FDA Submission

Your Name: Thomas Johnson

Name of your Device: Pneumonia DX Binary Classification Detection Algorithm

Algorithm Description

1. General Information

Intended Use Statement:

The algorithm is intended to assist radiologists and healthcare providers in identifying findings in chest radiographs that are suggestive of pneumonia in clinical settings such as hospitals and outpatient clinics. The algorithm is not intended to replace clinical judgment and must be used in conjunction with other diagnostic evaluations.

Indications for Use:

The algorithm is indicated for use in the detection of pneumonia in chest radiographs in both male and female patients between the ages of 1 and 90, with no existing conditions or one or more existing conditions in the following list:

- Atelectasis
- Cardiomegaly
- Consolidation
- Edema
- Effusion
- Emphysema
- Fibrosis
- Hernia
- Infiltration
- Mass
- Nodule
- Pleural_Thickening
- Pneumothorax

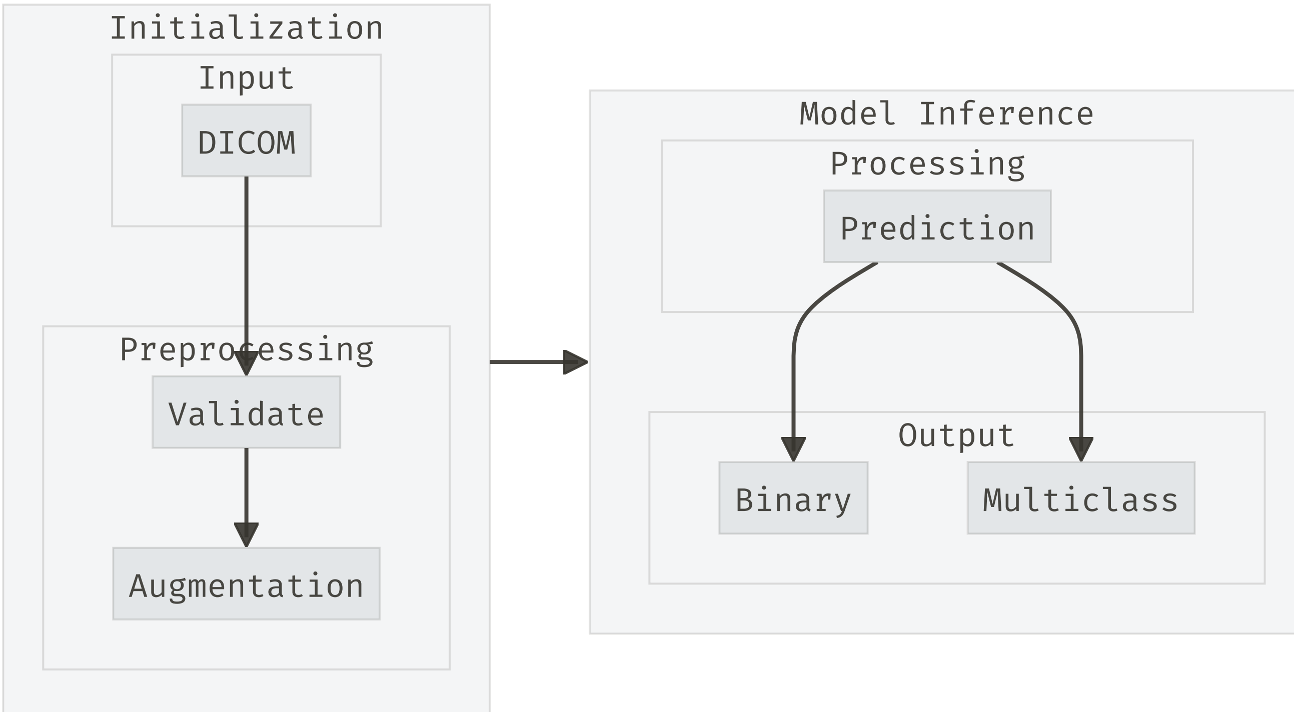
Device Limitations:

The algorithm requires a computing system with a minimum of 4 CPU cores, 16 GB ram, and a GPU to run. It is not intended to be used on mobile devices, and adequate thermal cooling must be present to ensure the GPU does not overheat. It is strongly advised to perform regular maintenance and inspection of the computing system to ensure optimal performance.

Clinical Impact of Performance:

The algorithm could enable faster identification of pneumonia in chest radiographs, leading to earlier treatment and improved patient outcomes. It could also help reduce the workload of radiologists by providing an initial assessment of chest radiographs, allowing radiologists to focus on more complex cases.

2. Algorithm Design and Function



DICOM Checking Steps:

During the DICOM checking step, the algorithm checks the DICOM header to ensure that the image is a chest radiograph (DX), the patient position is either AP (Anterior-Posterior) or PA (Posterior-Anterior), the body part examined is CHEST, and that the patient is between the ages of 1 and 90. If the image does not meet these criteria, the algorithm will output an error message and will not proceed with the classification.

Preprocessing Steps:

The algorithm preprocesses the input chest radiograph by resizing the image to 224×224 pixels, normalizing the pixel values to be between 0 and 1, and applying data augmentation techniques to increase the diversity of the training data.

CNN Architecture:

The algorithm uses an ImageNet pre-trained VGG16 model as the base architecture and adds additional dense layers on top of the base model. The VGG16 model, for reference, is architected as follows:

```
_________________________________________________________________
 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
 flatten (Flatten)           (None, 25088)             0
 fc1 (Dense)                 (None, 4096)              102764544
 fc2 (Dense)                 (None, 4096)              16781312
 predictions (Dense)         (None, 1000)              4097000
=================================================================
Total params: 138,357,544
Trainable params: 138,357,544
Non-trainable params: 0
_________________________________________________________________
```

3. Algorithm Training

Parameters:

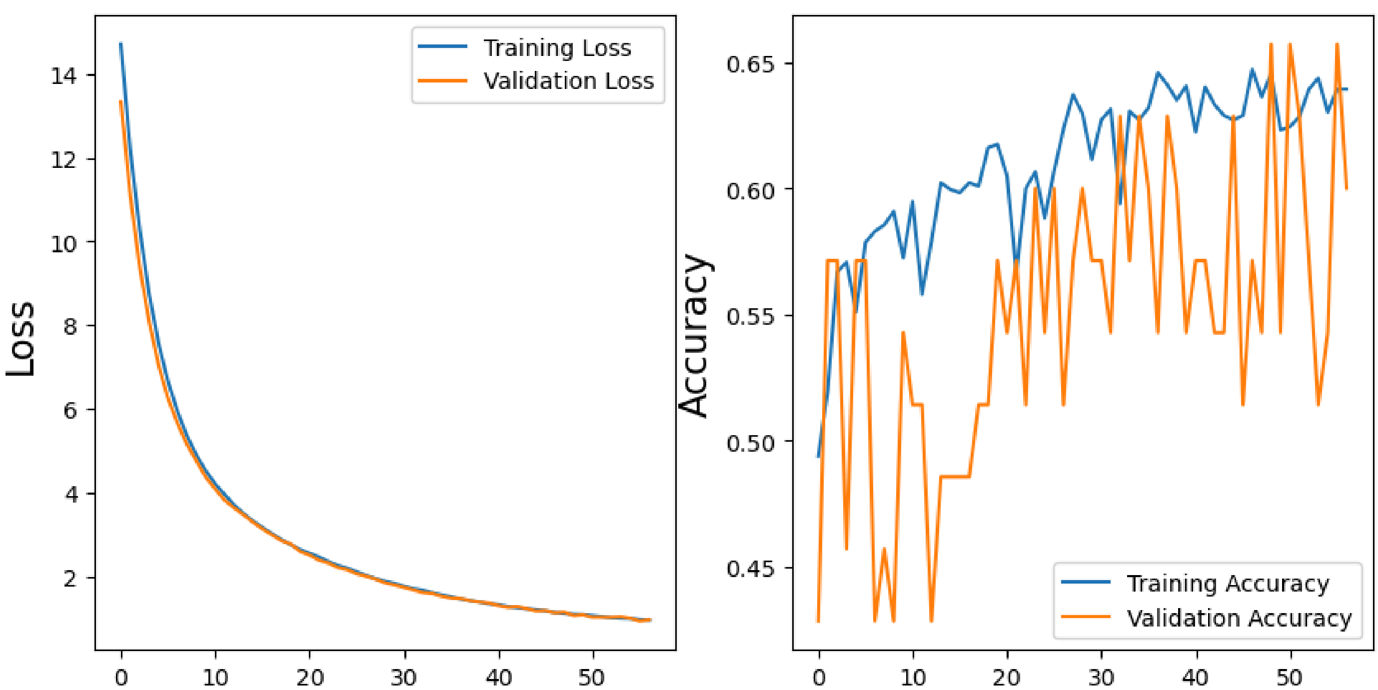
- Types of augmentation used during training:
  - Rotation
  - Horizontal flip
  - Height shift
  - Width shift
  - Shear
  - Zoom
- Batch size
  - 143 (Training set size / 16)
- Optimizer (Adam) learning rate
  - 0.0001
- Layers of pre-existing architecture that were frozen
  - All convolutional layers of the VGG16 model
- Layers of pre-existing architecture that were fine-tuned
  - 5 final layers of the VGG16 model
- Layers added to pre-existing architecture
  - flatten (Flatten)
  - dense (Dense 1024)
  - dropout (Dropout 0.5)

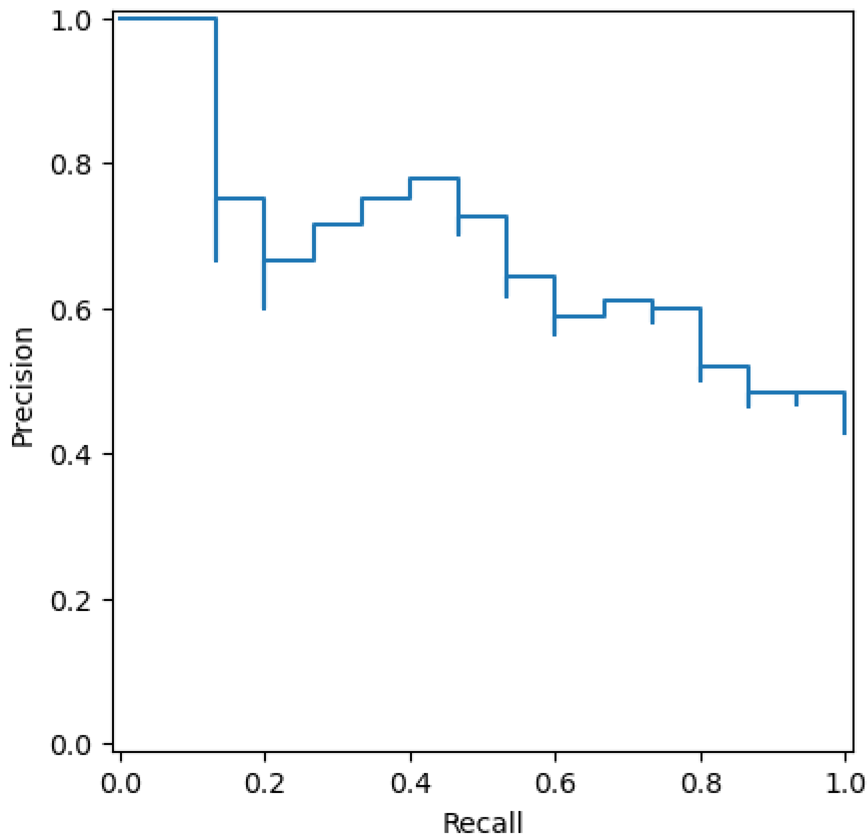
- dense_1 (Dense 512)
- dense_2 (Dense 15, for multi-class classification based on labels)
- dense_3 (Dense 1, for binary classification)

Algorithm architecture:

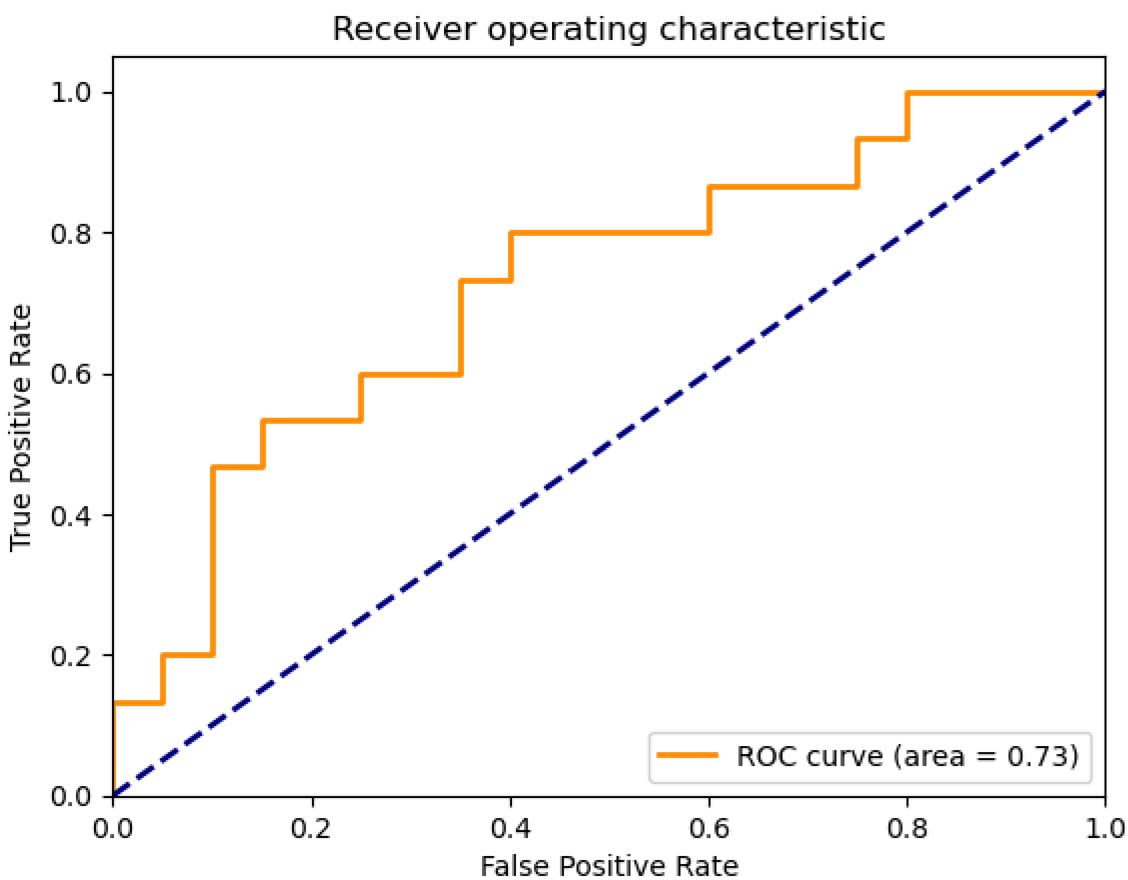
```
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 model (Functional)          (None, 7, 7, 512)         14714688
 flatten (Flatten)           (None, 25088)             0
 dense (Dense)               (None, 1024)              25691136
 dropout (Dropout)           (None, 1024)              0
 dense_1 (Dense)             (None, 512)               524800
 dense_2 (Dense)             (None, 15)                7695
 dense_3 (Dense)             (None, 1)                 16
=================================================================
Total params: 40,938,335
Trainable params: 28,583,455
Non-trainable params: 12,354,880
_________________________________________________________________
```

Optimizer : Adam

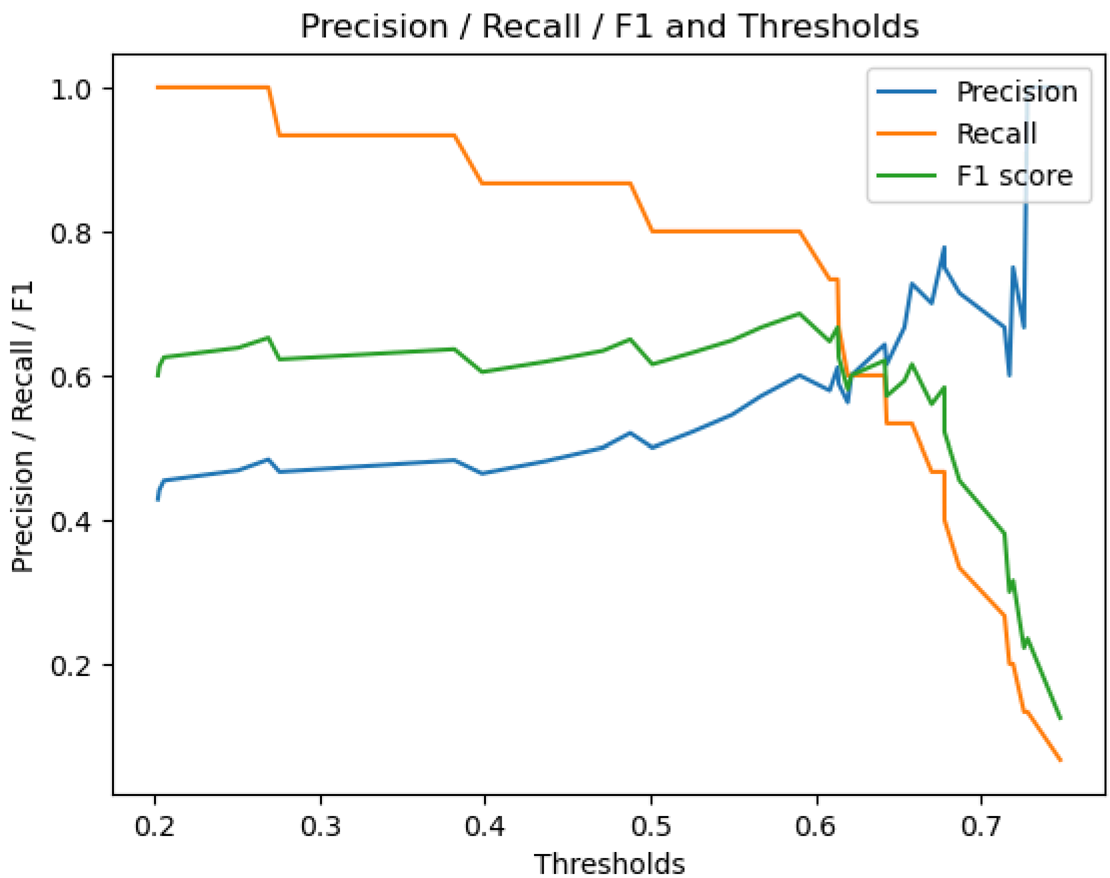




Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0



Receiver operating characteristic
True Positive Rate
False Positive Rate
ROC curve (area = 0.73)
0.0
0.2
0.4
0.6
0.8
1.0



Final Threshold and Explanation:

The final threshold for the algorithm was set at 0.6. This threshold was chosen based on the precision-recall curve of the algorithm on the validation dataset. The threshold was selected to maximize the F1 score of the algorithm, which balances the trade-off between precision and recall. A threshold of 0.6 was found to provide a good balance between precision and recall, resulting in an F1 score of 0.65 on the validation dataset.

Classification report:

```
              precision    recall  f1-score   support

           0       0.75      0.60      0.67        20
           1       0.58      0.73      0.65        15

    accuracy                           0.66        35
   macro avg       0.66      0.67      0.66        35
weighted avg       0.68      0.66      0.66        35
```

Confusion matrix:

```
[[12  8]
 [ 4 11]]
```

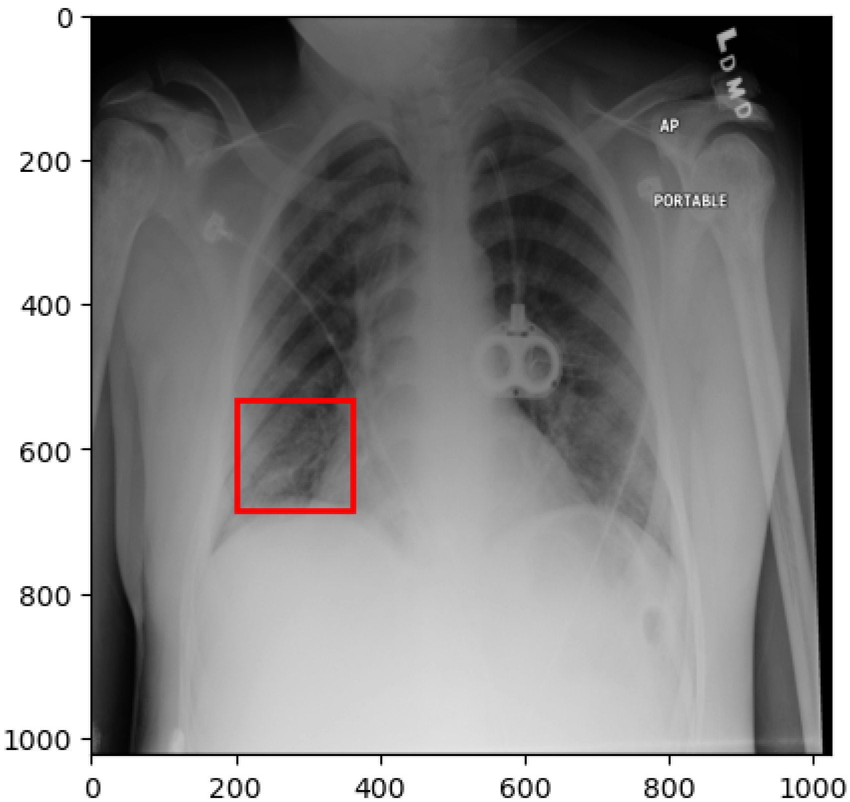
4. Databases

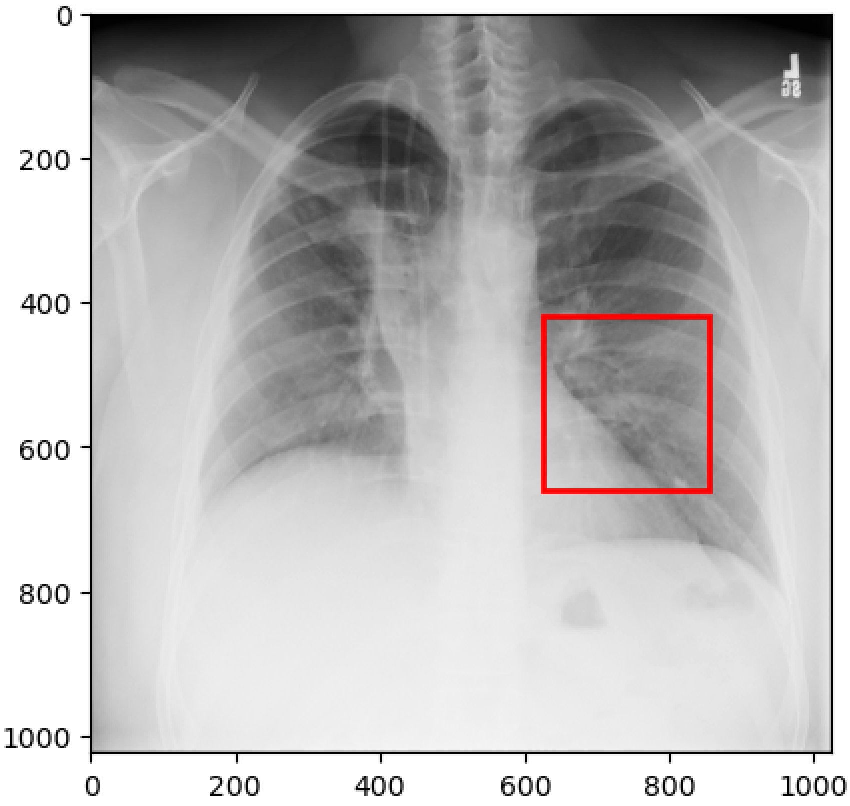
Description of Training Dataset:

The training dataset consists of 2290 chest radiographs from the NIH Chest X-ray dataset. The full NIH Chest X-ray Dataset is comprised of 112,120 X-ray images with disease labels from 30,805 unique patients. To create these labels, the authors used Natural Language Processing to text-mine disease classifications from the associated radiological reports. The labels are expected to be >90% accurate and suitable for weakly-supervised learning. The original radiology reports are not publicly available but you can find more details on the labeling process in this Open Access paper: "ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases." (Wang et al.)

The algorithm was trained on a subset of these images labeled as either "pneumonia present" or "pneumonia absent" based on the presence of pneumonia in the radiograph. The dataset is balanced, with an equal number of positive and negative cases.

Examples of plots of images from the training dataset, with bounding boxes for pneumonia regions:





Description of Validation Dataset:

The validation dataset consists of 572 chest radiographs from the NIH Chest X-ray dataset. The images are labeled as either "pneumonia present" or "pneumonia absent" based on the presence of pneumonia in the radiograph. The dataset is balanced, with an equal number of positive and negative cases.

5. Ground Truth

The ground truth for the algorithm was established by using the labels provided in the NIH Chest X-ray dataset. The labels were created by applying Natural Language Processing techniques to extract disease classifications from the corresponding radiological reports. With an anticipated accuracy exceeding 90%, the labels are deemed appropriate for weakly-supervised learning, and suitable for usage as ground truth in training and validating the algorithm. However, it should be noted that a limitation of the ground truth is that it is based on text-mined labels from radiology reports, and may not always be 100% accurate.

6. FDA Validation Plan

Patient Population Description for FDA Validation Dataset:

The FDA validation dataset will consist of the following populations:

- Male and female patients between the ages of 1 and 90
- Patients with no existing conditions
- Patients with one or more existing conditions from the following list:
  - Atelectasis
  - Cardiomegaly
  - Consolidation

    - Edema
    - Effusion
    - Emphysema
    - Fibrosis
    - Hernia
    - Infiltration
    - Mass
    - Nodule
    - Pleural_Thickening
    - Pneumothorax
- Patient body positions of either Anterior-Posterior (AP) or Posterior-Anterior (PA)
- An image type of chest radiograph (DX)
- Body part examined is CHEST

Ideally, the FDA validation dataset will include a diverse range of patients to ensure the algorithm's generalizability across different demographics and conditions. In addition, the dataset should represent real-world prevalence rates of pneumonia in chest radiographs. Sufficient sample sizes should be included to ensure statistical significance in the evaluation of the algorithm's performance.

Ground Truth Acquisition Methodology:

The ground truth for the FDA validation dataset will be established by expert radiologists who will independently review the chest radiographs and provide a binary classification of either "pneumonia present" or "pneumonia absent". The ground truth labels will be established by a consensus of the radiologists' assessments, which can be considered a silver standard for the purposes of the FDA validation.

Algorithm Performance Standard:

We reference the CheXNet paper (Rajpurkar et al., 2017) as a performance standard for the algorithm. The CheXNet paper reports an F1 score of 0.435 for the detection of pneumonia in chest radiographs. We aim to achieve similar or better performance metrics for our algorithm, and advise using the F1 score as the primary metric for evaluating the algorithm's performance.