FDA submission

Your Name: Jared Bowden

Name of your Device: Deep learning assisted detection of pneumonia from chest x-rays

---

Algorithm Description

1. General Information

Intended Use Statement:

The algorithm described here is intended to assist a radiologist with the detection of pneumonia (classification in the form of “pneumonia” vs. “pneumonia”) from chest x-rays.

Indications for Use:

The algorithm is intended to improve the efficiency with which radiologists are able to evaluate patient x-rays for the presence of pneumonia, and exclude negative cases for the purposes of screening. It is intended as an additional data source to facilitate the process of x-ray image interpretation, not as a replacement for the assessment process. The algorithm is indicated for use with patients between the ages of 10 and 85, and both postero-anterior (PA) and antero-posterior (AP) view positions. Ages and image positions outside of those specified here may yield false positives or false negatives that could adversely impact interpretation of patient information.

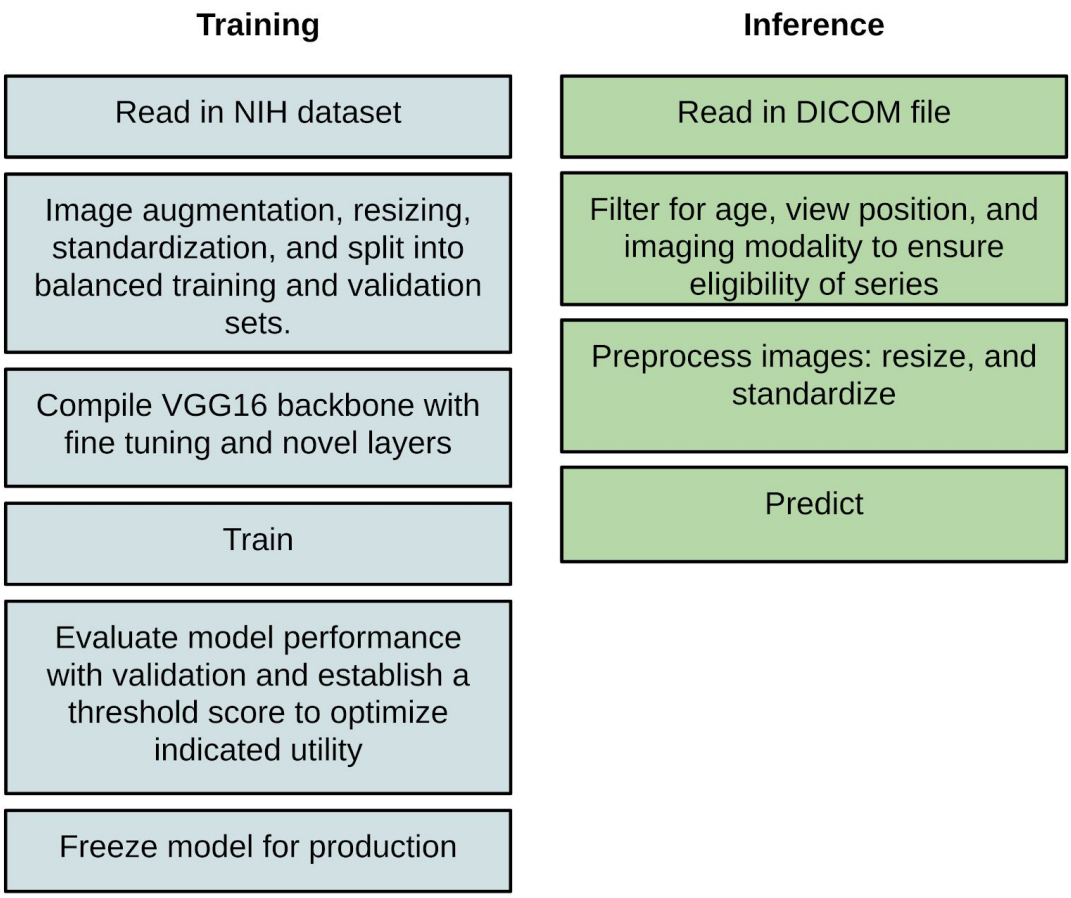
Device Limitations:

The algorithm will not function correctly if used on modalities beyond x-ray imaging. The speed with which the algorithm returns a result (“inference”) is hardware bound. Under conditions where image series exceed system memory (RAM) or processing (CPU), results will not be returned. or not returned in a timeframe that is appropriate for clinical use. A discrete graphics processing unit (CUDA capable) is strongly advised.

Clinical Impact of Performance:

Algorithm performance is specified in terms of the F1 metrics to facilitate interpretability with reference to existing literature; however, the final decision boundary threshold was informed by recall, with the intention of facilitating the confident identification of negative pneumonia cases -- as is necessary for the utility of high-throughput screening applications.

2. Algorithm Design and Function



DICOM Checking Steps:
The following checks are performed on the DICOM file prior to algorithm inference:

- Try/except block to ensure DICOM pixel array data can be read
- Filter to issue warning to indicate against inference under conditions where:
    - Patient age is greater than 85
    - Patient age is less than 10
    - Image modality is x-ray
    - Patient position is postero-anterior (PA) or antero-posterior (AP)
    - Body part examined is chest

Preprocessing Steps:
Prior to inference, the following preprocessing steps were conducted.

- Resize patient series to parameters required for algorithm in put (224,224)
- Standardization based on mean and standard deviation pixel intensity parameters acquired through training data.

CNN Architecture:
The algorithm was trained from a VGG16 backbone. Seventeen layers of this network were frozen, the remaining 2 were fine-tuned through training, with the inclusion of 4 additional layers, specified below.

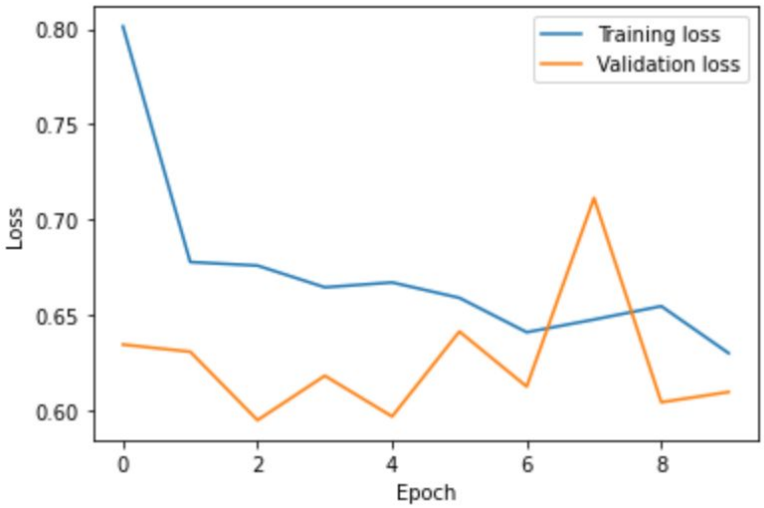
3. Algorithm Training

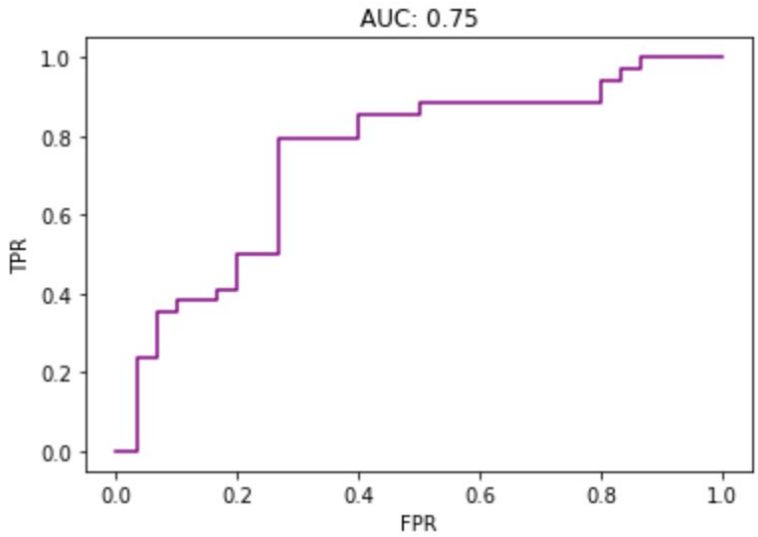
Parameters:

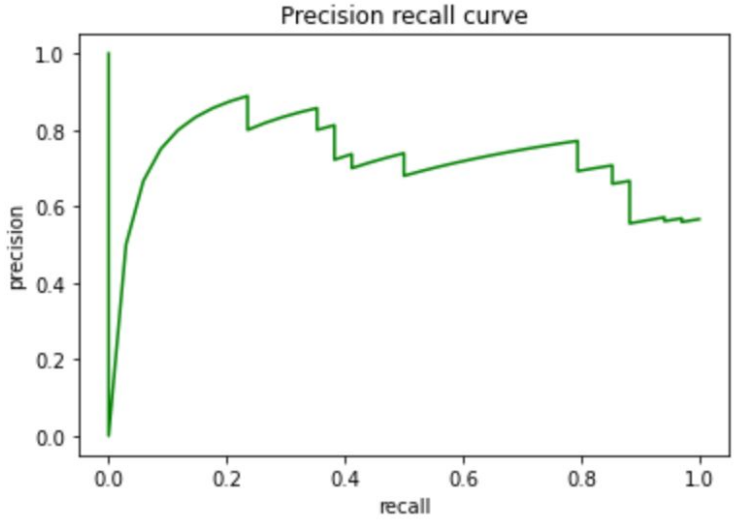
- Types of augmentation used during training
    - Image augmentation was performed with the ImageDataGenerator function of the Keras preprocessing library
    - `rescale=1. / 255.0,`
    - `horizontal_flip = True,`
    - `vertical_flip = False,`
    - `height_shift_range = 0.1,`
    - `width_shift_range = 0.1,`
    - `rotation_range = 20,`
    - `shear_range = 0.1,`
    - `zoom_range= 0.1`
- Batch size
    - `batch_size = 64`
- Optimizer learning rate
    - `Atom optimizer,`
    - `Learning rate of 1e-4`
- Layers of pre-existing architecture that were frozen
    - The first 17 layers of the VGG16 next work were frozen
    - `block1_conv1 False`
    - `block1_conv2 False`
    - `block1_pool False`
    - `block2_conv1 False`
    - `block2_conv2 False`
    - `block2_pool False`
    - `block3_conv1 False`
    - `block3_conv2 False`
    - `block3_conv3 False`
    - `block3_pool False`
    - `block4_conv1 False`
    - `block4_conv2 False`
    - `block4_conv3 False`
    - `block4_pool False`
    - `block5_conv1 False`
    - `block5_conv2 False`
- Layers of pre-existing architecture that were fine-tuned
    - The remaining 2 layers of the VGG16 network were fine-tuned
    - `block5_conv3 True`
    - `block5_pool True`
- Layers added to pre-existing architecture

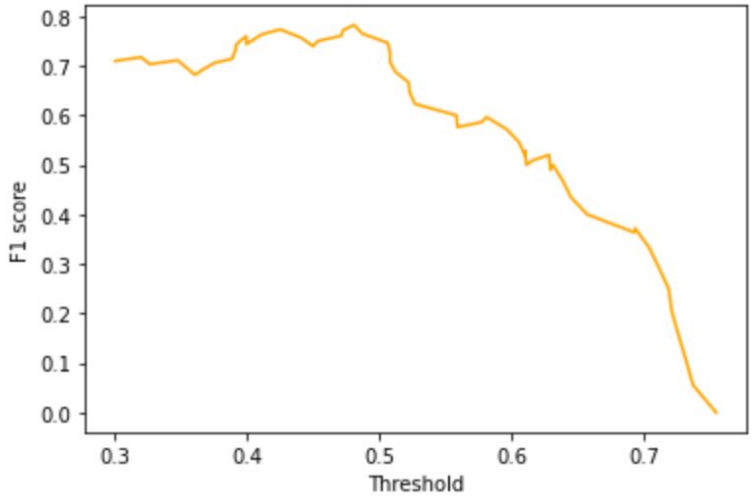
Layer (type)	Output Shape	Param #
model_9 (Model)	(None, 7, 7, 512)	14714688
flatten_4 (Flatten)	(None, 25088)	0
dropout_8 (Dropout)	(None, 25088)	0
dense_9 (Dense)	(None, 1024)	25691136
dense_10 (Dense)	(None, 1)	1025

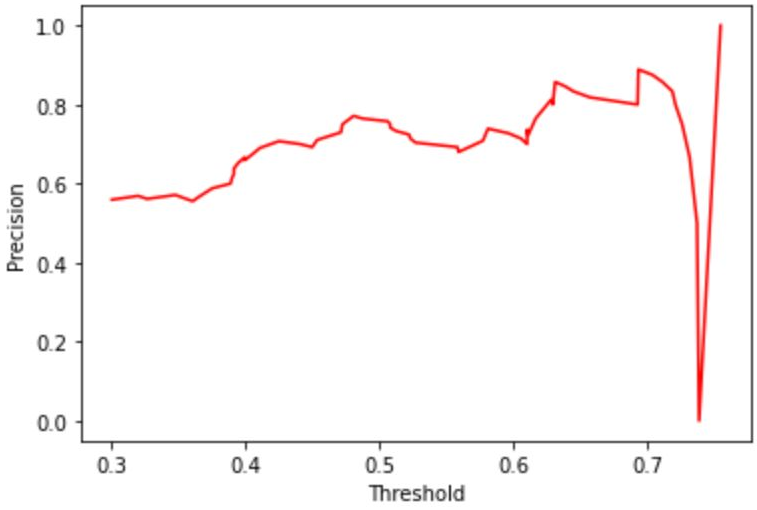
Total params: 40,406,849
Trainable params: 40,406,849
Non-trainable params: 0











Final Threshold and Explanation:
The final threshold (0.63) was largely informed by the performance of the F1 metric. This decision was justified through the ability of this metric to capture effects of both precision and recall, as well as precedent, and the ability to compare the existing algorithm with the performance of other attempts at this problem, within the literature. Additional adjustment was made to prioritize the recall metric, and ensure that the algorithm was

optimized to maximize the confidence in identifying negative pneumonia cases, as opposed to *correctly* classifying pneumonia cases.

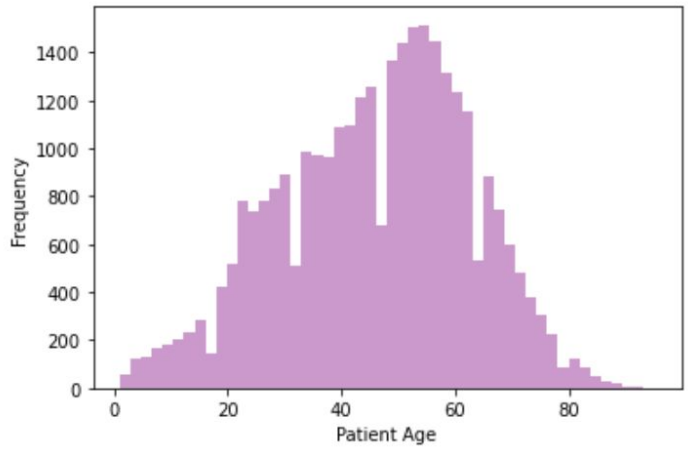
4. Databases

The image, labeling, and descriptive data used for the development of this algorithm was obtained from the NIH Chest X-ray Dataset. Data consists of 112,120 X-ray images taken from 30,805 patients and labeled for 14 different pathologies, one being pneumonia.

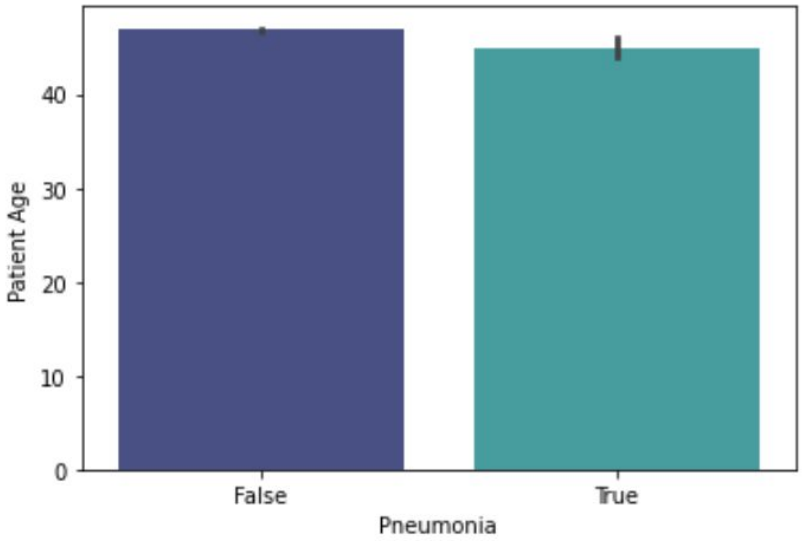
Patient age within the dataset ranges from 1 to 95 years old, with an average of 46.

```
count    30802.000000
mean        46.087559
std         16.692500
min          1.000000
25%         34.000000
50%         48.000000
75%         58.000000
max         95.000000
Name: Patient Age, dtype: float64
```

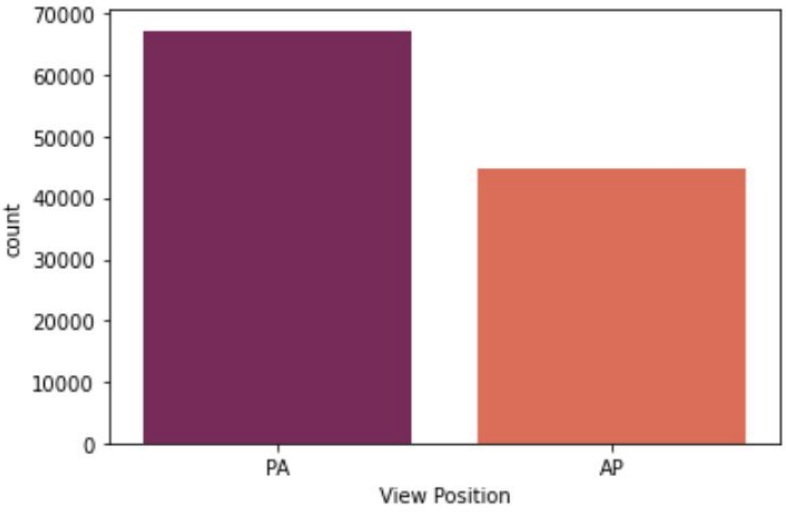
```
Text(0, 0.5, 'Frequency')
```



On average, the age of patients with pneumonia appears to be comparable to those without.

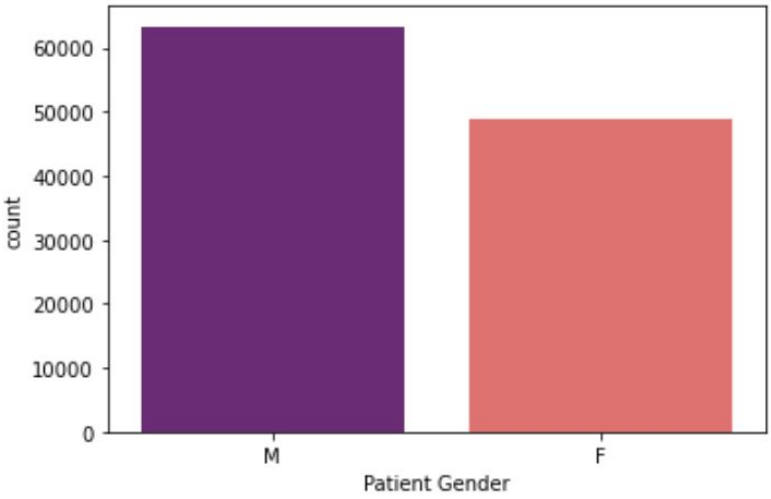


X-ray images were acquired in postero-anterior (PA) or antero-posterior (AP) orientations, with the majority being collected in the former.

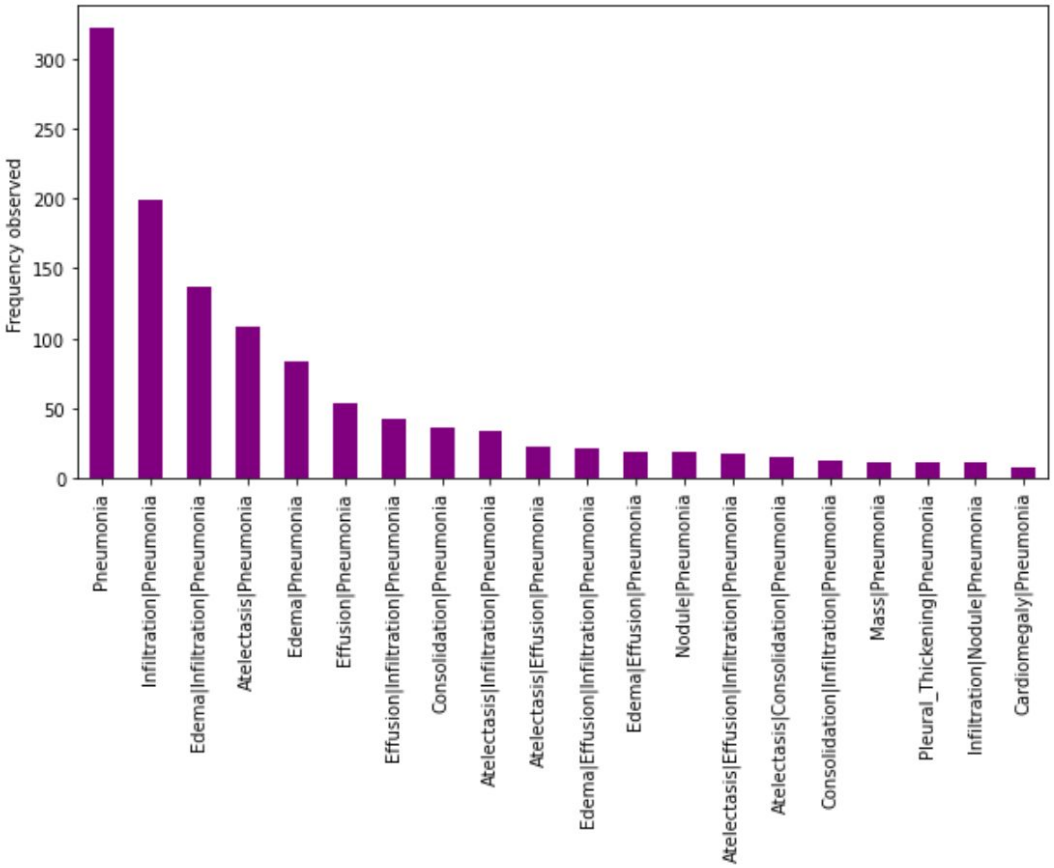


The gender of the population is predominantly male, however, the proportion of pneumonia labels appears to be fairly balanced across males and females.

```
Proportion of pneumonia label, male
False    0.98677
True     0.01323
Name: Pneumonia, dtype: float64
-------------------
Proportion of pneumonia label, female
False    0.987843
True     0.012157
Name: Pneumonia, dtype: float64
```

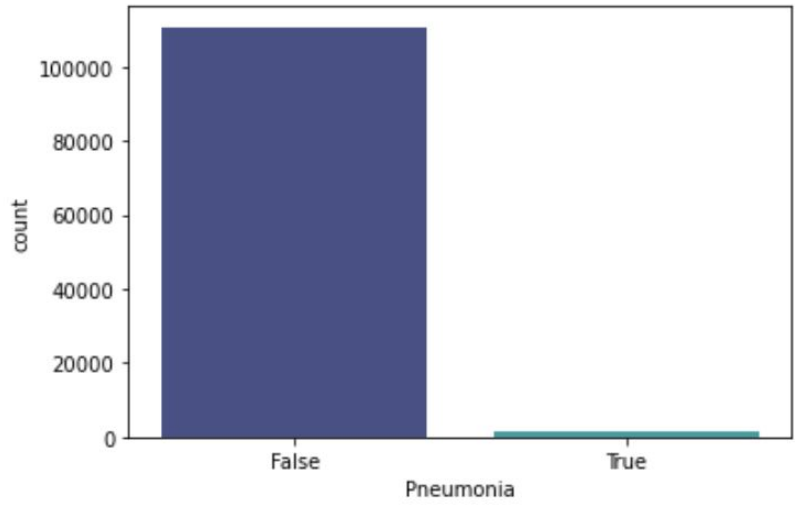


Comorbidities with the pneumonia label are quite common, the most prevalent appears to be infiltration. The combination of infiltration with pneumonia in ~75% as many cases as pneumonia alone.



Lastly, out of all total labels, pneumonia (and pneumonia comorbidities) account for only a small fraction of the total dataset.

```
False    0.987237
True     0.012763
Name: Pneumonia, dtype: float64
```



Description of Training Dataset:

Two features stand out as necessary design considerations for preparation of the training set. First and most prominently, the pneumonia class is highly under-represented within the dataset. Stratification is necessary to ensure that this population is represented proportionally across training and test sets, and reduction of

non-target classes is necessary to ensure equal representation with the target class. Secondly, owing to the fact that (in most cases) multiple images exist for each patient, there exists the possibility that images from a patient could exist in both the training and validation sets. Unaccounted for, this features of the dataset could result in leakage: the algorithm could learn patient-specific signals (unrelated to the label) that could be used to re identify the label within the validation set. To account for this, additional measures were taken within split functions to ensure participants could not appear in both training and validation sets.

Aside from the aforementioned design considerations (stratification of an unbalanced label, and blocking of patients across training or validation groups), splits were performed through conventional methods of random assignment.

Description of Validation Dataset:
As with training, the target label was stratified to ensure equal representation across training and validation sets, and randomized (while accounting for the need to avoid splitting participants between both training and test sets) After some review, the validation set was balanced (removing non-target cases to equalize) to ensure that performance metrics like accuracy could be meaningful interpreted after prototyping.

5. Ground Truth
Ground truth was established through radiologists reports, but extracted through natural language processing (NLP). The accuracy with which a radiologist can correctly identify the classifications that comprise this study is high; however, the NLP used to extract these labels from reports is purported to be >90% accurate. Accounting for human error and NLP accuracy, these limitations must be acknowledged as a potential weakness in the development process.

6. FDA Validation Plan
Patient Population Description for FDA Validation Dataset:
The dataset required for FDA validation should, to the extent possible, mirror the size and demographic composition of that specified above for development. Most notably, the validation population should endeavor to balance the presence of comorbid diseases observed within the original data set, specifically:

```
Atelectasis, Consolidation Infiltration/Pneumothorax,Edema, Emphysema,
Fibrosis, Effusion, Pneumonia, Pleural thickening, Cardiomegaly, Nodule Mass,
and Hernia.
```

Ground Truth Acquisition Methodology:
Criteria for ground truth within the FDA validation set can be improved. As opposed to parsing labels from radiologists reports, the validation set should instead endeavor to extract (and standardize) radiologists reports, directly. A further proposed elaboration would be a silver standard methodology: averaging classifications across multiple radiologists, and (optionally) weighting for years of experience. Additional improvements in ground truth could be achieved through the addition of lab tests to confirm the presence or absence of other comorbid conditions observed within patient series.

Algorithm Performance Standard:
As described by Rajpurkar et. al., (2017), an F1 score comparison of algorithm performance as compared to performance achieved by 4 (or more) more radiologists would provide a performance standard that is both

valid (radiologist are adept at identifying the target class from x-rays), as well as having established precedent for comparison within the literature.