FDA Submission

Your Name: Widya Puspitaloka

Name of your Device: AI for Pneumonia Detection from Chest X-Rays

Algorithm Description

1. General Information

INTENDED USE STATEMENT:
Assisting radiologists with identifying pneumonia from chest X-Rays.

INDICATIONS FOR USE:
The algorithm is intended for use on any person from all ages, from children to elderly, who have been administered a screening pneumonia study on X-Rays from the chest in AP or PA position, regardless whether they have other comorbids or whether they are first time patients.

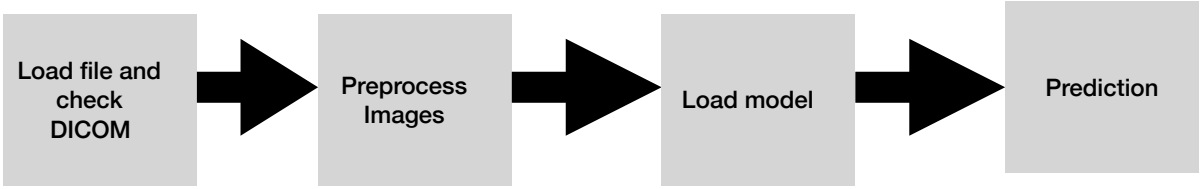
DEVICE LIMITATIONS:
The inference time depends on the processing speed of the CPU being used. The use of GPU is preferable to obtain the inference faster.

The presence of Edema and Infiltration might be a limitation of this algorithm since sometimes they occur together with pneumonia, thus it needs to be checked further when those two are present.

CLINICAL IMPACT OF PERFORMANCE:
The algorithm has high recall and low precision, simultaneously low F1 score. It indicates that there are more false positive case when the algorithm detects pneumonia, although it is not the case.

This result might be more acceptable than false negative in assisting the radiologist in determining pneumonia cases since it might be better to misclassify and to be checked further instead of missing the diagnosis leading to late treatment.

2. Algorithm Design and Function



DICOM CHECKING STEPS:
- Check image position: PA or AP
- Check image type/ modality: DX
- Check body part: Chest
- Show the finding to be compared with the prediction

- Show image

PREPROCESSING STEPS:
- Image is rescaled: 1/ 255
- Image is normalized by mean of 0 and standard deviation of 1
- Image is reshaped.

CNN ARCHITECTURE:
The algorithm used a pre-trained network, VGG16. To fine tune the model, some additional layers are added.
- The output of the model is flattened
- Dense (full-connected) layer is also added.
- Sigmoid activation function is added so output of the last layer is in the range of [0,1]
- Dropout layer is used which may prevent overfitting and improve generalization ability to unseen data.

Summary of the model is shown below:

```
_________________________________________________________________
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, 1)                 257
=================================================================
Total params: 41,062,209
Trainable params: 28,707,329
Non-trainable params: 12,354,880
_________________________________________________________________
```

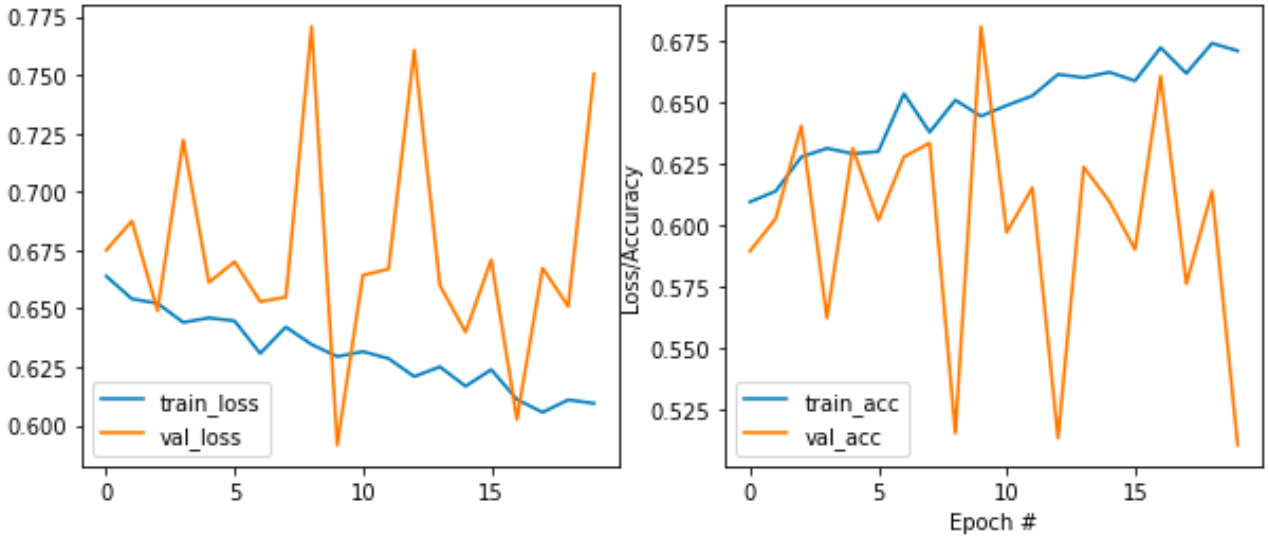
3. Algorithm Training

PARAMETERS:

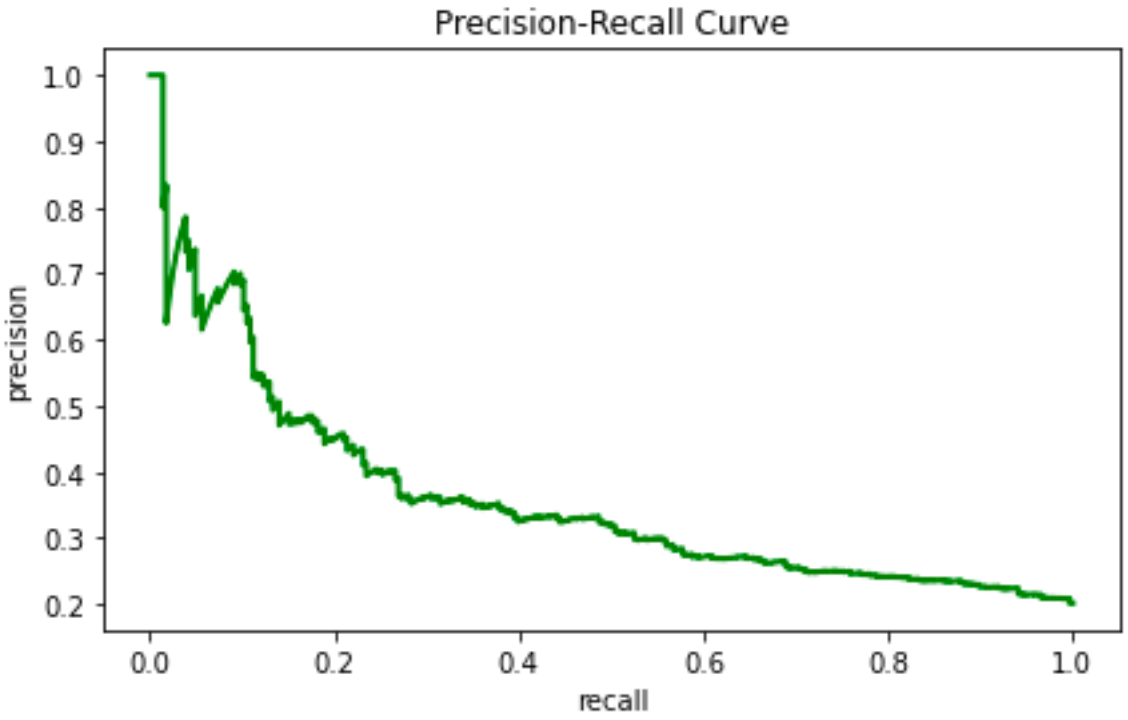
- Types of augmentation used during training
  - Horizontal flip
  - Height shift range 0.1
  - Width shift range 0.1
  - Rotation range 20
  - Sheer range 0.1
  - Zoom range 0.1
- Batch size = 32
- Optimizer learning rate = 1e-4
- Layers of pre-existing architecture that were frozen: freeze all (17 layers) but the last 2 last convolutional layer of VGG16.

- Layers of pre-existing architecture that were fine-tuned: last 2 convolutional layers which are block5conv3 and block5_pool.
- Layers added to pre-existing architecture: flatten, dense, and dropout layer.

ALGORITHM TRAINING PERFORMANCE VISUALIZATION

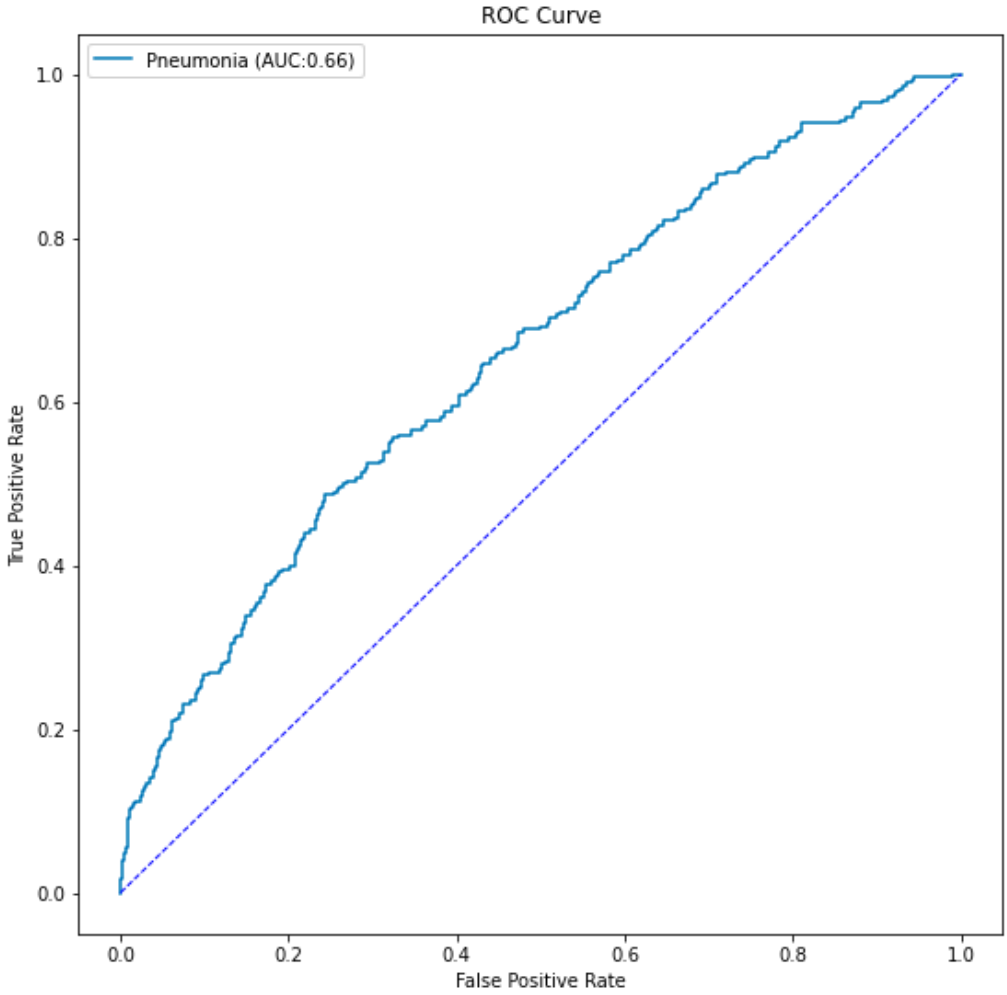


PRECISION-RECALL CURVE

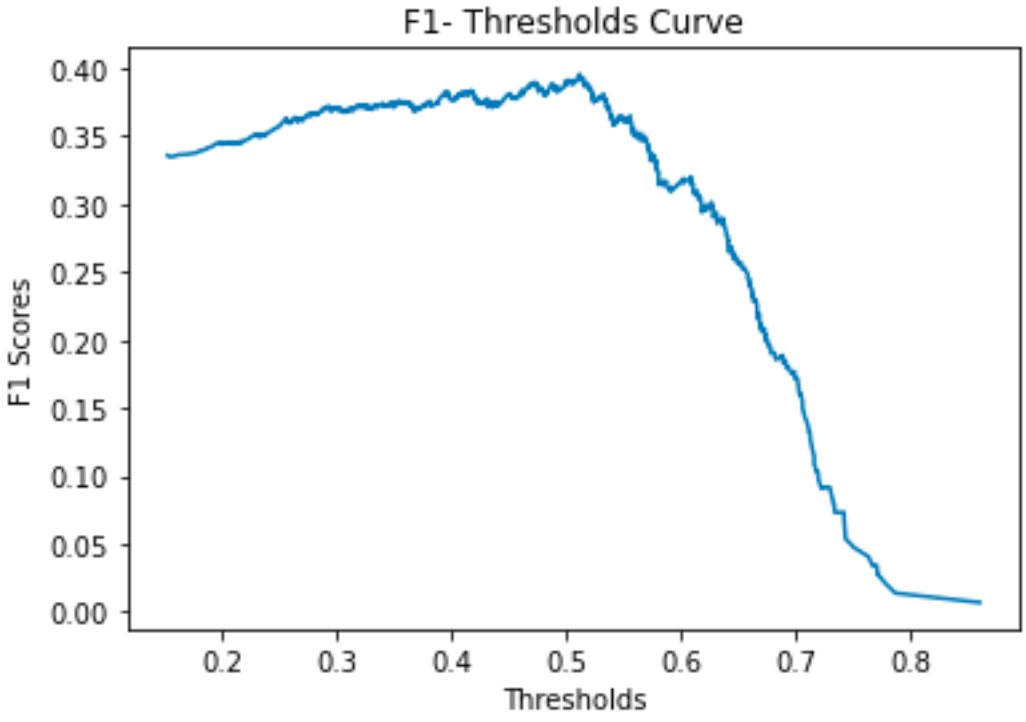


From the graph, as the recall increases, precision decreases.

AUC-ROC CURVE



FINAL THRESHOLD AND EXPLANATION



As shown from the graph, we get:

- Maximum f1: 0.3954
- Threshold: 0.5112

The maximum F1 score is 0.395 with the threshold of 0.51 as shown above.

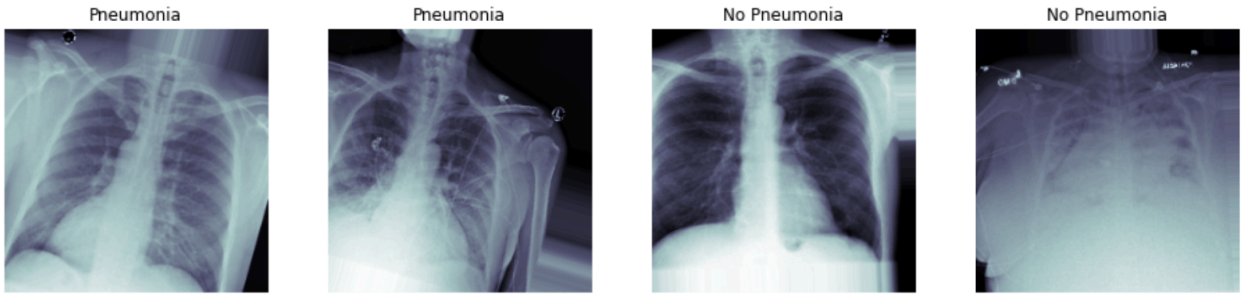
The final threshold used is 0.50.

4. Databases

The training and validation dataset is provided by Udacity from NIH Chest X-Ray dataset which can be downloaded from Kaggle.

DESCRIPTION OF TRAINING DATASET:
There are 2290 total samples for the training dataset. The dataset has equal number of positive pneumonia (1145 samples) and non-pneumonia (1145 samples).

A few examples of the images from the training dataset:



DESCRIPTION OF VALIDATION DATASET:
There are 1430 total samples for the training dataset. The dataset has 80% non-pneumonia cases (286) and 20% pneumonia cases (1144) to resemble the case in the real world setting — where no finding of pneumonia is much more common.

5. Ground Truth

Often times, the gold standard is unattainable for an algorithm developer. In this case, the gold standard for the detection of pneumonia is by obtaining sputum cultures to test the presence of the pneumoccocus bacteria that causes pneumonia.

Since the algorithm is intended to be used to assist radiologist and not for a replacement, thus a silver standard approach can be used as the ground truth by taking weighted judgment of several radiologists to each make their own diagnosis of an image. The final diagnosis is then determined by a *voting* system across all of the radiologists' labels for each image.

On top of the ground truth, the algorithm must be compared to a performance standard from other research.

6. FDA Validation Plan

PATIENT POPULATION DESCRIPTION FOR FDA VALIDATION DATASET:
To validate the algorithm, we want to make sure that the FDA Validation Dataset includes all age group and gender, with the X-Ray image taken from the chest from an AP or PA position.

We want to make sure the pneumonia case in the dataset is reflective of the distribution of those cases that are seen in the real world.

GROUND TRUTH ACQUISITION METHODOLOGY:
Since the algorithm is intended to be used to assist radiologist, thus a silver standard approach can be used as the ground truth by taking weighted judgment of several radiologists to each make their own diagnosis of an image.

ALGORITHM PERFORMANCE STANDARD:
The standard performance of the algorithm is necessary to support our justification on using this algorithm. Literature searching can be conducted to find the sources as the standard that we can compare to.

This following paper CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning showed F1 score from four radiologists and the F1 score with the mean of 0.387 and from their algorithm (CheXNet) which is 0.435 as shown below.

	F1 Score (95% CI)
Radiologist 1	0.383 (0.309, 0.453)
Radiologist 2	0.356 (0.282, 0.428)
Radiologist 3	0.365 (0.291, 0.435)
Radiologist 4	0.442 (0.390, 0.492)
Radiologist Avg.	0.387 (0.330, 0.442)
CheXNet	0.435 (0.387, 0.481)

The algorithm that we have created should, at least, be better than the radiologist average from the paper. It is shown that the maximum F1 score this algorithm can have is 0.40 which is comparable to CheXNet and better than the average of radiologist.