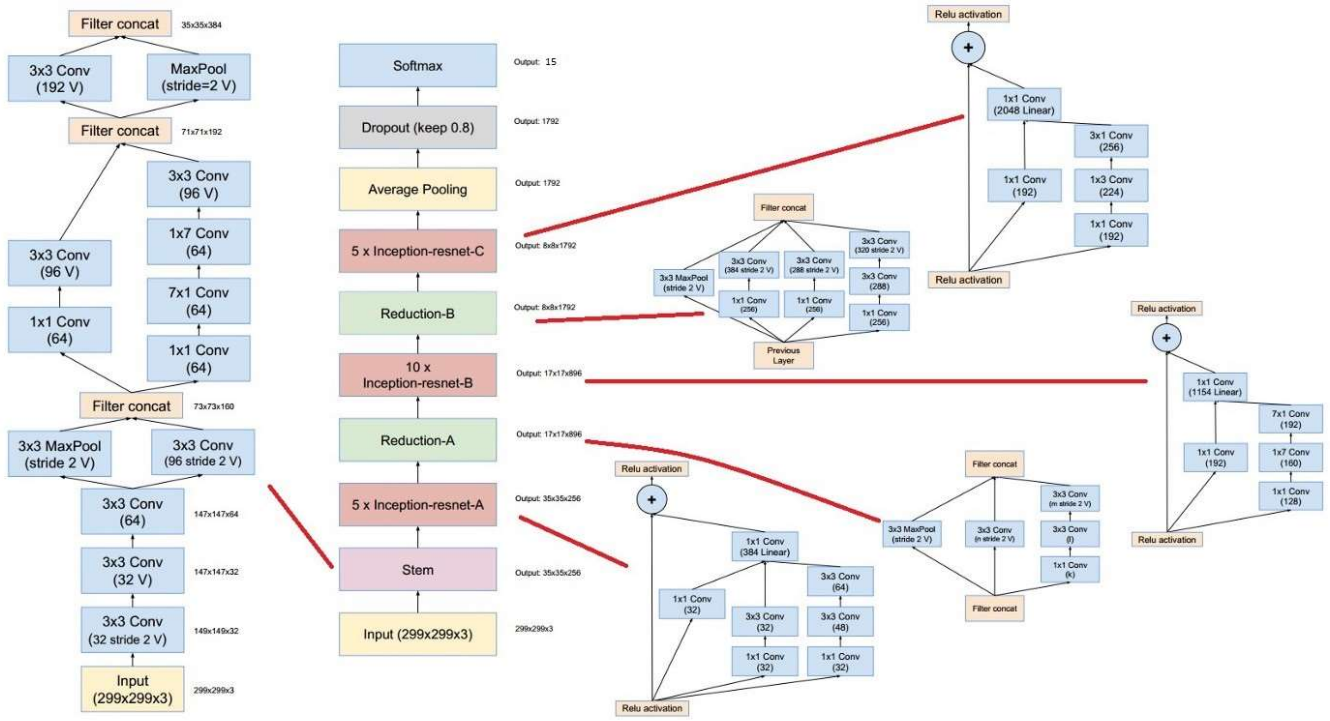
Clinical Impact of Performance:

- False positives do not negatively affect a patient, but can slow down the queue and increase the waiting time for other patients.
- False negatives may increase the waiting time for the patient.

2. Algorithm Design and Function

CNN Architecture:

The classifier is a state of the art CNN structure: inceptionResNetV2:



DICOM Checking Steps:

- Check image modality.

Expected modalities:

```
    * DX - Digital Radiography
    * CT - Computed Tomography
    * MR - Magnetic Resonance
```

- Check body part examined.

Expected value: CHEST

- Check patient position

Expected values: AP or PA

Preprocessing Steps:

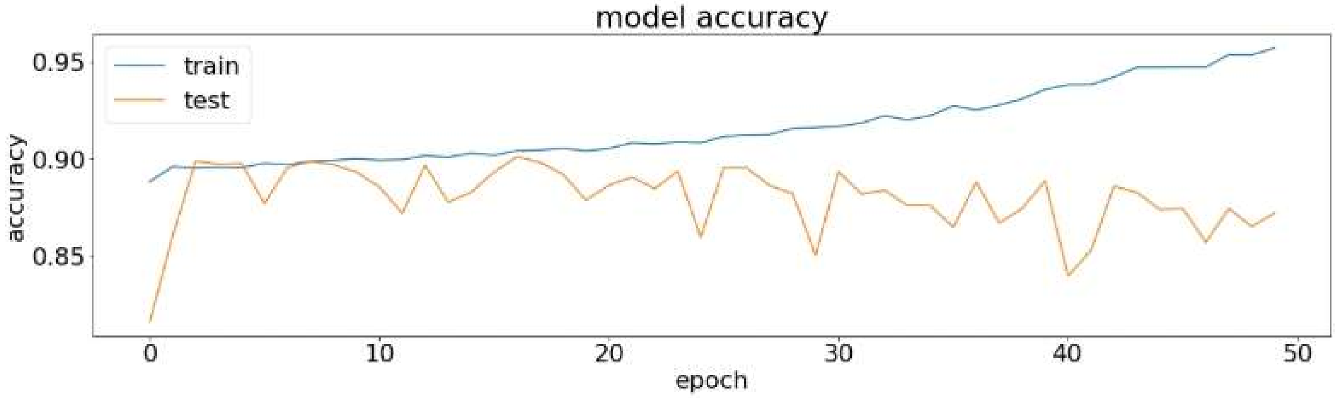
- Resizing image to size 256 x 256
- Divide image array values by 255 to get pixel values between 0 and 1
- Apply sample-wise standartization

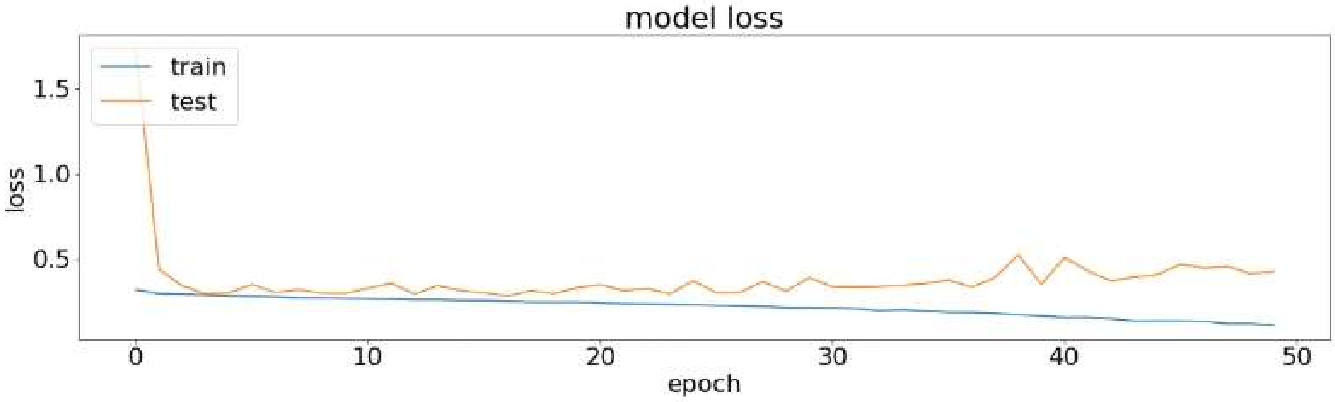
3. Algorithm Training

Parameters:

- Types of augmentation used during training
    - Horizontal flip
    - Horizontal shift range: 5%
    - Vertical shift range: 10%
    - Rotation range: 5 degrees
    - Shear range: 10%
    - Fill mode: reflect
    - Zoom range: 10%
- Batch size: 32
- Optimizer: Adam
- Optimizer learning rate: default (0.001)
- Layers of pre-existing architecture that were frozen: None
- Layers of pre-existing architecture that were fine-tuned: All
- Layers added to pre-existing architecture: Last dense layer size was set to number of classes(15)

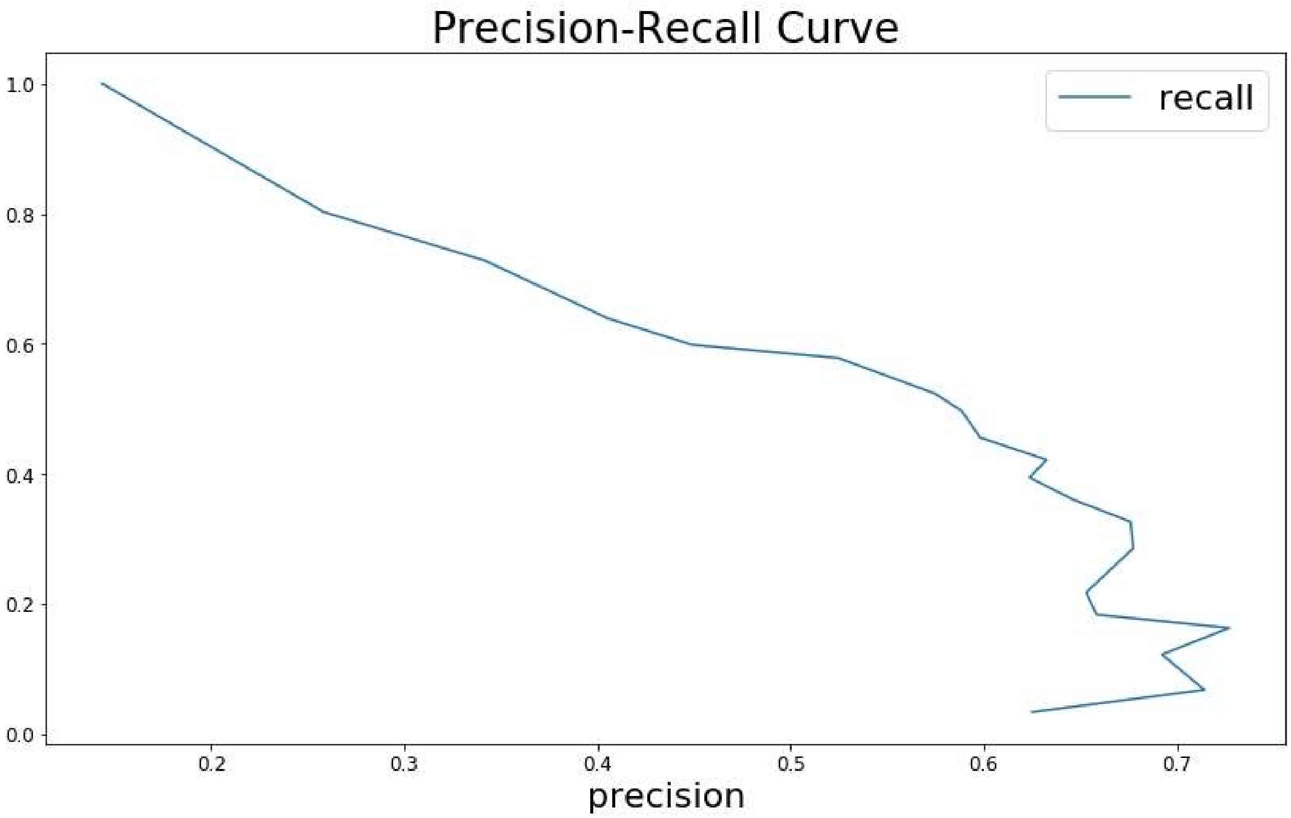
Training performance visualization:





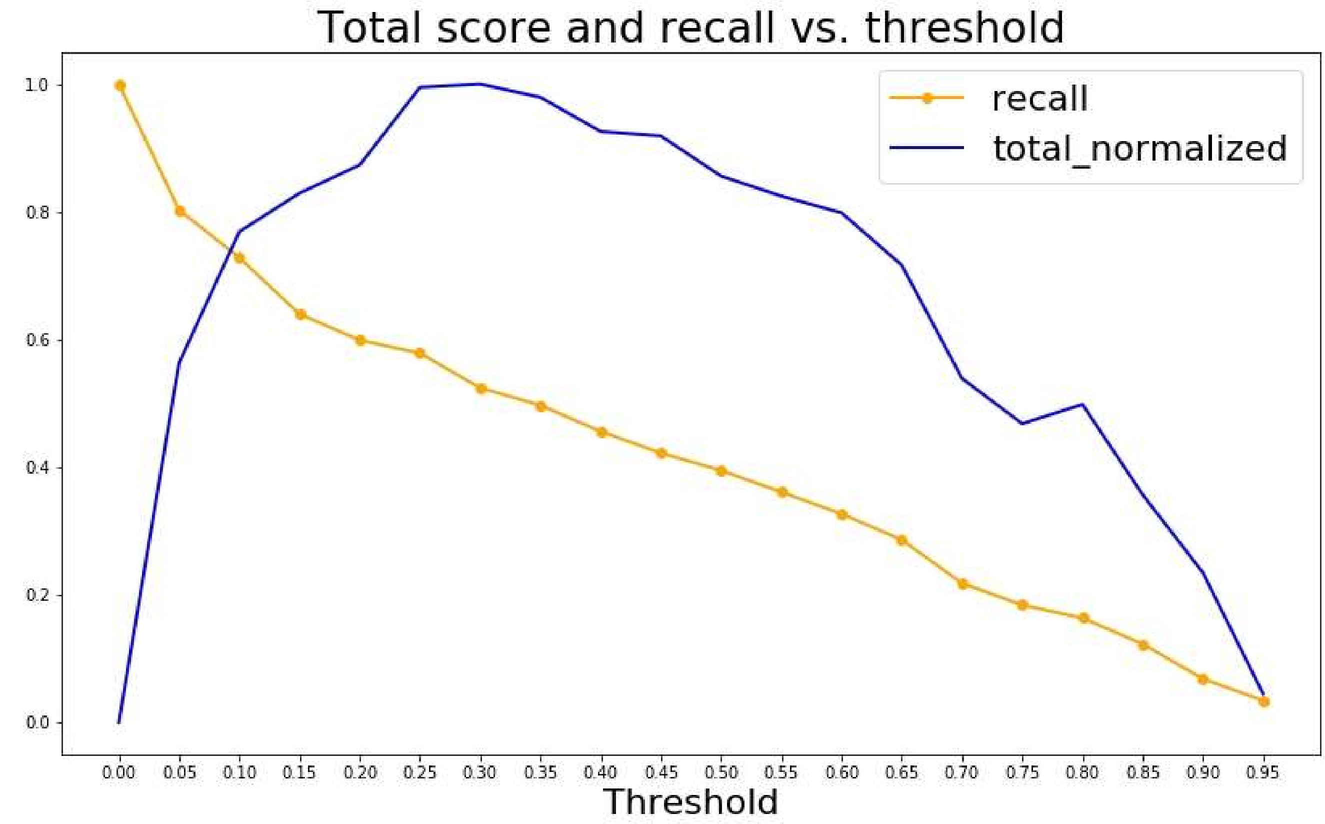
After 17 epochs the model is only overfitting to the training data. The model that is used for inference is saved after 17th epoch and has a highest validation accuracy.

P-R Curve:



Final Threshold and Explanation:

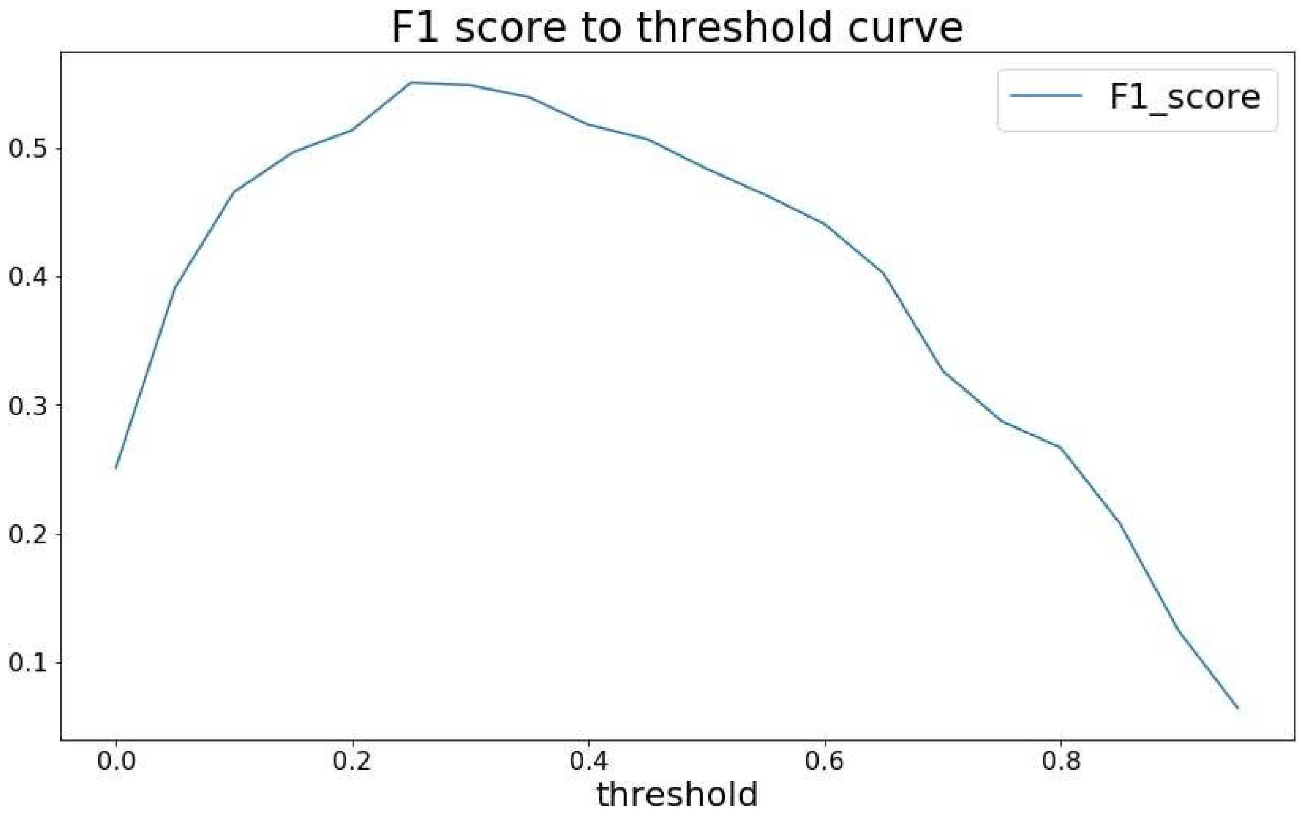
For the worklist prioritization the most important metric is recall. Here is the plot of recall vs threshold and a normalized total score(recall+precision+f1+accuracy) vs threshold:



Based on this plot, we can set two thresholds:

1. If we want to have highest recall while still keeping relatively high total score, we should pick threshold around **0.05**
2. If we want to maximize total score while still keeping relatively high recall, the threshold should be around **0.25**

F1 score is more balanced metric. Here is a plot of F1 score vs threshold:



This plot also indicates 0.25 as the best threshold.

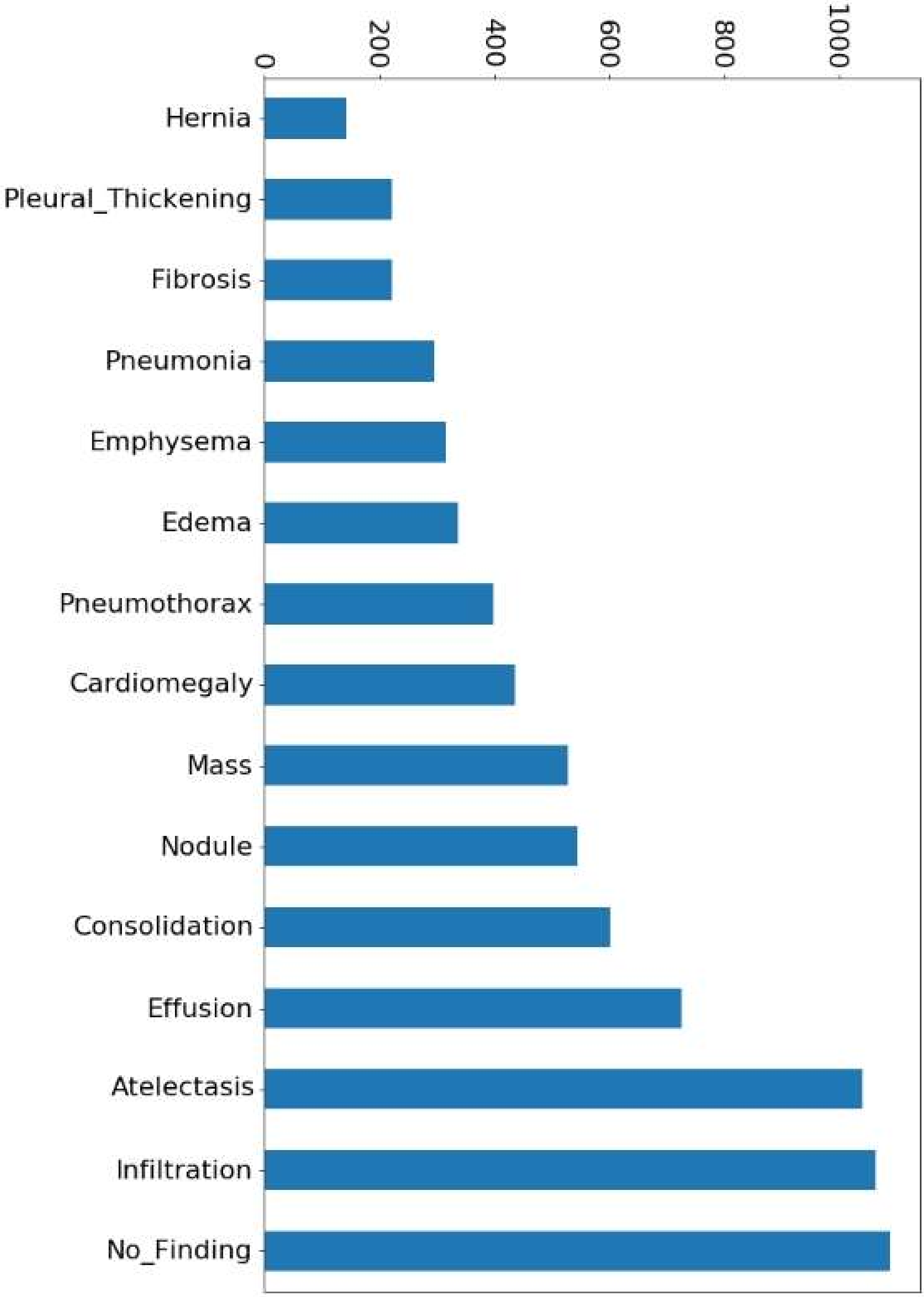
So, there are two thresholds for two different situations:

1. Threshold = 0.25 for high-loaded worklist prioritization tasks. If there are many patients and the hospital has to decide which patients to treat first.
2. Threshold = 0.05 otherwise. If the number of patients is low, but the patient with detected presence of pneumonia should be treated first.

4. Databases

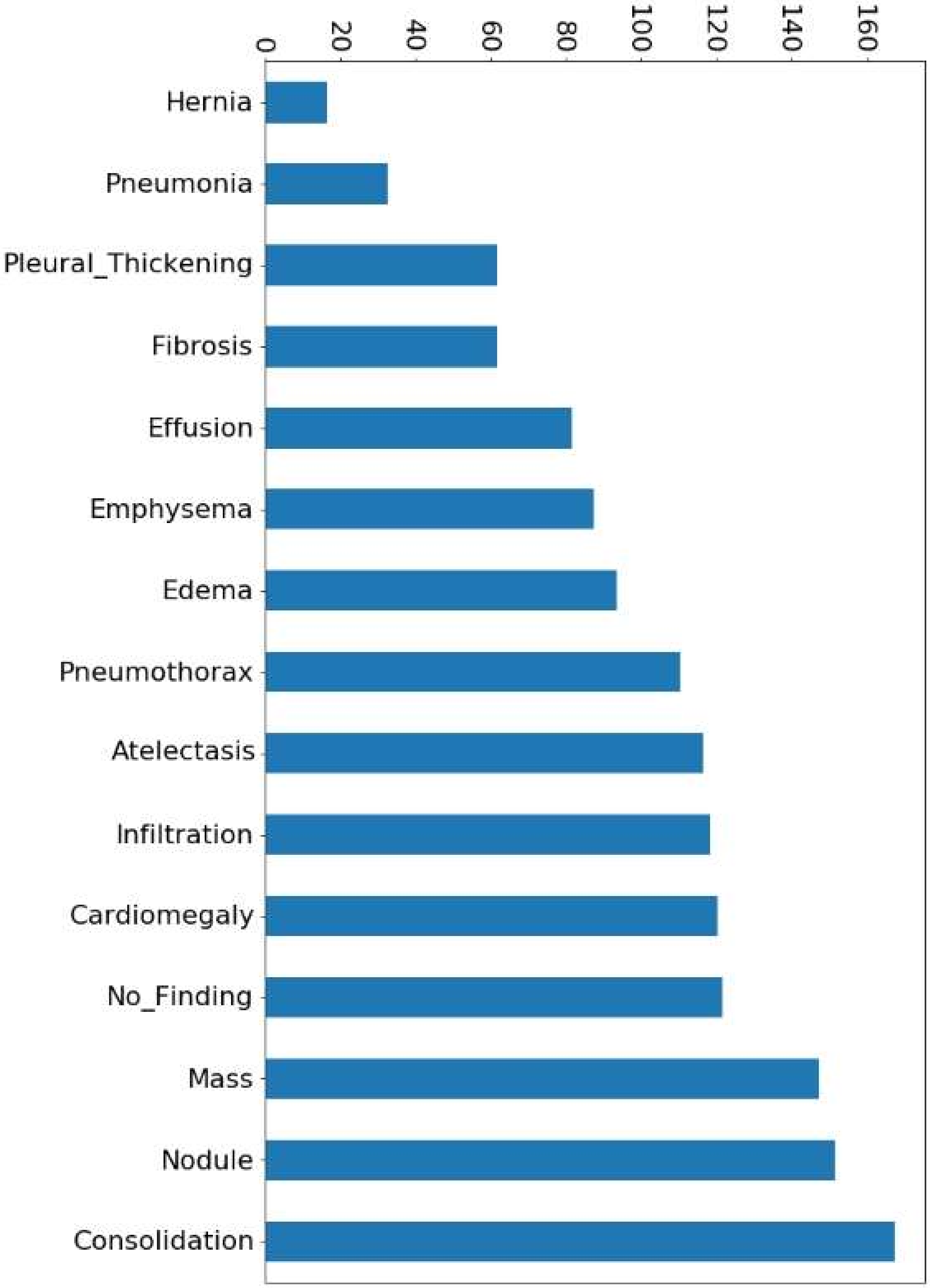
Description of Training Dataset:

Number of samples: 7930 Distribution between desease classes:



Description of Validation Dataset:

Number of samples: 1481 Distribution between desease classes:



5. Ground Truth

The dataset was mined from the PACS system, and the ground truth was obtained from the radiological reports. These are the huge benefits, because it is hard to collect such a large dataset from PACS and have it labeled based on radiological reports.

The limitation is that the ground truth labels were mapped by the NLP algorithm so there would be some erroneous labels but the NLP labelling accuracy is estimated to be >90%.

6. FDA Validation Plan

Patient Population Description for FDA Validation Dataset: Not specified

Ground Truth Acquisition Methodology:

Ideally, we should get a dataset of about 300-1000 chest x-rays from a clinical partner. The dataset should represent the natural distribution of images between all common findings. Every image of the dataset should have a set radiological reports received from a group of experienced radiologists. Then, the silver standard ground truth labels are determined by a voting system across all of the radiologists' labels for each image taking into account the radiologists' experience levels.

Algorithm Performance Standard:

The best metric for this task is F1 score. According to this study (https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1002686#pmed.1002686.s008), the average F1 score for the radiological report in pneumonia detection is between 0.22 and 0.51 Our algorithm's