

FDA Submission

Your Name: Felix Glushchenkov

Name of your Device: Pneumonia Detection Assistant from X-Ray Images

Algorithm Description

1. General Information

Intended Use Statement: To assist the radiologist in detecting pneumonia in X-Ray images.

Indications for Use:

Screening X-Rays for pneumonia.

Patient aged: 1-95

Patient 25th to 75th age quantiles: 35-60.

Patient gender: Men and Women

X-Ray properties:

- position: AP/PA

- modality: DX

- body part scanned: chest

Device Limitations:

The presence of other diseases is sometimes problematic. If the truth is there is no pneumonia, but there's effusion, consolidation, or edema, then the algorithm may think there's pneumonia.

Specificity in these cases is: 0.35, 0.3, and 0.15 respectively.

If there is pneumonia, and also either atelectasis, nodule, or fibrosis then the algorithm may think there is no pneumonia. Recall in such cases is: 0.5, 0.54, 0.45 respectively.

Clinical Impact of Performance:

As users, we are interested in the validity of what the model outputs - that is, if it says it's positive for pneumonia, how likely is that?

Similarly, if it says there's no pneumonia, with what probability is that true?

Note: tn: true negative, tp: true positive, fn: false negative, fp: false positive.

The first metric is called precision, and it is equal to:

$$\text{number of true positives} / \text{number of cases we think are positive} = tp / (tp + fp).$$

For model 1 and threshold=0.39: precision = $19/(19+38)=0.33$.

For model 2 and threshold=0.37: precision = $18/(18+43)=0.29$

The second metric is known as the negative predictive value and is calculated as:

$$\text{"number of true negatives} / \text{number of cases we think are negative} = tn / (tn + fn)"$$

For model 1: npv = $62/(62+9)=0.87$.

For model 2: npv = $57/(57+10)=0.85$

Another key metric is recall: among the positive cases in the data (tp & fn), how many are identified as positive?

Model 1 recall = $tp / (tp + fn) = 19/(19+9)=0.68$

Model 2 recall = 0.64

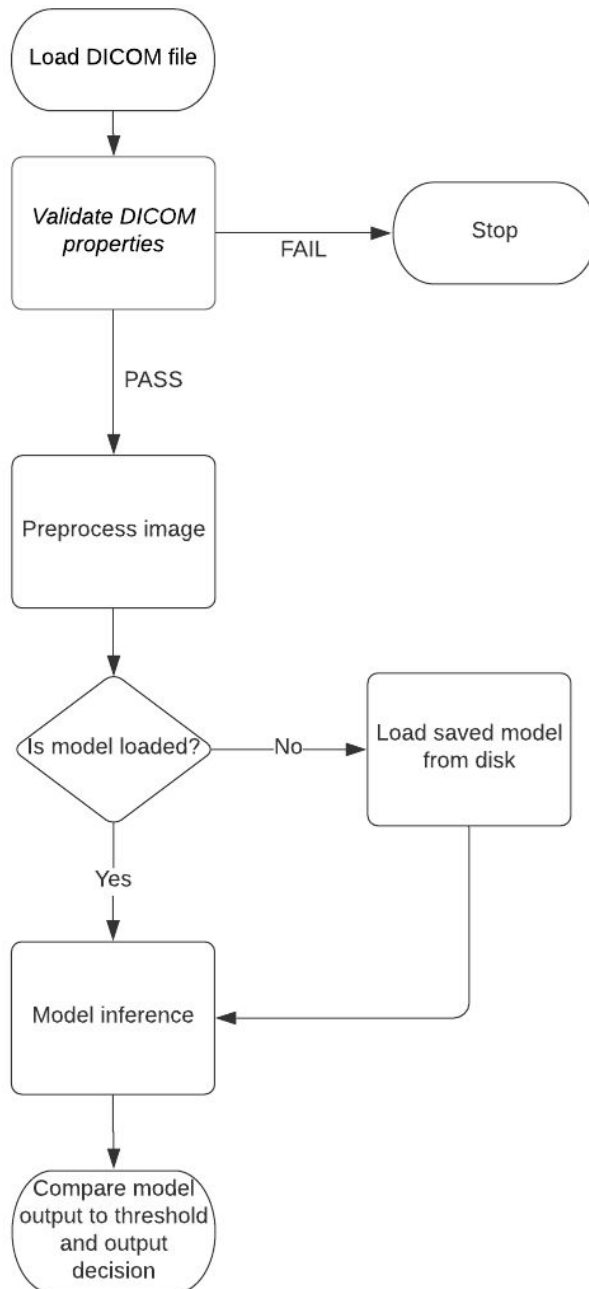
So, model 1 has slightly better metrics.

We are able to identify 68% of the positive cases in the dataset, but that comes with a lot of false positives due to low precision. However, we can identify 87% of the negative cases correctly with few false negatives. Thus, the value of the model lies in its detection of non-pneumonia images.

2. Algorithm Design and Function

Both models are VGG 16 pretrained networks but #1 has its last convolutional layer unfrozen. #2 is all frozen. Both models have 2 fully connected layers of 512 and 256 neurons and 1 output neuron stacked on top of the CNN. Since the first model had better performance, it will be described below.

Algorithm Flowchart



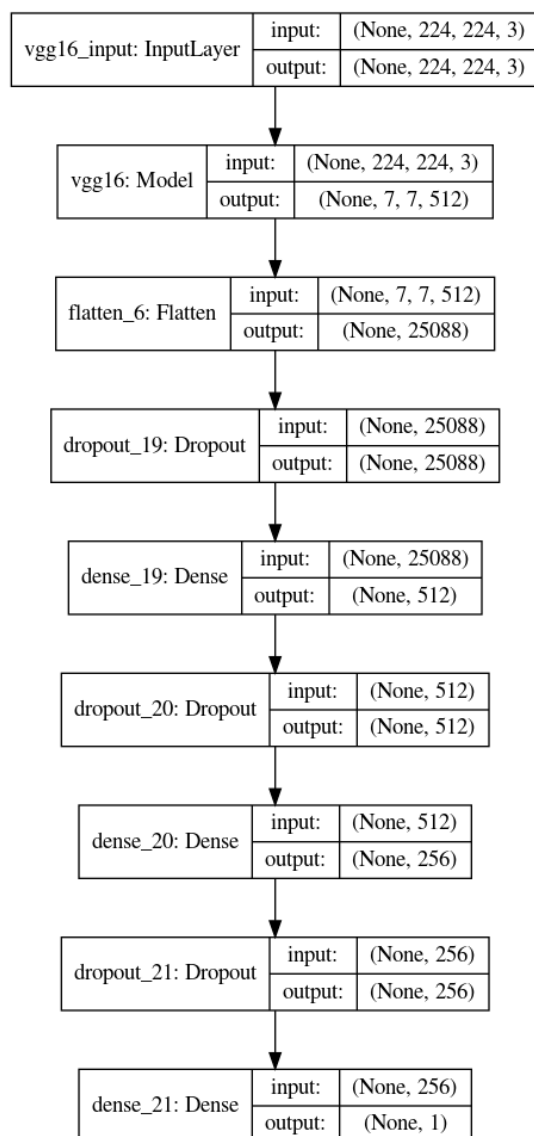
DICOM Checking Steps:

1. Check modality is an X-Ray scan (DX)
2. Check it's a scan of the chest
3. Check the view angle is PA or AP.

Preprocessing Steps:

1. Scale pixels to the interval [0,1]
2. Reshape and resize the image to (batch, height, width, channels)=(1, 224, 224, 3)
3. Pass the image through the Keras preprocess function for VGG 16 networks. It converts RGB to BGR, each channel is zero centered with respect to ImageNet, and does no scaling.

CNN Architecture:



3. Algorithm Training

Parameters:

Types of augmentation used during training:

- rescale pixels to [0,1]
- horizontal flip
- height shift range: 0.1
- width shift range: 0.1
- rotation range: 10
- shear range: 0.1
- zoom range: 0.1
- vgg16.preprocess_input()

Batch size: 32

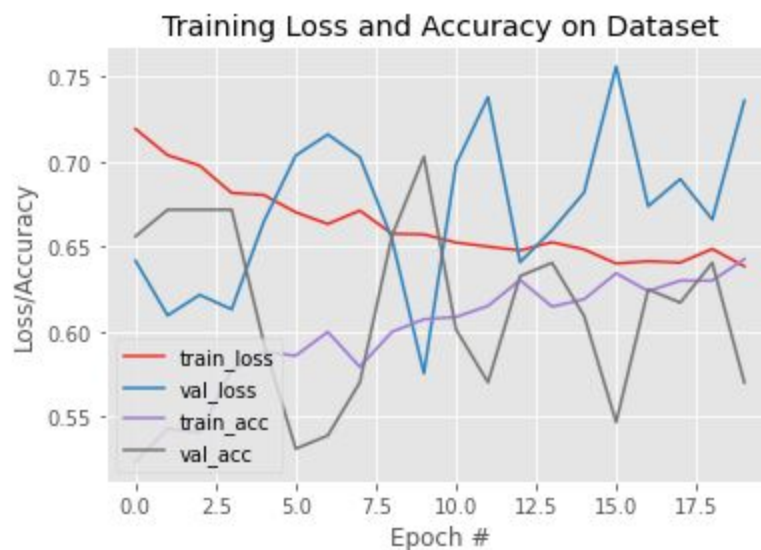
Optimizer learning rate: 0.00001

Layers of pre-existing architecture that were frozen: all layers of VGG16, except the very last convolutional layer

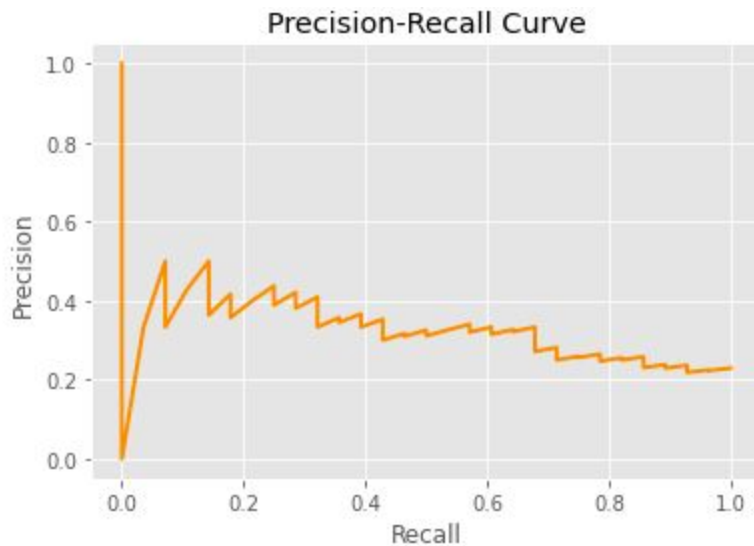
Layers of pre-existing architecture that were fine-tuned: the very last convolutional layer

Layers added to pre-existing architecture: flatten, dropout 0.4, dense 512, dropout 0.4, dense 256, dropout 0.4, dense 1.

Algorithm training performance visualization

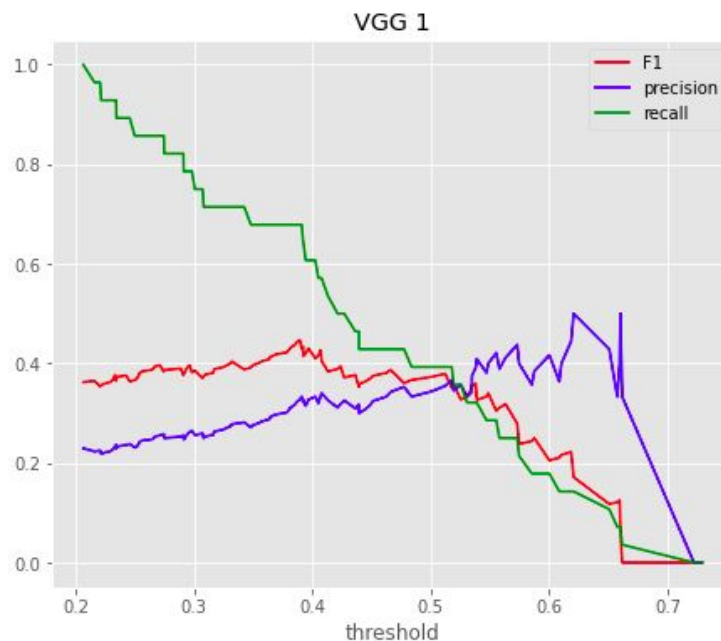


P-R curve



Final Threshold and Explanation:

The final threshold was 0.39. This struck a good balance between F1, precision, and recall as seen below:



Any lower, and F1 and precision decrease but recall increases. In fact false positives increase substantially as the threshold is dropped which may waste the radiologist's time. A general feature of the model is its ability to throw out negative cases due to its high negative predictive value, rather than correctly detect positive cases. At a threshold of 0.39, the NPV is at its maximum.

4. Databases

The dataset is 112120 images. A training and validation dataset were created from it by random sampling. The prevalence of pneumonia in this set is 1.3%. The minimum and maximum ages are 1 and 95, respectively. There were some outliers (ages that clearly don't make sense) that were filtered out. The majority of ages are between 20 and 60, however. In fact, the 25th and 75th percentiles are 35 and 59. Those with pneumonia are most often between 20 and 70, and 837 out of 1431 pneumonia-positive cases are males - that is slightly more than half. However, the overall dataset consists of more male xrays than female xrays with 63328 out of 112120, or 56%, being male.

Description of Training Dataset: The training set contains a 50-50 positive-negative split of pneumonia cases with 2290 images in total. This was created by selecting by first splitting the Pneumonia cases into 80-20 train-validation sets. Then for the train set, a random sample of positive-for-pneumonia quantity of negative-for-pneumonia cases were selected and thrown out, thus creating a 50-50 training set split.

Description of Validation Dataset: The validation set contains a 20-80 positive-negative split of pneumonia cases with 1430 images. This was created by randomly choosing negative-for-pneumonia cases that were 4 times larger in number than the positive-for-pneumonia cases.

5. Ground Truth

From the [Kaggle data source](#):

This 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.

This means the labels are mostly accurate for training. However, a more accurate alternative would be to have 3 radiologists verify the diagnoses and label the images, possibly weighted by years of experience. This would be a very lengthy and costly procedure for 112,120 images. Given the training required, one cannot simply crowdsource the labels like one can for common objects such as those found in ImageNet. Thus, if the NLP approach is 90% accurate, it is likely a good compromise.

Disease labels include: atelectasis, consolidation, infiltration, pneumothorax, edema, emphysema, fibrosis, effusion, pneumonia, pleural thickening, cardiomegaly, nodule, mass, and hernia.

6. FDA Validation Plan

Patient Population Description for FDA Validation Dataset:

Men and women, aged 1 to 95, whose chest was scanned with an x-ray machine in the PA/AP positions. Those patients with effusion, consolidation, or edema may not be suitable for ruling out pneumonia. Those patients with atelectasis, nodule, or fibrosis may not be suitable for ruling in pneumonia. The validation data set should contain a 20% prevalence of pneumonia.

Ground Truth Acquisition Methodology:

An optimal ground truth can be obtained via a sputum test (phlegm from a deep cough), pleural fluid culture test (fluid sample from the space between the tissues lining the lungs and chest cavity), or a bronchoscopy (tube with camera inserted into windpipe, with optional biopsy). However, these tests take a long time to conduct and are costly.

Algorithm Performance Standard:

This [paper](#) reports an F1 score of 0.435 (confidence interval: 0.387, 0.481) for their model and an average of 0.387 (CI: 0.330, 0.442) for 4 radiologists. They found their model is statistically significantly better than the radiologist performance by calculating the difference between the average F1 score of their model and the average F1 score of the radiologists. The 95% CI on the difference did not include 0, and so they concluded their model was a statistically significant improvement over the radiologist's performance. My model's maximum F1 score was 0.447, however the 95% confidence intervals were not calculated.