FDA Submission

Your Name: Xinyue Yao

Name of your Device: PneuDetect

Algorithm Description

1. General Information

Intended Use Statement:
Assisting a radiologist in detecting pneumonia in chest x-ray scans.

Indications for Use:

- PneuDetect is intended for use in detecting the presence of pneumonia in patients' x-ray scans stored in DICOM format.
- Intended population: female and male patients of ages 1-95.
- Chest X-ray image should be taken from either 'PA' or 'AP' viewing position.

Device Limitations:

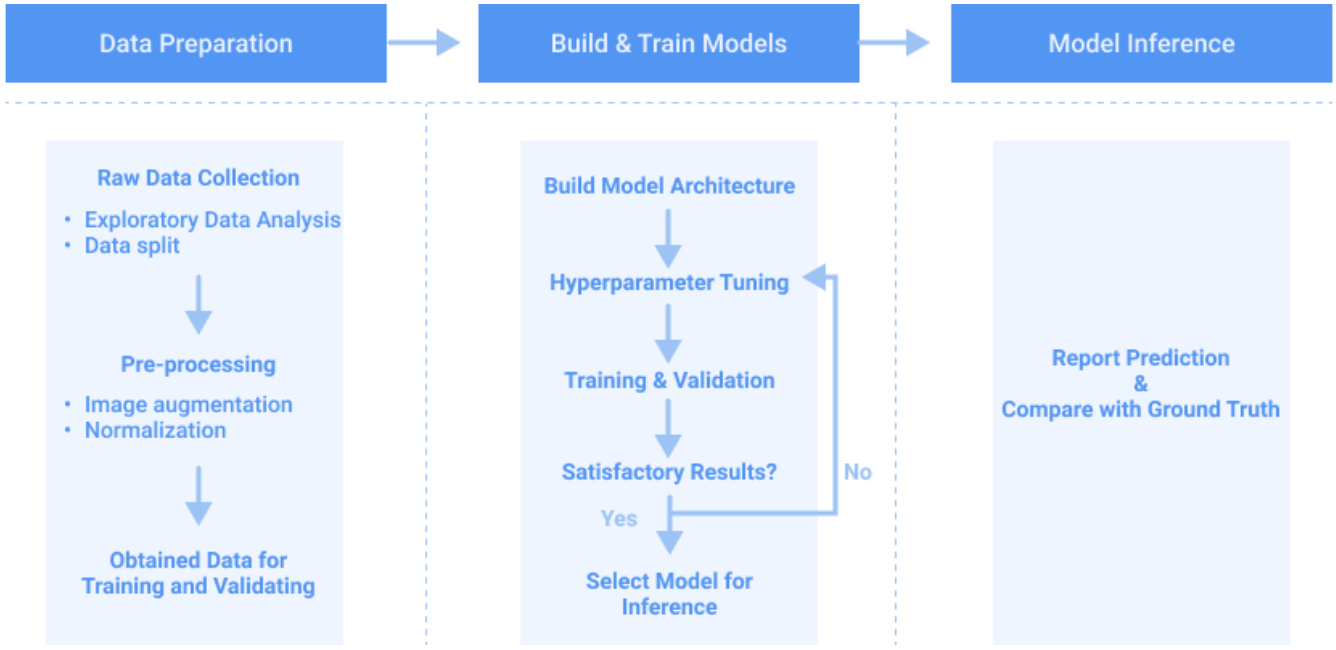
- Computational: Running PneuDetect on a GPU is preferred, although classification can be made using a standard CPU despite its much slower computing speed.
- Algorithmic:
    - PneuDetect has a high recall rate and low precision rate, which means PneuDetect is relatively more confident in negative results than positive ones.
    - PneuDetect may confuse the presence of pneumonia with other labels that present similar intensity distribution. These include Infiltration, Effusion, Mass, and Pleural Thickening. In the case of co-occurrence, PneuDetect may mistake these labels with pneumonia.

Clinical Impact of Performance:
PneuDetect can only assist a diagnosis and should never be used as the primary interpretation. PneuDetect may have an increased likelihood of producing false positive or false negative. False-positive may lead to unnecessary additional examinations for patients, while false negative may lead to severe consequences to a patient's safety. Thus, users need to review and validate original x-ray scans concurrently before coming to a final clinical conclusion of a case.

The threshold used in model validation and inference is favoring recall, which will try to lower the false negative. This allows PneuDetect to be more confident in predicting negative results. Therefore, PneuDetect serves a better role as a screening tool.

2. Algorithm Design and Function

Figure 1a. Algorithm Flow Diagram



DICOM Checking Steps:

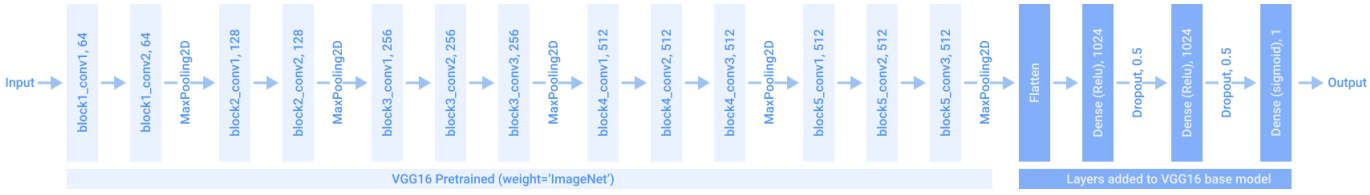
- The examined body part is the chest ('CHEST')
- The image type (modality) is digital radiography ('DX').
- The view position is either 'AP' or 'PA'.

Preprocessing Steps:

- Resize images to meet the image size requirement of the VGG16 model.
- Normalize the image such that the mean is zero and the standard deviation is 1.

CNN Architecture:
The model uses VGG16 as the base model, with the top fully connected layers being cut off. Dense layers and dropout layers were added for use in training to classify pneumonia. The architecture of the model is shown as below:

Figure 1b. CNN Architecture

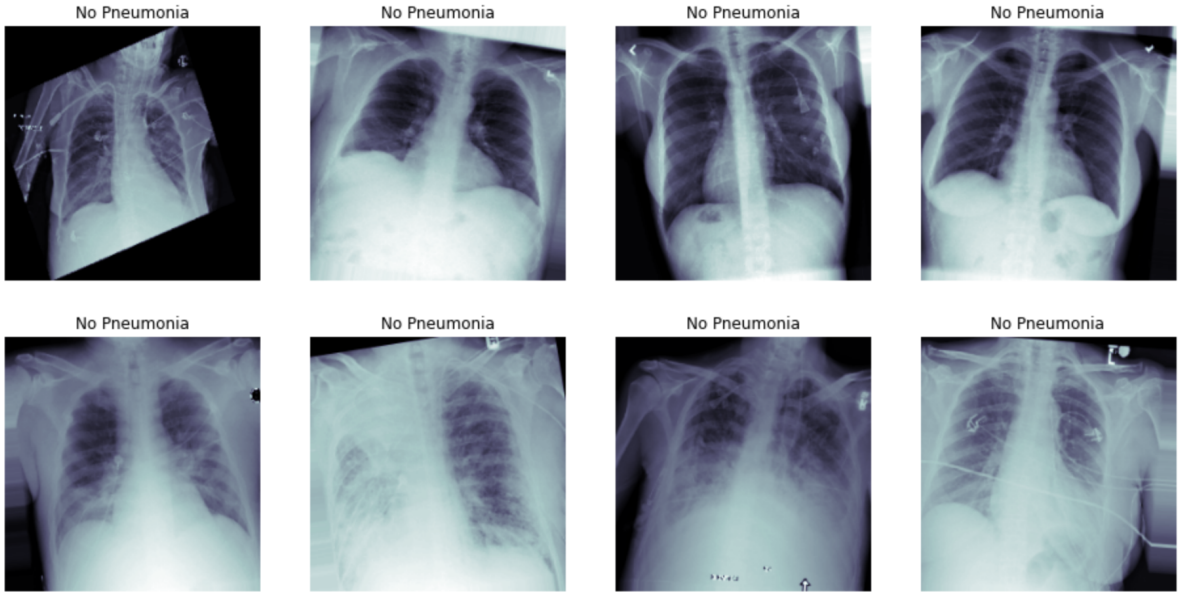


3. Algorithm Training

Parameters:

- Types of augmentation used during training:
  - rescale: 1./255.0
  - horizontal_flip: True
  - vertical_flip: False
  - rotation_range: degree range [0,10]
  - width_shift_range: possible values in the interval [-0.1, 0.1]
  - shear_range: 0.05 (counter-clockwise degrees)

- zoom_range: [1-0.05. 1+0.05]
- brightness_range: (0.3, 0.9)
- preprocessing_function=preprocess_input (data normalization in keras) Examples of augmented images:



- Batch size: 16
- Optimizer learning rate: 1e-4
- Loss function: binary cross-entropy loss.
- Layers of pre-existing architecture that were frozen: the first 10 convolution layers and max pooling layers.
- Layers of pre-existing architecture that were fine-tuned: the last three convolution layers and the last max pooling layer.
- Layers added to pre-existing architecture: one flatten layer, three dense layers, and two dropout layers are added.

Model Metrics

Figure 2a. Model training history (loss and accuracy)

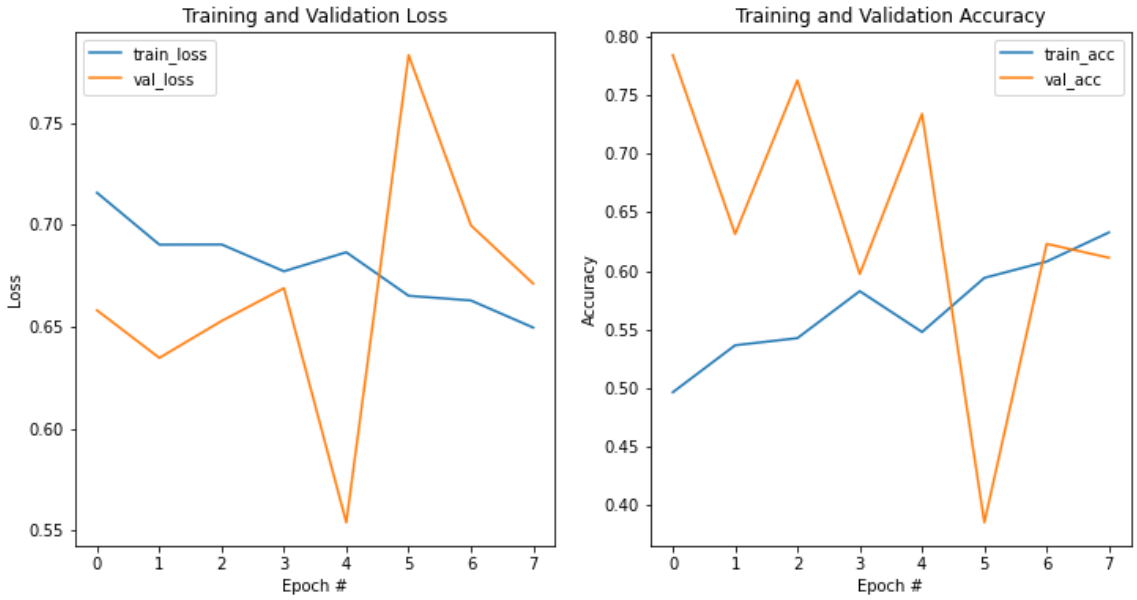


Figure 2b. ROC Curve with AUROC score used in choosing threshold

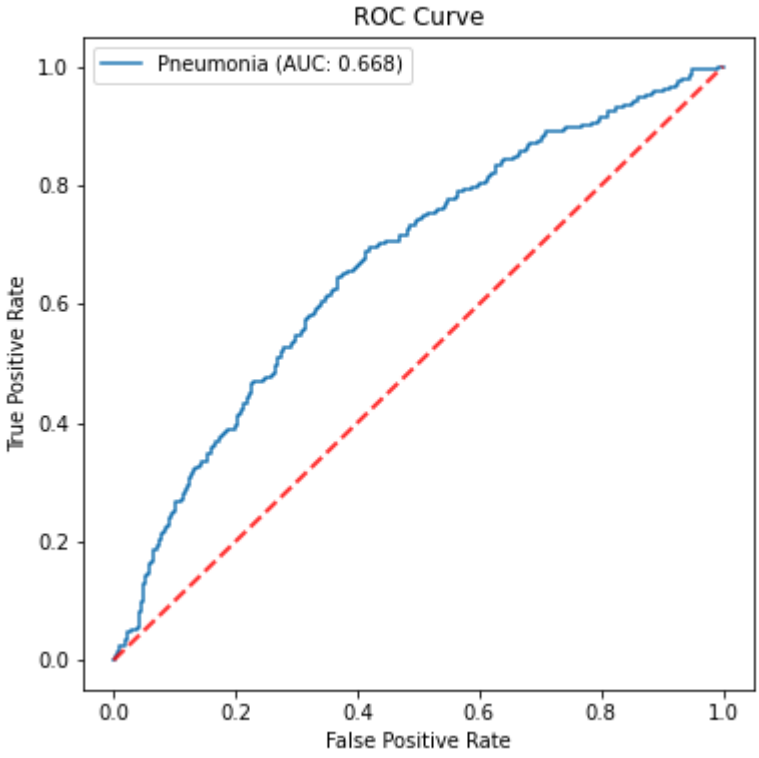
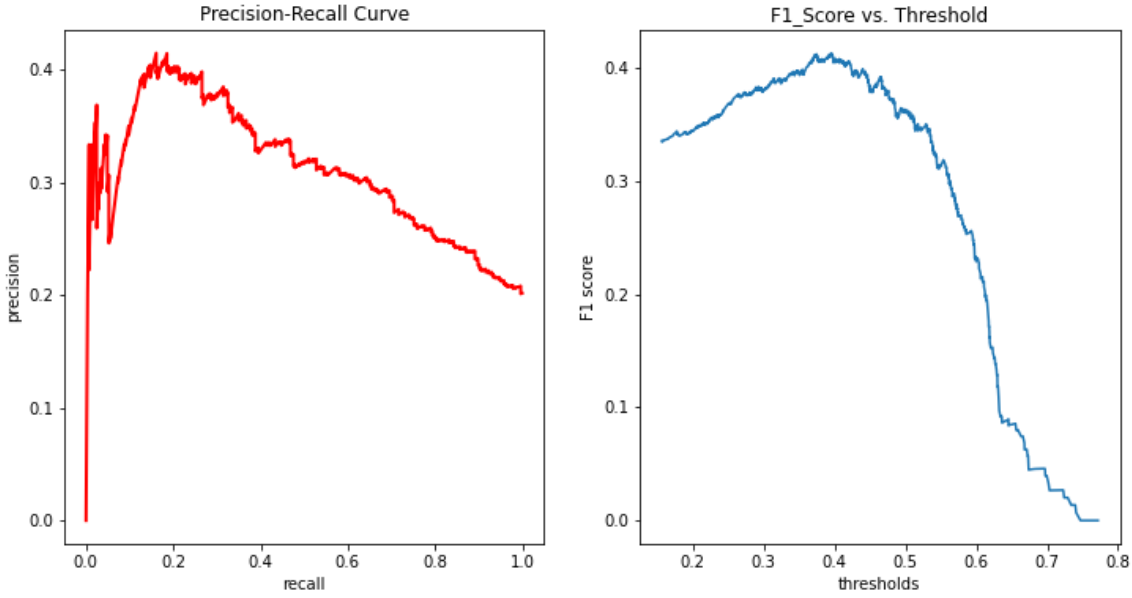


Figure 2c. Precision-Recall Curve and F1 Score vs. Thresholds



Hyperparameter Tuning and Model Selection

During training, number of layers and learning rate are adjusted to refine the model. Tested models (only added layers) are listed as below.

- Model 1:
  (Flatten)(Dense, 1024)(Dropout, 0.5)(Dense, 1024)(Dropout, 0.5)(Dense, 1), learning_rate = 1e-3
- Model 2:
  (Flatten)(Dense, 1024)(Dropout, 0.5)(Dense, 1024)(Dropout, 0.5)(Dense, 1), learning_rate = 1e-4

- Model 3:
  (Flatten)(Dense, 1024)(Dropout, 0.5)(Dense, 1), learning_rate = 1e-4

Model 2 is chosen because it achieves the highest AUROC score (0.668) and the highest Max F1 score (0.413). From the three sets of hyperparameters, we can see that decreasing the learning rate is a good approach to improve the model's performance.

Final Threshold and Explanation:
Final threshold: 0.302
This threshold value is chosen to optimize the recall rate, for this is more meaningful for its negative predictive value. When the model favors precision rate, the output recall and associated F1 score are too low to draw any meaningful conclusion. Thus, I decided to optimize for recall and try to increase PneuDetect's sensitivity to negative cases. The corresponding F1 score at threshold 0.302 is 0.381.

4. Databases

The ChextX-ray8 dataset consists of 112120 chest x-ray scans with disease labels from 30805 unique patients. The labels consist of 14 common thoracic pathologies and the 'No Finding' label.

There are only 1431 pneumonia positive images in the entire dataset, which leads to data imbalancing issue. To compensate for this data imbalance, the training dataset is sampled such that there are equal amount of positive and negative labels.

One pneumonia positive image is removed from training and validation because the corresponding patient is associated with an age outlier.

The demographic of patients:

Figure 3a. Patient Age

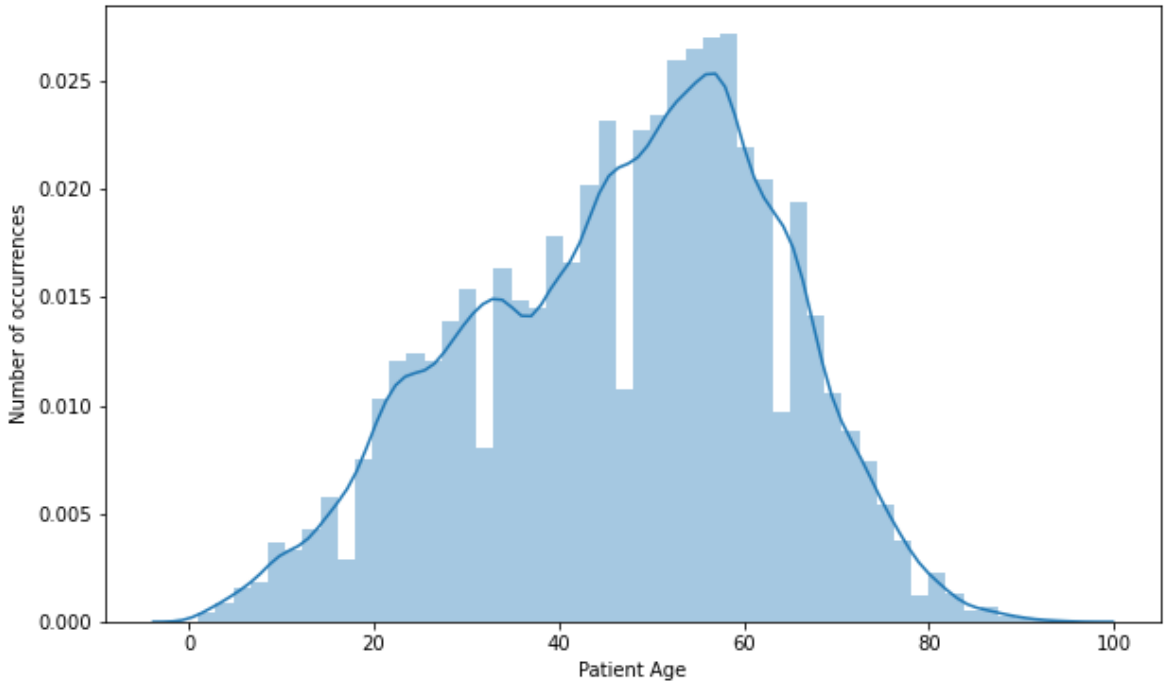


Figure 3b. Patient Gender

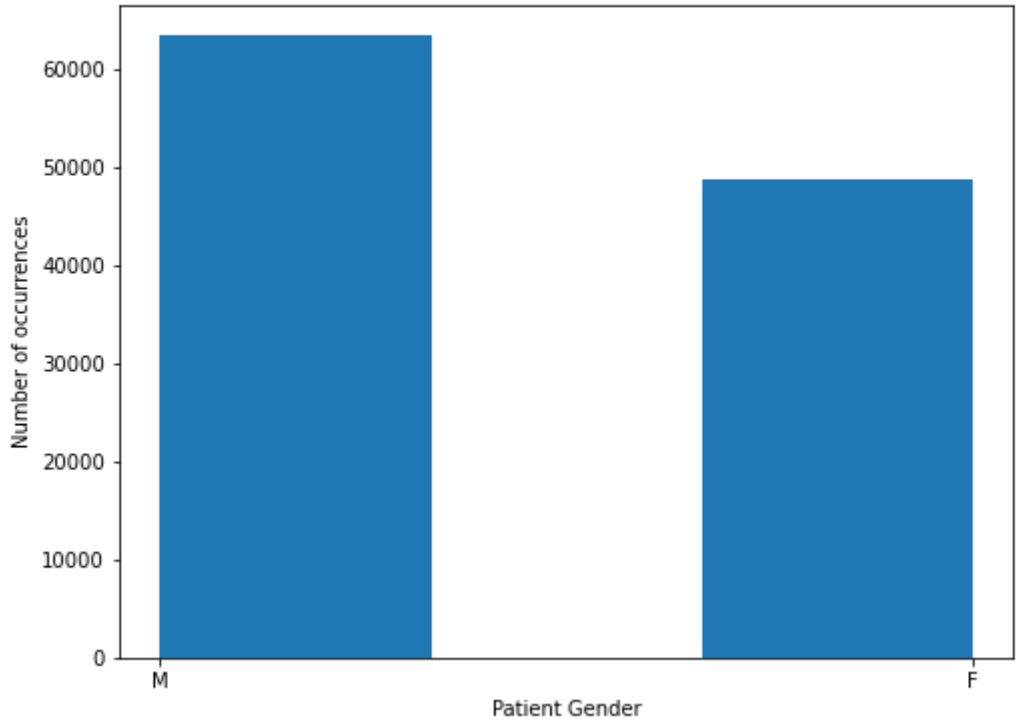
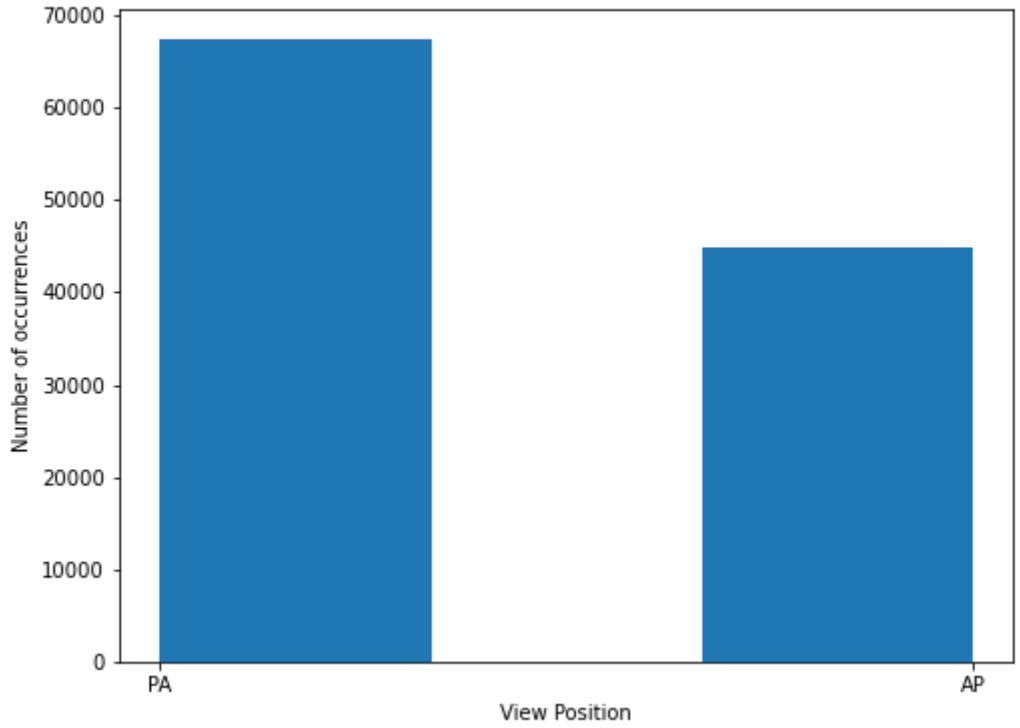


Figure 3c. View Position



Description of Training Dataset:

The training dataset consists of 2288 images in total, with 1144 pneumonia positive cases (positive to negative ratio is 1:1).

Description of Validation Dataset:
The validation dataset consists of 1430 images, with 286 pneumonia positive cases (positive to negative ratio is 1:4). This ratio reflects the occurrence of pneumonia in the real world.

5. Ground Truth

The true labels are used as the ground truth for model evaluation. These labels were created through text-mining radiologists' reports using Natural Language Processing (NLP). This approach may lead to erroneous labels although the estimated NLP accuracy is larger than 90%.

6. FDA Validation Plan

Patient Population Description for FDA Validation Dataset:
The patient population should follow the following demographic features:

- Ages of 1-95, following a normal distribution.
- The ratio of males to females is approximately 1.2:1.
- Patients can be previously diagnosed with lung diseases.

Ground Truth Acquisition Methodology:
Ideally, the ground truth should be a clinical conclusion summarized from multiple experienced radiologists' validations, along with patients' medical history and lab examination results.

Algorithm Performance Standard:
Rajpurkar et. al. (2017) reports an average F1 score, 0.387 from four practicing radiologists' labeling. The algorithm should achieve at least an F1 score of 0.387 for acceptable performance.