FDA Submission

Author Name: Saeed Sheikh

Device Name:
PneumoClassifier: Algorithm for the identification of Pneumonia within chest X-rays

Algorithm Description

1. **General Information**
    a. **Intended Use**
    This algorithm is intended to be used for the identification of Pneumonia within chest X-rays.

    b. **Indications for Use**
    This algorithm is intended for use on men and women from the ages of 1-95 who have been administered a chest X-ray, where posteroanterior (PA) and anteroposterior (AP) views exist of the chest have been captured.

    c. **Device Limitations**
    This algorithm is only effective in identifying Pneumonia within DICOM images of chest X-rays and is ineffective in identifying any other diseases/abnormalities present within any other images of any other part of a patient's anatomy.

    d. **Clinical Impact of Performance**
    This algorithm could be integrated into a clinical workflow on DICOM images produced after image acquisition. The output from this algorithm along with the original DICOM image could be used to help guide a radiologist or other clinician in making their final diagnosis. False positive identifications of Pneumonia may lead to inappropriate or unnecessary treatment being administered if they are not caught prior to the final diagnosis being determined and delivered to the patient. False negatives may lead to critical treatment being withheld from a patient. Hence this algorithm should not solely be relied upon as a source for final determination of applying or withholding treatment.

2. **Algorithm Design and Function**

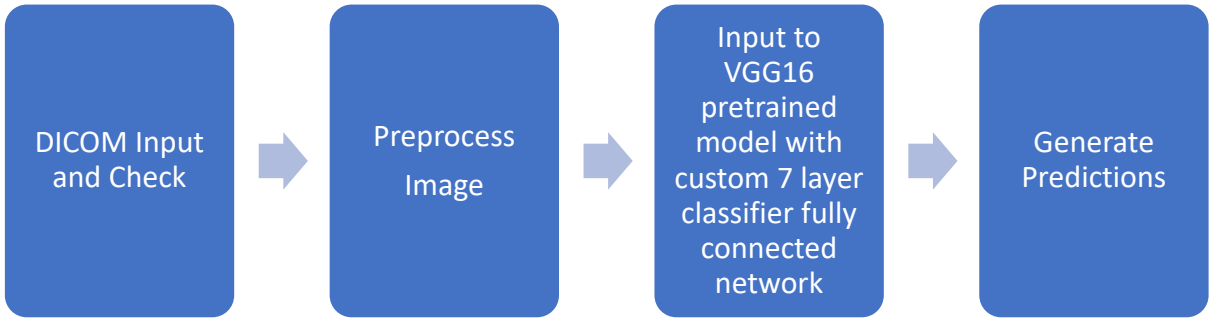

Figure 1: Algorithm Flowchart

a. **DICOM checks**
The Algorithm reads in DICOM images and checks if they contain the digital radiography ('DX') Modality code, if the age of the patient falls within the bounds for the algorithm, as established above, if the imaged anatomy is of the chest, and if the image views are either posteroanterior (PA) or anteroposterior (AP).

b. **Image Preprocessing**
The images are then normalized using the mean and standard deviation of the pixel information for the given image.

c. **CNN Architecture**
After preprocessing the normalized images are inputted into a deep learning model. The fully trained deep earning model along with weights obtained from training are loaded into the algorithm. The model is then used to predict the presence of Pneumonia within the inputted image.
The fully trained model is built using the VGG16 convolutional neural network. See https://neurohive.io/en/popular-networks/vgg16/. The architecture of the VGG16 neural network is frozen after seventeen layers and custom seven-layer classifier network is used, where each layer is half the size of the previous layer.

3. **Algorithm Training**
   a. **Parameters:**
   Training Augmentations
   The following Image augmentations were used to facilitate image variations for the model, these augmentations were selected after experimentation:

```
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

Validation Augmentation
Images were only rescaled for validation
rescale=1. / 255.0

Batch Size
A Batch size of 32 was chosen to help speed up training and due to the large size of the dataset being trained on. This increased batch size did not impact performance when the model was trained using smaller batch sizes

Optimizer learning rate
The optimizer learning rate was set to 1e-6 to help improve performance during training.

Layers of pre-existing architecture that were frozen
Seventeen layers from the pretrained VGG16 model were frozen.

Layers of pre-existing architecture that were fine-tuned
The Block5 layer from the VGG16 model was used as the transfer layer and was fine tuned.

Layers added to pre-existing architecture
A layer to Flatten the input was added after the Block5 layer from the VGG16 network. Next a 1024 sized dense layer was added with a ReLu activation, followed by a dropout layer, with a dropout of 0.5. This was followed by a 512 sized dense layer with a ReLu activation, followed by another dropout layer, with a dropout of 0.5. Finally, a 256 sized dense layer was added with a ReLu activation, followed by a dense layer with size 1 having a Sigmoid activation.

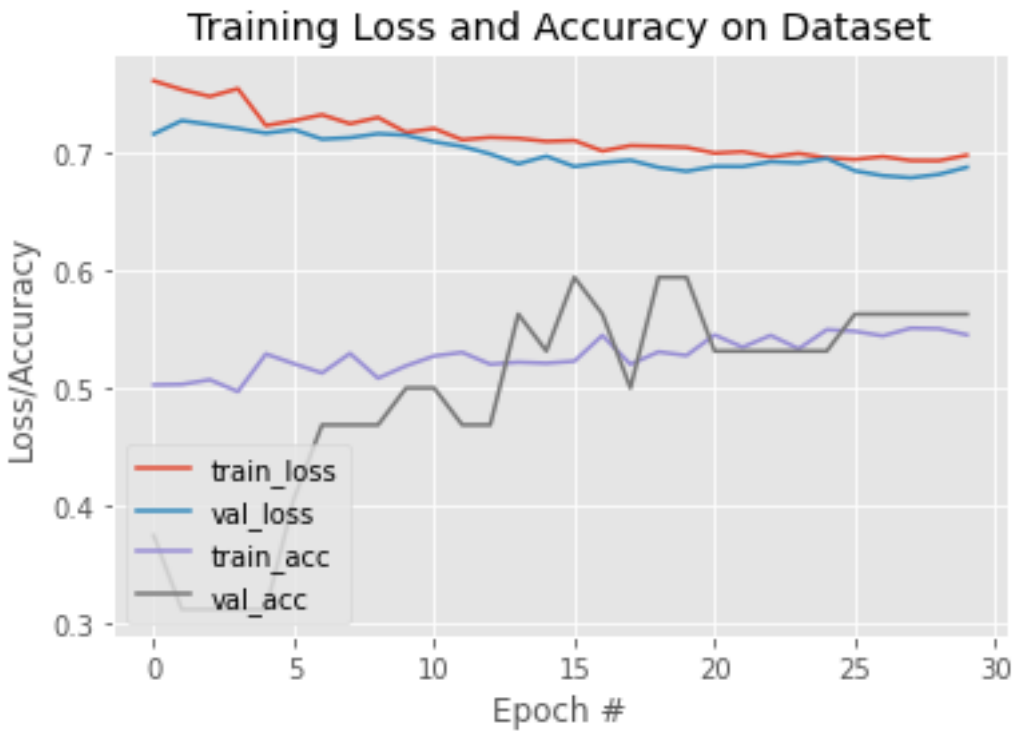

Figure 2: Plot of Model Training History

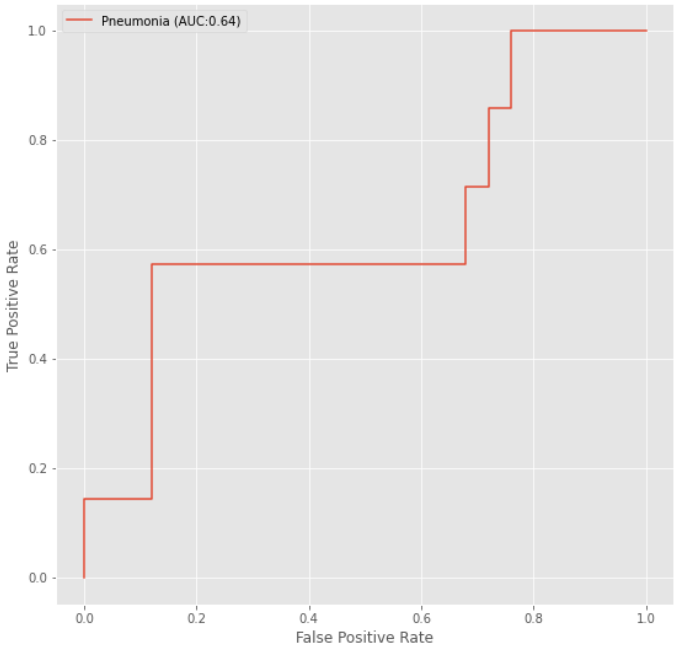

Figure 3: Plot of AUC

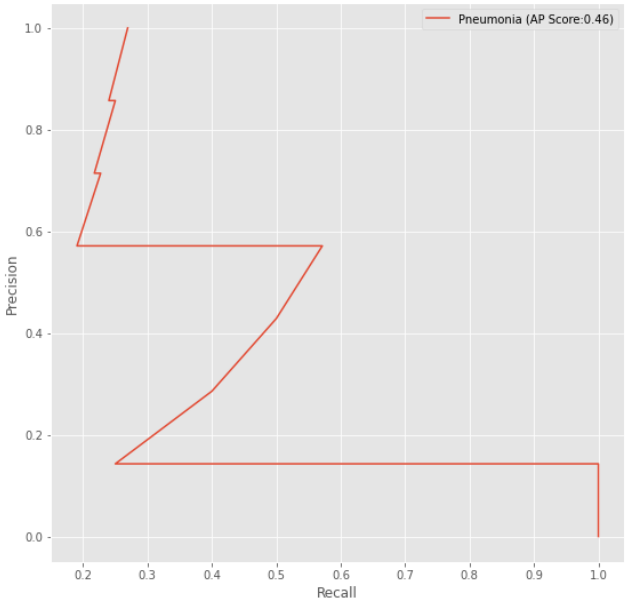

Figure 4: Precision-recall curve plot

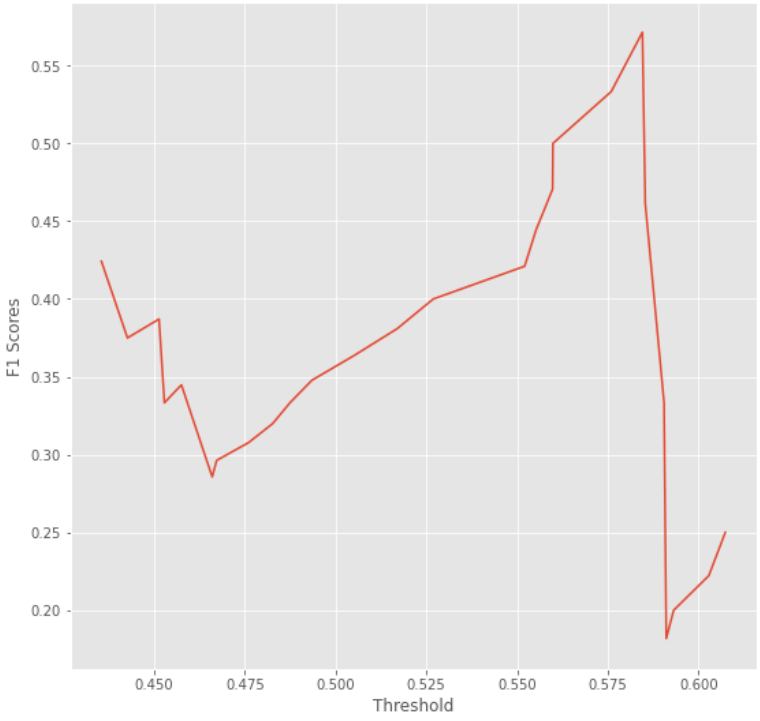

Figure 5: F1 vs Threshold plot

Final Threshold and Explanation

After 30 epochs of training it seemed the training and validation loss were not improving, and training was halted. The final performance after training was a training loss of 0.6925 a binary accuracy of 0.5450, a validation loss of 0.6869 and a validation binary accuracy of 0.5625

The threshold for classification was 0.58 and was chosen based on the above precision recall curve and F1-threshold plot, leading to a precision of around 0.57, a recall of around 0.57 and F1 score of around 0.57. The threshold was chosen to yield the best F1 score while providing an optimal threshold for classification.

4. Databases

The NIH chest X-ray dataset was used for analysis and training the model. The dataset consists of 112,120 X-ray images from 30,805 unique patients. The patient population for these images was between 1 and 95 years old. The images were captured through Digital Radiography ('DX') modality and can be split into the following view types: posteroanterior (PA) and anteroposterior (AP). Within the X-ray images, chest X-rays with Pneumonia findings were sometimes comorbid with the following other thoracic diseases: Atelectasis, Consolidation, Infiltration, Pneumothorax, Edema, Emphysema, Fibrosis, Effusion, Pleural thickening, Cardiomegaly, Mass, Nodule, and Hernia. In general, out of 112,120 X-ray images in the dataset only 1431 contained findings of Pneumonia, in addition, the top three thoracic diseases comorbid with Pneumonia were Infiltration (199 cases), Edema (137 cases) and Atelectasis (108 cases).

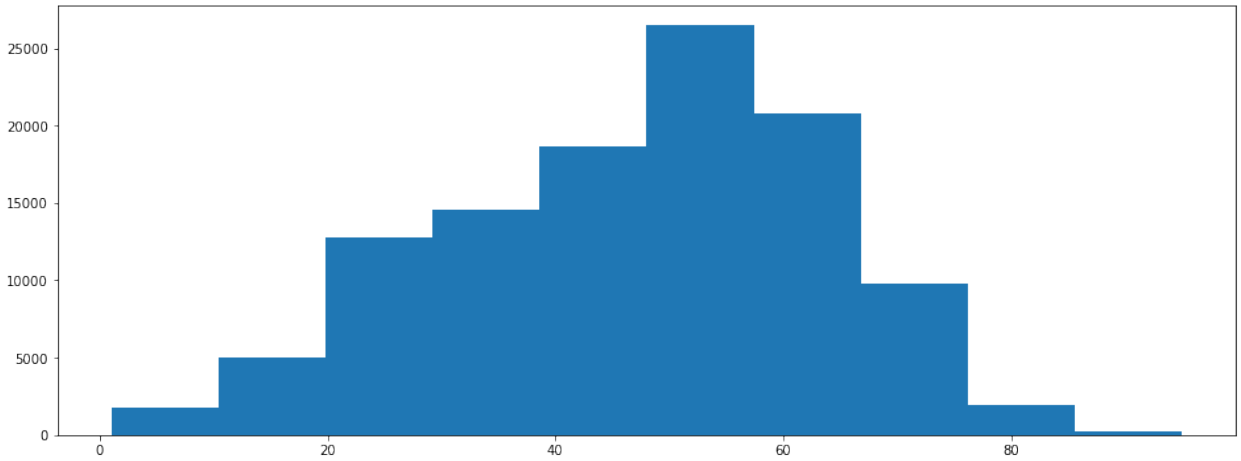

Figure 6: Age Distribution within NIH chest X-ray dataset

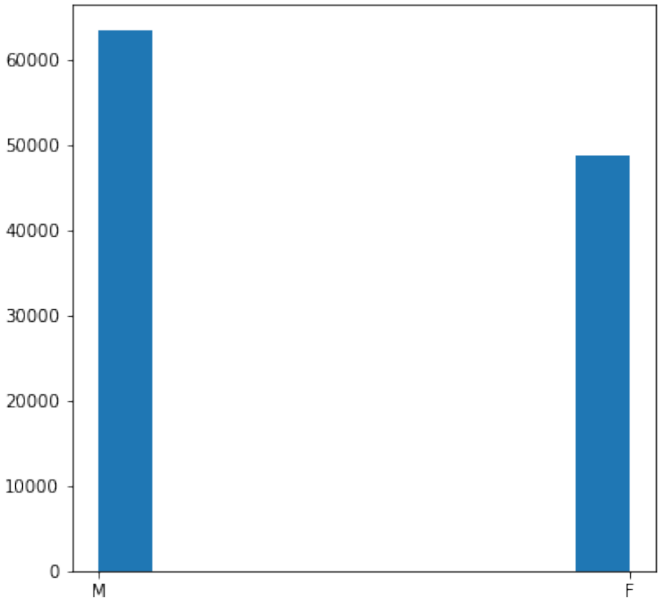

Figure 7: Gender Distribution within NIH chest X-ray dataset

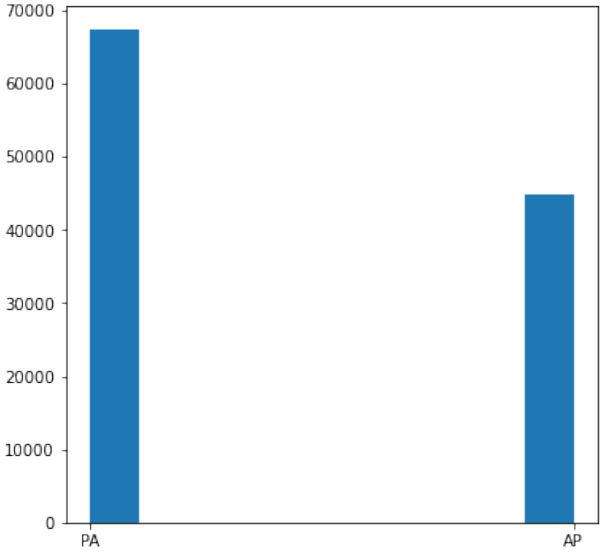

Figure 8: Distribution of image views within NIH chest X-ray dataset

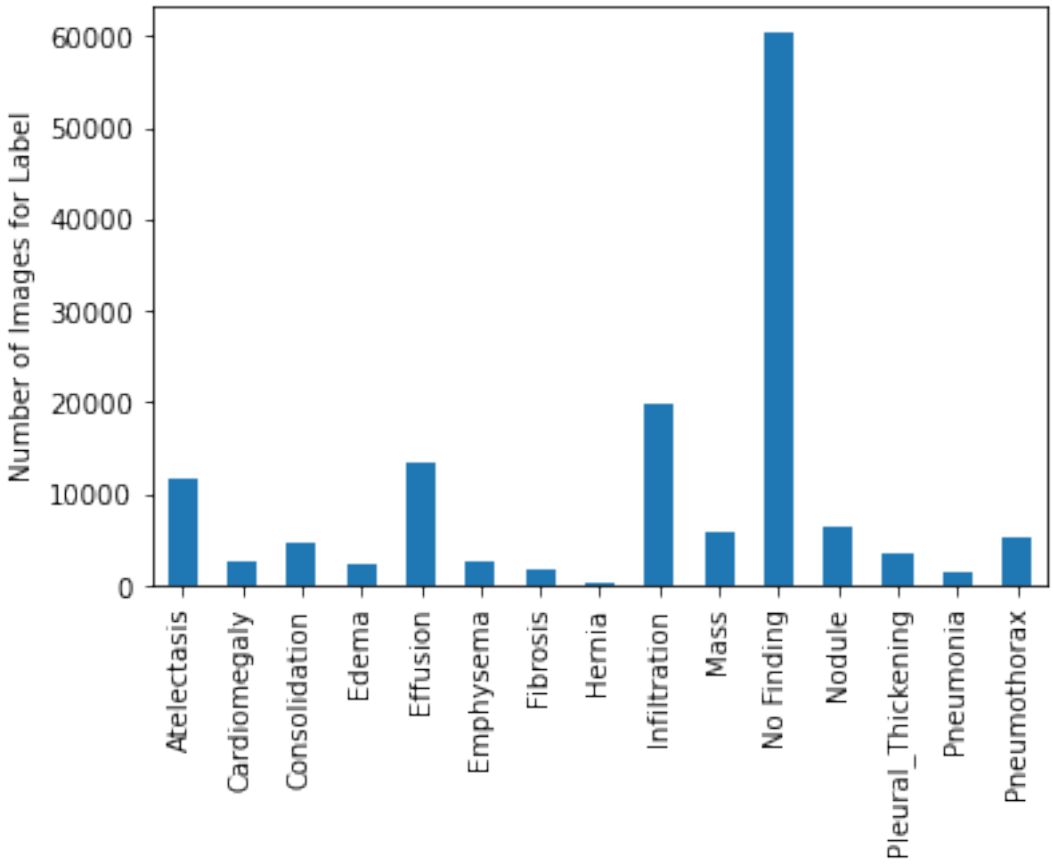

Figure 9: Distribution of finding labels within NIH chest X-ray dataset

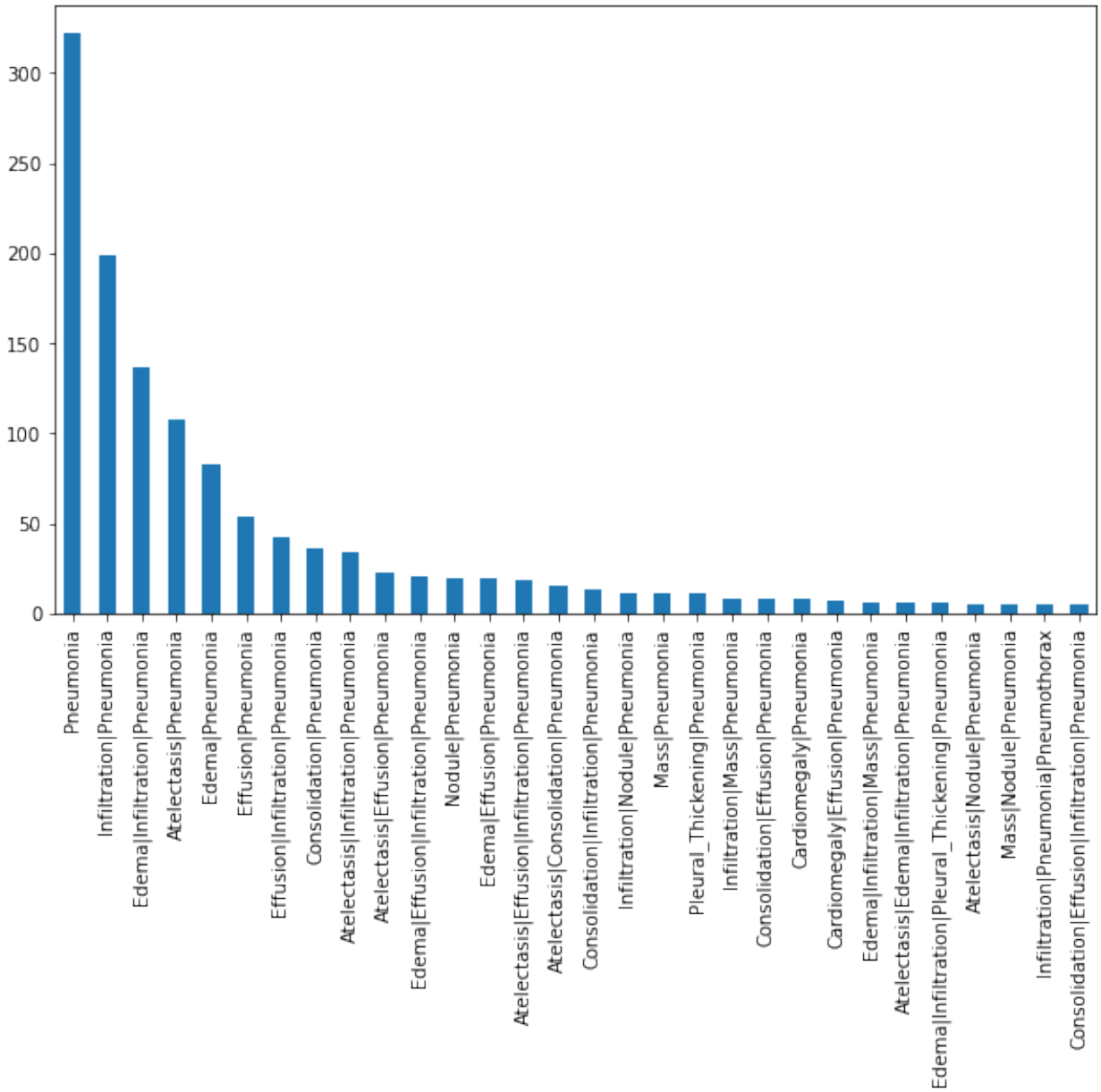

Figure 10: Top 30 Pneumonia finding labels within NIH chest X-ray dataset

Description of Training Dataset:

For training of 80% of the NIH chest X-ray dataset was selected. From these images it was determined that only 1145 images contained finding labels indicating Pneumonia, therefore the training dataset was reduced in size to 1145. An additional 1145 images with finding labels that did not contain Pneumonia were added to the dataset to create a balanced dataset for training.

Description of Validation Dataset:
For validation 20% of the NIH chest X-ray dataset or 22424 images were used. Of these images, only 1144 were used for the validation dataset. From the 1144 images, 286 contained finding labels of Pneumonia, to create an imbalance in the validation dataset to mimic how Pneumonia may appear in a random sample of chest X-rays, 3 times the number of Pneumonia cases were added to the 286 Pneumonia images, or 858 images. These non-Pneumonia images plus the Pneumonia images generated the dataset size of 1144.

5. Ground Truth

The image labels for the NIH chest x-ray dataset were (Natural Language Processing) NLP-extracted so there could be some erroneous labels, but the NLP labeling accuracy is estimated to be >90%.

6. FDA Validation Plan

Patient Population Description for FDA Validation Dataset

A Validation image dataset would include digital radiography posteroanterior (PA) and anteroposterior (AP) chest x-ray images on male and female patients between the age of 1 and 95. Pneumonia does not need to be prevalent in the dataset and finding labels in the dataset can indicate comorbidity with the following thoracic diseases: Atelectasis, Consolidation, Infiltration, Pneumothorax, Edema, Emphysema, Fibrosis, Effusion, Pleural thickening, Cardiomegaly, Mass, Nodule, and Hernia.

Proposed Ground Truth

As described in this paper: https://arxiv.org/pdf/1711.05225.pdf, a silver standard consisting of a group of independent practicing radiologists could be utilized to generate finding labels for a subset of chest X-ray images to obtain a better ground truth than NLP generated finding labels.

Performance criteria

F1 Scores for these radiologists could then be calculated within a given confidence interval, for example, 95%, and the algorithm's F1 score could be compared to the average of these F1 scores to determine the performance of the algorithm compared to the radiologists