# FDA Submission

Author: Kaden Strand

Pneumonia Early Detection Application

## Algorithm Description

### 1. General Information

**Intended Use:** This algorithm is intended to be used as an early screening aide to detect Pneumonia for further referral to radiologist.

**Indications for Use:**

- Patient Age: 10 to 75
- Patient Gender: M and F
- Chest X-Ray with view positions: AP and PA

Clinical Setting: The algorithm is intended for integration into the workflow of diagnostic clinics. This is not an emergency detection scenario, so x-ray images may be sent to a remote server for processing. DICOM format following HIPAA rules must be used for all X-Ray images. Each X-Ray DICOM metadata is first verified for correct patient information (see *DICOM Checking Steps*). If the X-Ray passes the DICOM verification step, the image is then pre-processed and input to the machine learning algorithm, yielding a prediction. After the prediction is complete, the result is sent to a radiologist. The radiologist will give the final diagnosis based on their own independent analysis of the image and the prediction given by the algorithm.

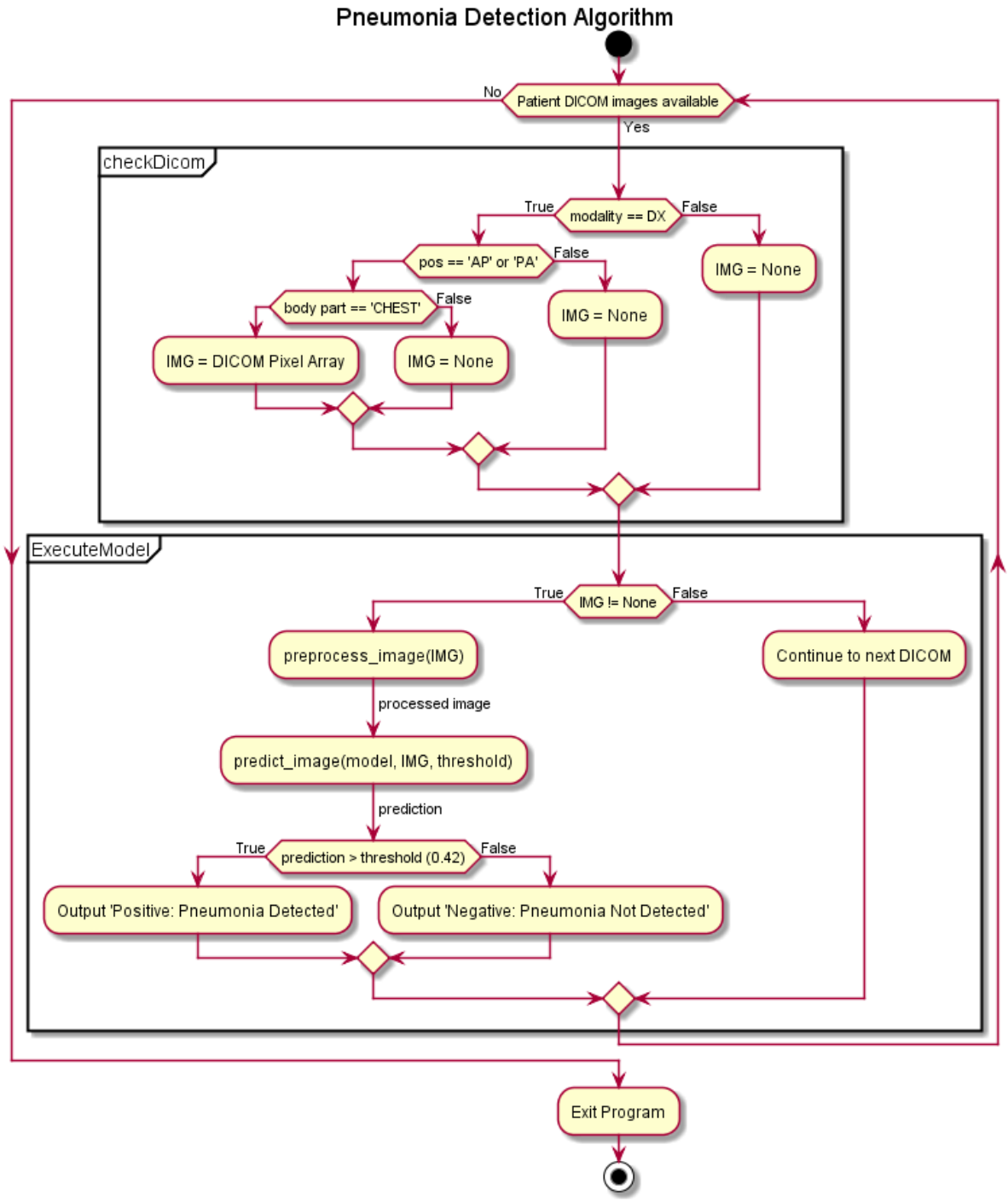
**Device Limitations:**

- The device (algorithm) does not achieve 100% accuracy. Therefore, algorithm results must be only be used as supplemental data to an expert radiologist who will determine the final diagnosis. Algorithm results may not be trusted individually for final diagnosis.
- May require GPU if handling many simultaneous screenings. (Image pre-processing and ResNet50 CNN model execution may be accelerated with GPU)

**Clinical Impact of Performance:** The early detection algorithm attempts to minimize false negative rates in Pneumonia classification, and has a higher likelihood of a false positive reading. All readings, false positives and false negatives, must be reviewed by a radiologist.

- **False Positives** incorrectly indicate the presence of Pneumonia. Further review by a radiologist my waste time and resources for the patient and hospital, but a final diagnosis is confirmed by the radiologist and there is no life-threatening Pneumonia that has gone undetected in the patient.
- **False Negatives** incorrectly indicate no presence of Peneumonia. This is a life-threatening scenario for the patient and it is vital that a radiologist review the result and make a final diagnosis incorporating their own expert analysis. False negative occurances are severe failures of the algorithm, so the algorithm has been optimized to attempt to minimize false negative readings.

2. Algorithm Design and Function




**DICOM Checking Steps:** Verify DICOM metadata:

- *Body Part* must be CHEST
- *Modality* must be DX
- *Patient Position* must be AP or PA

**Preprocessing Steps:** Image Pre-processing:

1. Resize image to 224x224 pixels
2. Convert from grayscale to RGB (Image dimensions: 224x224x3 pixels)
3. Apply ResNet image preprocessing which recenters pixel values around zero (subtract pixel mean) but does not normalize range (do not divide by pixel standard deviation)

**CNN Architecture:** The ResNet50 model architecture is used as a basis for the Pneumonia detection model, with a flatten layer and two additional fully connected layers appended. Dropout and batch normalization layers are also included to improve training.

## 3. Algorithm Training

**Parameters:**

- Types of augmentation used during training:
  - horizontal flip,
  - height/width shift with range 0.1
  - rotation range 20
  - shear range 0.1
  - zoom range 0.1
- Batch size
  - 32 images
- Optimizer learning rate
  - Adam optimizer, learning rate 1e-4
- Layers of pre-existing architecture that were frozen
  - All layers except final 2D convolutional layer and average pooling were frozen
- Layers of pre-existing architecture that were fine-tuned
  - Final 2D convolutional layer 'conv5_block3_3_conv' and following layers were fine-tuned:

```
Layer (type)                 Output Shape              Param #
=================================================================
conv5_block3_3_conv (Conv2D)    (None, 7, 7, 2048)   1050624
_________________________________________________________________
conv5_block3_3_bn (BatchNormali (None, 7, 7, 2048)   8192
_________________________________________________________________
conv5_block3_add (Add)          (None, 7, 7, 2048)   0


_________________________________________________________________
conv5_block3_out (Activation)   (None, 7, 7, 2048)   0
_________________________________________________________________
avg_pool (GlobalAveragePooling2 (None, 2048)         0
=================================================================
```

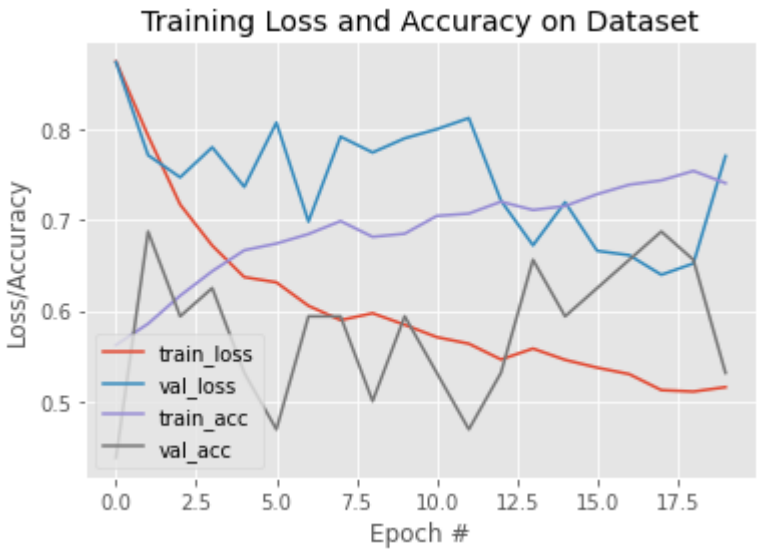
- Layers added to pre-existing architecture
  - Flatten, dropout, dense, batch normalize, dense

Model Architecture: "Pneumonia_Detection_Model"

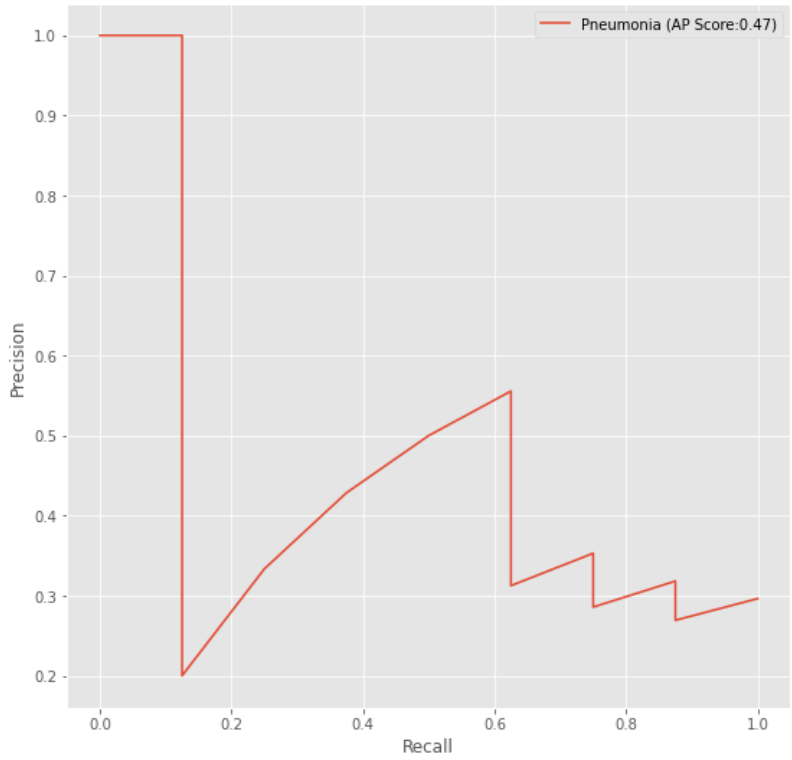
```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
```

```
ResNet50 (Model)                 (None, 7, 7, 2048)        23587712
_________________________________________________________________
flatten_1 (Flatten)              (None, 100352)            0
_________________________________________________________________
dropout_1 (Dropout)              (None, 100352)            0
_________________________________________________________________
dropout_2 (Dropout)              (None, 100352)            0
_________________________________________________________________
dense_1 (Dense)                  (None, 64)                6422528
_________________________________________________________________
batch_normalization_1 (Batch     (None, 64)                256
_________________________________________________________________
activation_1 (ReLu Activation)       (None, 64)            0
_________________________________________________________________
dense_2 (Dense, Sigmoid Activation) (None, 1)              65
=================================================================
Total params: 30,010,561
Trainable params: 7,477,441
Non-trainable params: 22,533,120
```

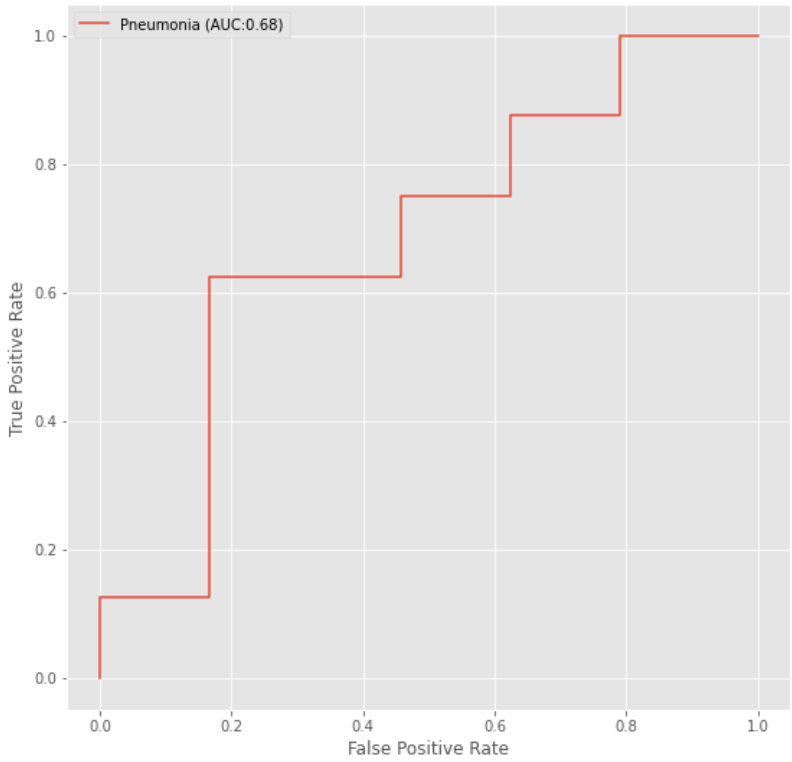
Training Performance:




Precision-Recall Curve:




ROC Curve:




**Final Threshold and Explanation:** It is preferred the the model performance favors **recall**, so that the model has a fewer number of false negatives. All readings, false positives and false negatives, must be reviewed by a radiologist.

- **False Positives** incorrectly indicate the presence of Pneumonia. Further review by a radiologist my waste time and resources for the patient and hospital, but a final diagnosis is confirmed by the radiologist and there is no life-threatening Pneumonia that has gone undetected in the patient.

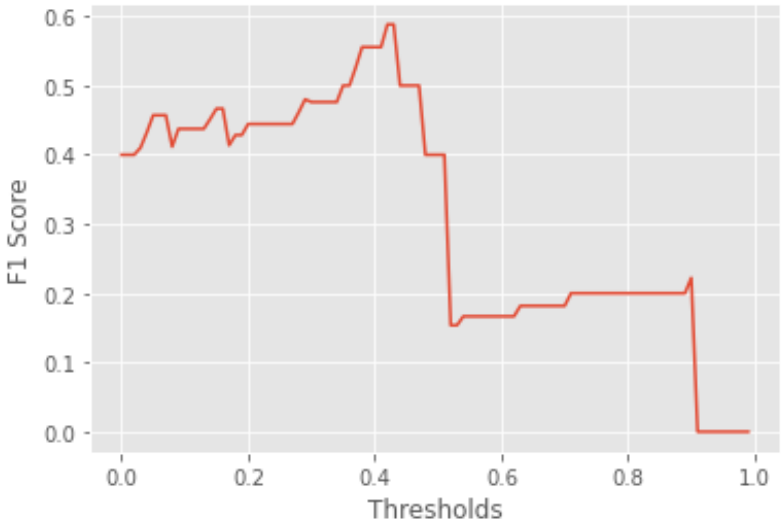
- **False Negatives** incorrectly indicate no presence of Peneumonia. This is a life-threatening scenario for the patient and it is vital that a radiologist review the result and make a final diagnosis incorporating their own expert analysis. False negative occurances are severe failures of the algorithm, so the algorithm has been trained to attempt to minimize false negative readings.

The classification threshold is determined as follows: For each threshold in the range of thresholds from 0.0 to 1.0 at increments of 0.01:

- Make predictions using threshold and the validation set, and calculate and save the F1 score
- Save the best (maximum) F1 score and associated threshold.

With this method, we choose a final classification threshold of 0.42. At a threshold of 0.42, F1 score is maximized with a value of 0.588, Recall is favored with a value of 0.625, and Precision maintains a value of 0.56 when tested against the validation set.

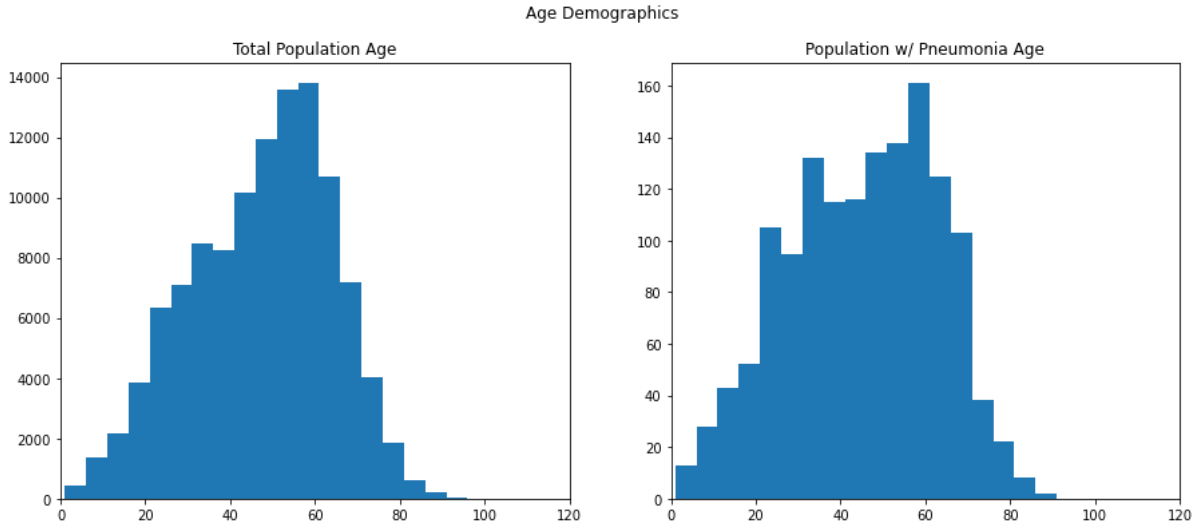
F1 Score vs Threshold is shown below:



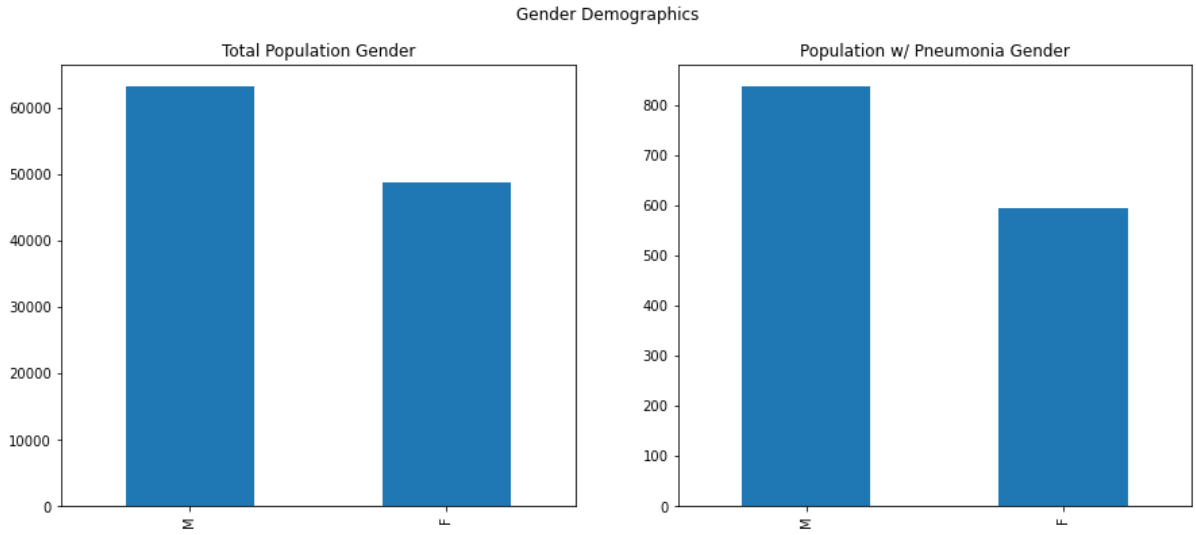

## 4. Databases

In the NIH master dataset, a total of 1431 images samples present with Pneumonia. The training and validation datasets are developed by splitting the NIH dataset and stratifying by the presence of Pneumonia. The rate of Pneumonia in the total population (NIH dataset) is around 1%. All samples cases with Pneumonia are included within Train and Validation sets. The age and gender demographics of the total population and the subset population with Pneumonia are similar, so the training datasets follow the demographics of the total population without further stratification.

Age Demographics:




Gender Demographics:




**Description of Training Dataset:** The training dataset contains 80% of the available samples with Pneumonia, and holds an equal number (50/50 split) of image samples with and without Pneumonia sampled from the NIH dataset.

The training set size is: (1145 samples with Pneumonia + 1145 samples without Pneumonia) = 2290 total samples

**Description of Validation Dataset:** The validation dataset contains the remaining 20% of the available samples with Pneumonia, and holds a 20/80 split of image samples with and without Pneumonia sampled from the NIH dataset, e.g. the probability of randomly sampling a non-Pneumonia case from the validation dataset it 80%. This 20/80 split is designed to mimic a clinical setting where it would be expected that roughly 20% of the scans may present with Pneumonia, even though the total population rate of Pneumonia is far lower (around 1% in NIH dataset). No samples from the validation dataset are included in the training dataset.

The validation set size is: (286 samples with Pneumonia + 1144 samples without Pneumonia) = 1430 total samples

## 5. Ground Truth

An NIH chest x-ray dataset is used for ground truth data. This dataset was not specifically acquired for pneumonia analysis, so the percentage is samples in this dataset is very low:
Samples with Pneumonia: 1431
Samples without Pneumonia: 110689
Fraction: (1431 / 110689) = 0.013

The ground truth labels are developed using Natural Language Processing software, which may incorrectly categorize some image labels.

Additionally, the NIH dataset samples show many comorbidities for Pneumonia, primarily Infiltraion and Edema. Since the model is designed to predict any occurance of pneumonia including in cases where Pneumonia is present alongside other diseases, this ground truth dataset is acceptible. However, the model would likely be improved by training on a dataset with many more cases of Pneumonia present and without the presense of comorbidities.

## 6. FDA Validation Plan

**Patient Population Description for FDA Validation Dataset:** The ideal FDA validation dataset would follow the gender, age, view position, and findings demographics of the NIH chest x-ray dataset and meet the algorithm indications for use:

- Patient Age: 10 to 75
- Patient Gender: M and F
- Chest X-Ray with view positions: AP and PA

Additionally, Pneumonia would present at a rate of 20% in the FDA validation dataset, to match the expected clinical setting where Pneumonia is suspected at a higher rate than in the general population.

**Ground Truth Acquisition Methodology:** A silver standard approach in which multiple radiologists review and assign ground truth labels for presence of Pneumonia is preferred, since it is a visually challenging task even for trained radiologists.

**Algorithm Performance Standard:** According to Rajpurkar et al., radiologists average an F1 score of 0.387. The proposed algorithm achieves an F1 score of 0.588 on the validation dataset, which exceeds the radiologist performance standard. Additionally, the algorithm is optimized for recall for use in early detection of Pneumonia.