FDA Submission

Your Name: Katie Geary

Name of Device: A Machine Learning Algorithm for the Classification of Pneumonia in X-ray Images

Algorithm Description

1. General Information

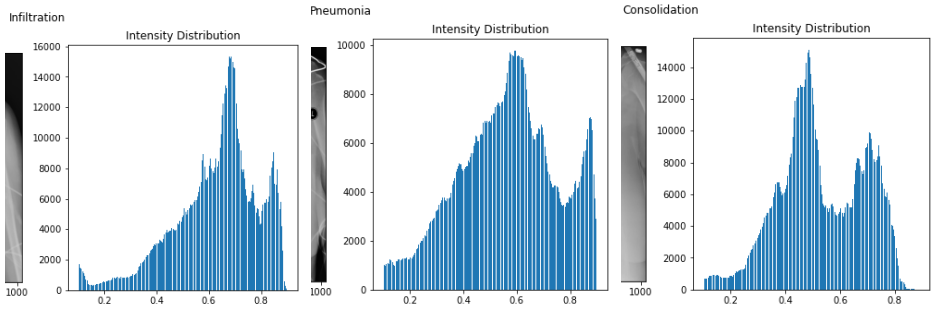
Intended Use Statement:
This algorithm is intended for assisting the radiologist in the detection of pneumonia from chest x-rays.

Indications for Use:
This algorithm is indicated for use in files of

- males and females between the ages of 1 and 95 years old
- Chest x-ray taken while the patient is in the anteriorposterior (AP) or posterioranterior (PA) position.

Device Limitations:

- While training of the algorithm requires GPU, processing of images can be completed in a timely fashion using CPU.
- False positives are likely in situations where diseases with similar intensity plots are present. Infiltration and consolidation are two examples identified from this data set.

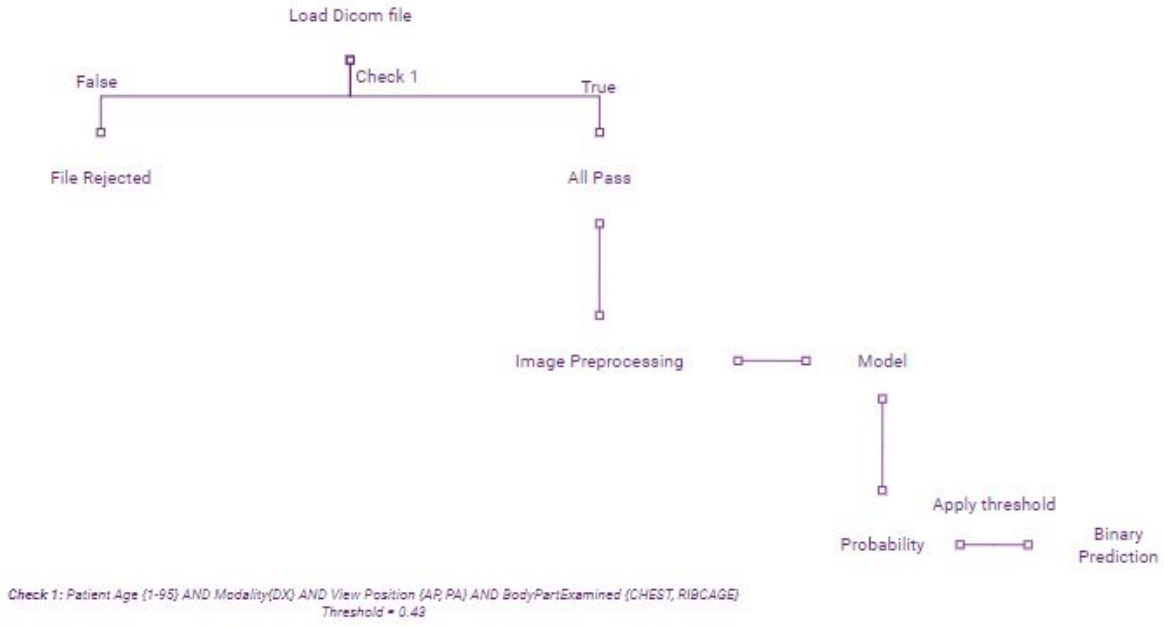


Clinical Impact of Performance:
The threshold for this model strongly favors recall over precision. Because of this, false positives are much more likely than false negatives. This algorithm will not be a good fit as a diagnostic tool, but could be extremely useful as a prioritization tool to draw the radiologist's attention to the images that are potentially pneumonia. If the model predicts the image is negative for pneumonia, there is a very good chance it actually is -- there will be very very few false negatives. However, if the model predicts the image is pneumonia positive, there is a decent chance it actually is not (false positive). However, the risk to this is mitigated because the radiologist will be viewing the image and making the final call. There is no risk to pushing the

false positives above the true negatives in the workflow. The key benefit will come from the true positives which are recognized quicker by being prioritized.

2. Algorithm Design and Function

Algorithm Flowchart



DICOM Checking Steps:
Before proceeding to the Image Preprocessing step, the algorithm checks to make sure

- The Modality is DX (x-ray)
- The body part examined is the Chest
- The patient was radiographed in position AP or PA

Preprocessing Steps:
Once an image proceeds to the preprocessing stage, the following occur:

- The image is converted to greyscale (if not already in greyscale)
- The image is reshaped to be 1,224,224,3
- Normalization of intensities

CNN Architecture:
The core of the algorithm is the pre-trained VGG16 Neural Network. The first 18 layers (0:17) were frozen to prevent retraining, while the final convolutional and pooling layers were left unlocked to be connected to and retrained with the 13 new layers that were added. These 13 layers included a layer to flatten the final output from VGG16 before adding dropout and dense

layers leading up to a final output for binary classification. A screen capture of the network architecture can be seen below:

```
Downloading data from https://github.com/fchollet/deep-learning-models/releases/download/v0.1/vgg16_weights_tf_dim_ordering_tf_kernels.h5
553467904/553467096 [==============================] - 10s 0us/step
VGG16 layers:
input_1 False
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
block5_conv3 True
block5_pool True
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
model_1 (Model)              (None, 7, 7, 512)         14714688
_________________________________________________________________
flatten_1 (Flatten)          (None, 25088)             0
_________________________________________________________________
dropout_1 (Dropout)          (None, 25088)             0
_________________________________________________________________
dense_1 (Dense)              (None, 1024)              25691136
_________________________________________________________________
dropout_2 (Dropout)          (None, 1024)              0
_________________________________________________________________
dense_2 (Dense)              (None, 512)               524800
_________________________________________________________________
dropout_3 (Dropout)          (None, 512)               0
_________________________________________________________________
dense_3 (Dense)              (None, 256)               131328
_________________________________________________________________
dropout_4 (Dropout)          (None, 256)               0
_________________________________________________________________
dense_4 (Dense)              (None, 128)               32896
_________________________________________________________________
dropout_5 (Dropout)          (None, 128)               0
_________________________________________________________________
dense_5 (Dense)              (None, 64)                8256
_________________________________________________________________
dropout_6 (Dropout)          (None, 64)                0
_________________________________________________________________
dense_6 (Dense)              (None, 1)                 65
=================================================================
Total params: 41,103,169
Trainable params: 28,748,289
Non-trainable params: 12,354,880
_________________________________________________________________
```

3. Algorithm Training

Parameters:

Types of augmentation used during training include:

- Samplewise_center = True
- Samplewise_std_normalization = true

- Horizontal flips = True
- Height shift (range =0.1)
- Width shift (range = 0.1)
- Rotation (range = 20)
- Shear (range = 0.1)
- Zoom (range = 0.1)

Types of augmentation used during validation include:

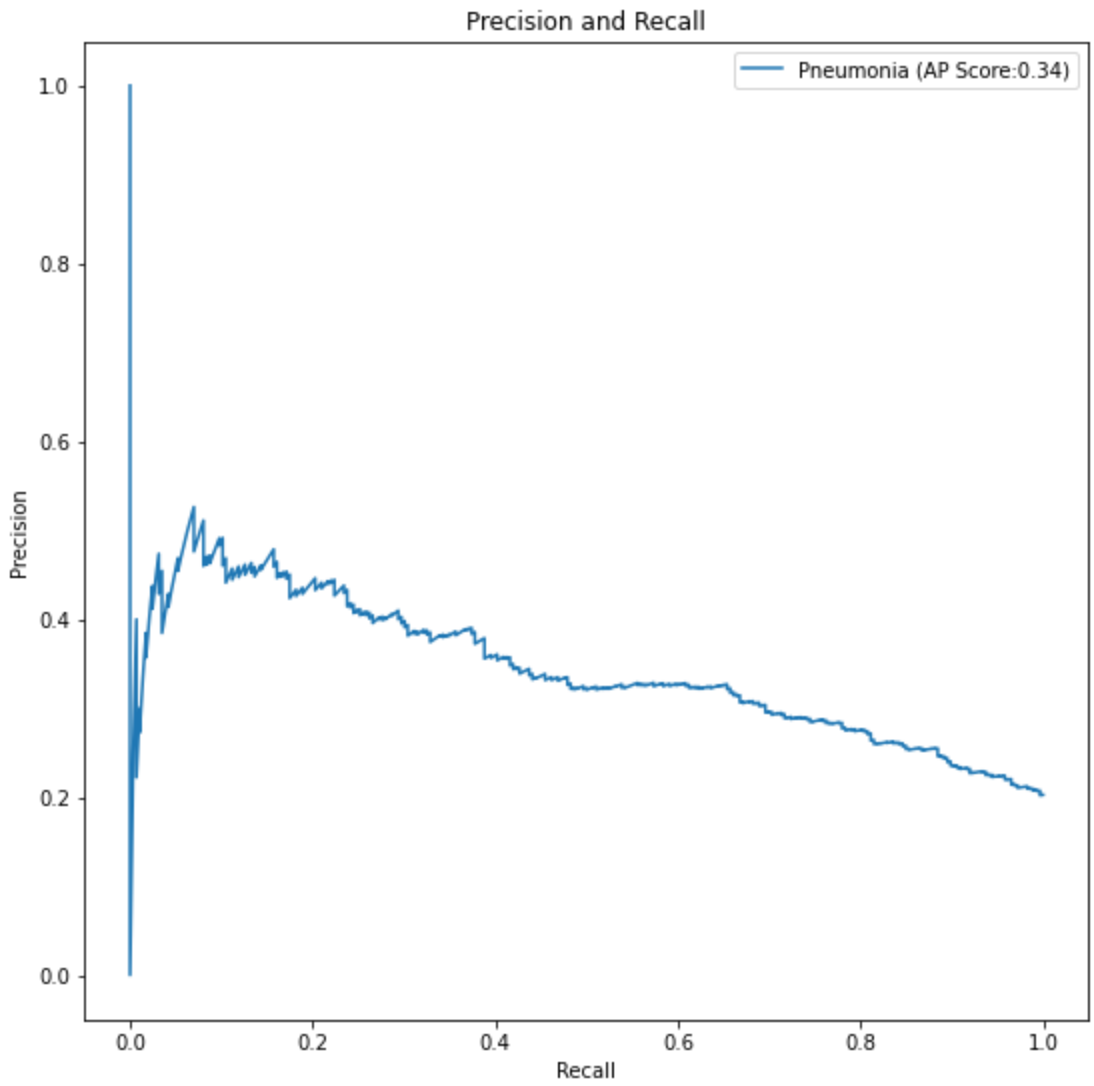
- Samplewise_center = True
- Samplewise_std_normalization = true

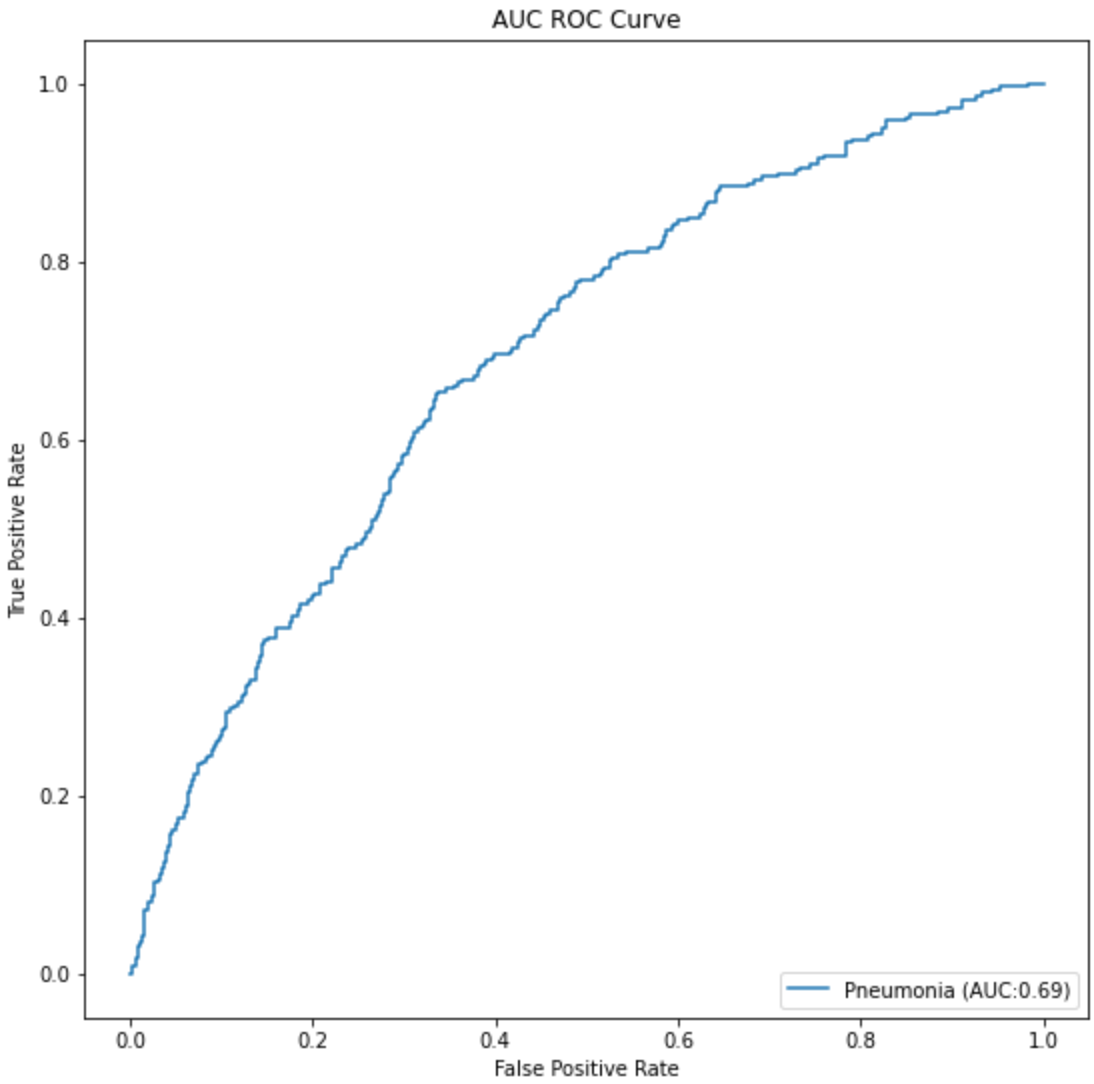
The batch size for training was 64, while the batch size for validation was 1430

The optimizer learning rate was set to 1e-4

Layers 0:17 of pre-existing (VGG16) architecture were frozen for training
Layers 18 and 19 of pre-existing (VGG16) architecture were not frozen, and fine-tuned during training.
A series of Flatten, Dense, and Dropout layers were added to the end of the pre-existing architecture, as seen in the figure above.

Insert algorithm training performance visualization, P-R curve and AUC ROC curve


Precision and Recall
Pneumonia (AP Score:0.34)
Precision
Recall
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0



Final Threshold and Explanation:
Threshold: 0.43003806
F1 Score: 0.45945945945945954

The F1 score was used as the metric of choice to determine the optimal threshold. Based on a plot of F1 score versus threshold (as seen below), the F1 score was at a maximum when the threshold was at 0.4300. This threshold favors recall over precision, making it ideal for an assistive algorithm, as it will be extremely confident in its negative predictions and abundant with positive identifications to insure the radiologist reviews with significant potential to be pneumonia. In a study completed by researchers at Stanford University, the average radiologist F1 score when determining the presence of pneumonia is 0.387 (Rajpurkar, 2017). While this is not a significant increase, nor is it near accurate enough to operate on its own, this metric shows the algorithm could be highly beneficial in focusing the efforts of the radiologist.

Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., ... & Ng, A. Y. (2017). Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*.

4. Databases

(For the below, include visualizations as they are useful and relevant)

The database used for the training and validation of this algorithm is called the NIH Chest X-ray Dataset. The dataset contains 112,120 chest x-ray images taken from more than 30,805 different individuals, making it one of the largest publicly available datasets of chest x-rays. Each image is a PNG file with associated metadata including image index, finding labels, follow-up number, patient ID (all de-identified), patient age, patient gender, view position, image size (original) and Image pixel spacing. Of all patients in the dataset with pneumonia, 41.5% were female and 58.6% were male. The data was split into two sets, a training and validation set, sampling from data with and without pneumonia to create specific ratios of each. Before splitting, all patients over the age of 100 were removed from the dataset as preliminary EDA revealed these cases to be outliers with every case over 100 having an actual age over 150 years.

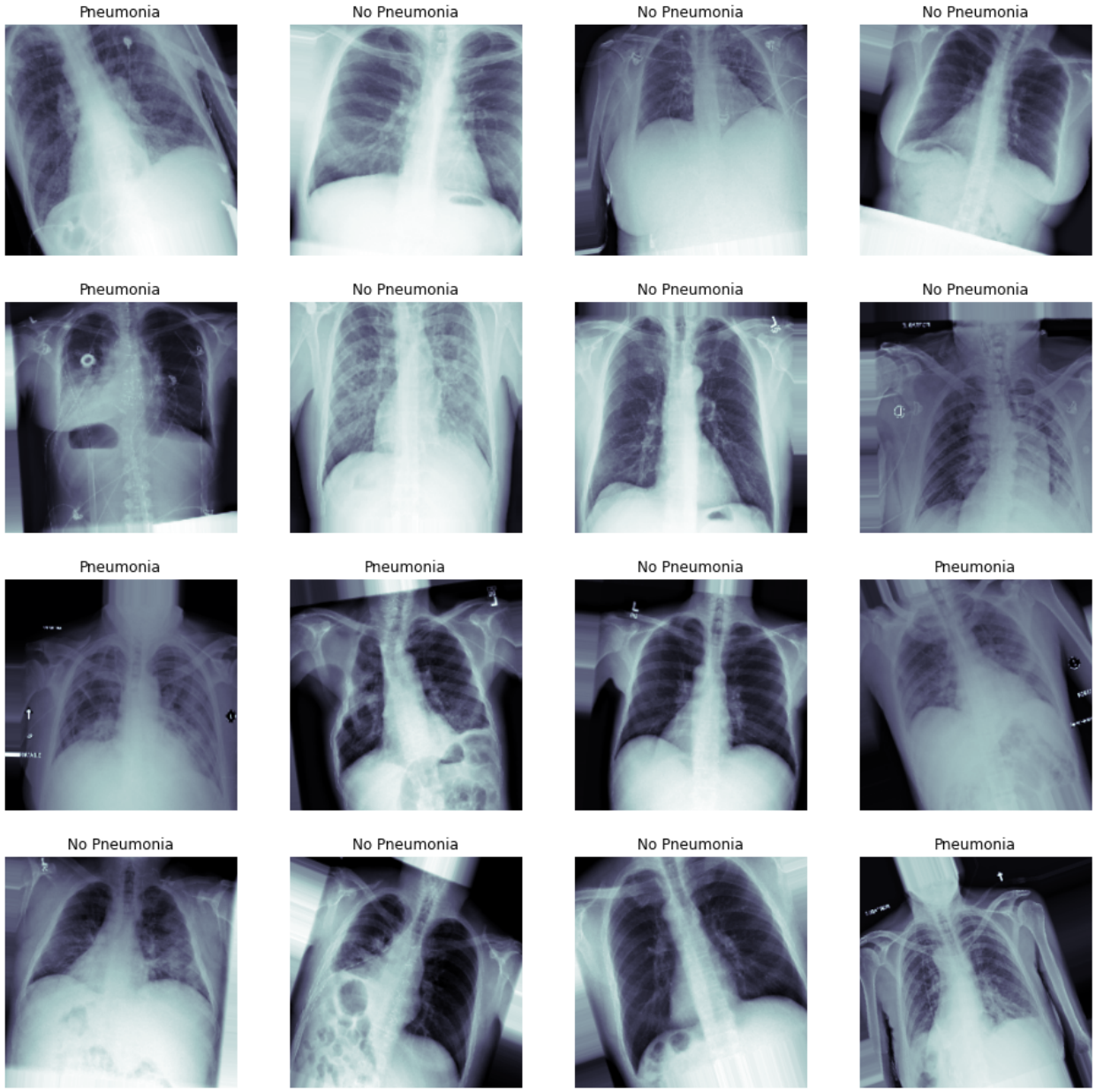
Description of Training Dataset:

The final training dataset was split 50/50 between pneumonia and non-pneumonia images (based on their ground truth labels) to give the algorithm sufficient chance to view images with and without pneumonia to learn with as little bias as possible. However, it is important to note that the non-pneumonia class was a combination of other ailments excluding pneumonia as well as cases in which no findings were recorded. In the pneumonia class, other ailments could be present in addition to pneumonia. This was an effort to have the machine learning algorithm be prepared to ascertain the presence of pneumonia with and without comorbidities to prepare for real life scenarios.

There were 2,288 images in the training set; 1,145 with pneumonia and 1,145 without. The minimum age in the training set was 2 years old, and the maximum was 90 years old. Of the subjects in the training set, approximately 40.7% were female and 59.3% were male; in the subset of subjects with pneumonia 41.2% were female and 58.9% were male. The age distribution within the training set was normally distributed, but like the original/complete data set, slightly skewed to the right with a peak at approximately 60 years old.This pattern of age distribution was found in both the patients with and without pneumonia in the training dataset. In the training data set, the most common finding, as in the full original dataset, was predominately “no finding,” with over 600 cases reporting no presence of a finding. The second most common finding, unlike in the original set, was Pneumonia. This makes sense as we forced the training set to be 50% pneumonia cases. During EDA, it was determined that pneumonia occurs most often on its own. This trend followed in the training subset and Infiltration and Edema with Infiltration were the two most common comorbidities with pneumonia. In the non-pneumonia cases no-finding was most common followed by Infiltration and Atelectasis with less than 20% of the number of “no finding” cases.

The training set was augmented using ImageDataGenerator to standardize and zero mean the images, normalize their intensities to be between 0 and 1, horizontal flips (but no vertical) were

introduced, heights and width were shifted within a range of 0.1, up to 20 degrees of rotation were allowed, as well as a shear and zoom range of 0.1. A sample of the training set can be seen below.



Description of Validation Dataset:
The final validation set was split 20/80, with 20% of the images having pneumonia present (according to their ground truth labels). This split was chosen in an effort to reflect pneumonia in "the wild." Like the training set, pneumonia positive images could include comorbidities, while pneumonia negative images could have non-pneumonia findings or no finding at all.

There were 1430 validation images with 286 representing pneumonia and 1,144 lacking the presence of pneumonia. The minimum age in the validation set was 2 years old, and the

maximum was 90 years old. Of the subjects in the validation set, approximately 43.5% were female and 56.4% were male; in the subset of subjects with pneumonia 42.6% were female and 57.3% were male. The age distribution within the validation set was normally distributed, but like the original/complete data set, slightly skewed to the right with a peak at approximately 55 years old.This pattern of age distribution was found in both the patients with and without pneumonia in the validation dataset. In the validation data set, the most common finding, as in the full original dataset, was predominately “no finding,” with over 600 cases reporting no presence of a finding. Pneumonia was the third most common finding. This makes sense as we forced the training set to be 20% pneumonia cases by including 4 times as many non-pneumonia cases. During EDA, it was determined that pneumonia occurs most often on its own. This trend followed in the training subset and Infiltration and Edema with Infiltration were the two most common comorbidities with pneumonia. In the non-pneumonia cases no-finding was most common followed by Infiltration and Effusion with less than 10% and 20% (respectively) of the number of “no finding” cases.

The validation set was only augmented to standardize and zero-mean the images in the dataset by using a function in the Keras package to alter the images to have an intensity between 0 and 1 with a close distribution. A sample of the validation set can be seen below. The number pairings above the images are (truth, prediction).

5. Ground Truth

The ground truth labels for all images in this dataset were created using a Natural Language Processing system (NLP). The NLP was used to review the associated radiological reports for each file and to extract the radiologist's diagnosis of the x-ray. The NLP's accuracy is estimated to be >90%, making them sub-optimal and opening the potential to impact the algorithms performance if ~10% of the images are mislabeled. NLP mislabeling is typically the result of conflicting positional words in the radiology report such as "it is unlikely, given the findings, that pneumonia is present." There is always a chance the NLP will only pick up on "pneumonia is present." The potential labels include 14 common thoracic pathologies in addition to 'No finding.'

- Atelectasis

- Consolidation
- Infiltration
- Pneumothorax
- Edema
- Emphysema
- Fibrosis
- Effusion
- Pneumonia
- Pleural thickening
- Cardiomegaly
- Nodule
- Mass
- Hernia
- No finding

6. FDA Validation Plan

Patient Population Description for FDA Validation Dataset:
To qualify for inclusion in the FDA validation set, the image must be taken from

- Man or woman between the ages of 1 and 95 years old
- Must be a chest x-ray (DX)
- Must be taken in the AP or PA position

Ground Truth Acquisition Methodology:
According to the Infectious Diseases Society of America, the gold standard for determining pneumonia is with chest x-rays (Bartlett, 2000). In compliance with this, the ground truth acquisition method would be to have an average of 3 different radiologists.

Algorithm Performance Standard:
In compliance with the F1 scores reported in Rajpurkar, 2017, the algorithms F1 score should be on par or better than the average radiologist's score of 0.387.