FDA Submission

Your Name: SETH RAMACHANDIRAN

Name of your Device: Pneumonia ASSISTANCE DETECTOR

Algorithm Description

1. General Information

Intended Use Statement: Assisting radiologists in detecting pneumonia in Chest X-ray images with PA/AP views. It is intended to be used with radiologist review of X-Ray, Sputum cultures and review of medical history for diagnostic validation

Indications for Use: It is used for both male and female for 1-100 years old with or without the diseases in comorbid with Pneumonia. It may be used for other assistance with co-morbid like Atelectasis, Cardiomeagaly, Edema, Effusion, Emphysema, Fibrosis, Infiltration, Nodule, Pleural Thickening, Pneumothorax etc to the extent X-Rays are used in detecting these diseases.

Device Limitations: Require high-power processing GPU card to run the algorithm faster. The solution is run in both high-powered cloud system as well as a Lenovo P620 Thinkstation with one single GPU card is used for comparison. The single P620 Thinkstation is usable solution providing technology independence from cloud and the model is run in such system. The model took ~300 secs in this case.

The performance of the device is limited by the fact many of the diseases exhibit similar X-Ray characteristic and can lead to wrong classification – for example the infiltration is very close to the Pneumonia. Hence in such cases to rule out other diseases important to have additional tests.

Training LIMITATION: Based on the training accuracy and the validation **accuracy is at a lower 60%.** If the accuracy is at 95% then the Device could be much more powerful tool but with this accuracy needs to be used with caution.

Clinical Impact of Performance:

The goal of this device is to assist the radiologist and a radiologist will be reviewing the result and validate the results. When the device predicts false positives, the sputum culture test or other tests/medical history can be used to validate the condition. False negative result will adversely impact the patient and hence the radiologist review is very important. The clinician can order for the sputum test and verify the results along with his/her own analysis.

2. Algorithm Design and Function

LOAD MODEL

Dicom CHECKING

PREPROCESSING IMAGE

PREDICT WITH MODEL.

MODEL: VGG19 MODEL ARCHITECTURE WITH TRANSFER LEARNING.

The VGG19 is one of the old CNN algorithms. It is trained with publicly available X Ray, which is labelled, and 112000 x-rays were used to train the model. The VGG19 is used with its own weights from its pre-trained model but last few layers are customised and used for the purposes of this exercise.

The model was trained in the Udacity cloud with GPU and with the Lenovo P-Thinkstation. Each epoch in the cloud took about 56s whereas in Lenovo P620 Thinkstation about 240s.

DICOM Checking Steps: Perform 3 different DICOM checks

- Check patient position: Only AP or PA view will be processed
- Check image type (modality): Only DX type will be processed
- Check body part examined: Only chest taken image will be process

Preprocessing Steps: Rescale the image by dividing by 255.0, then normalize the image with the mean and standard deviation retrieved from the training data. Finally, resize the image to (1, 224, 224, 3) to fit in the network.

CNN Architecture:

- VGG19 model architecture is used for transfer learning.

- Several custom layers are also added to the VGG19.

3. Algorithm Training

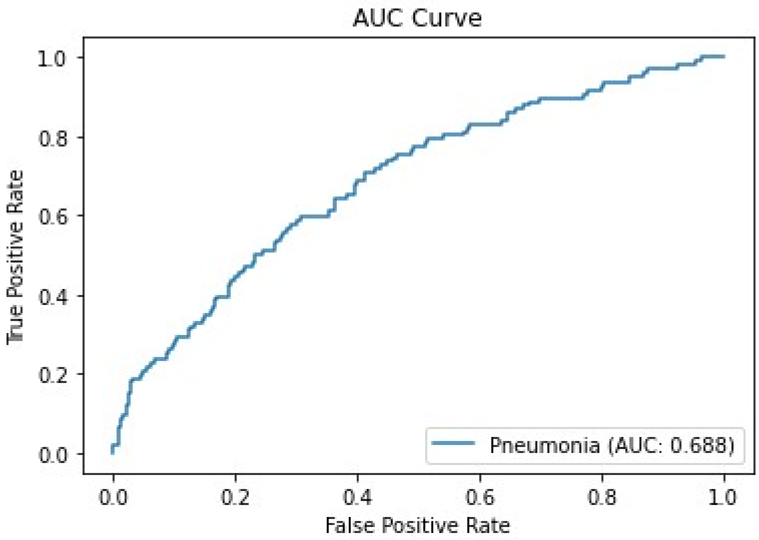
Parameters:

- Augmentation used:
    - Horizonal Flip

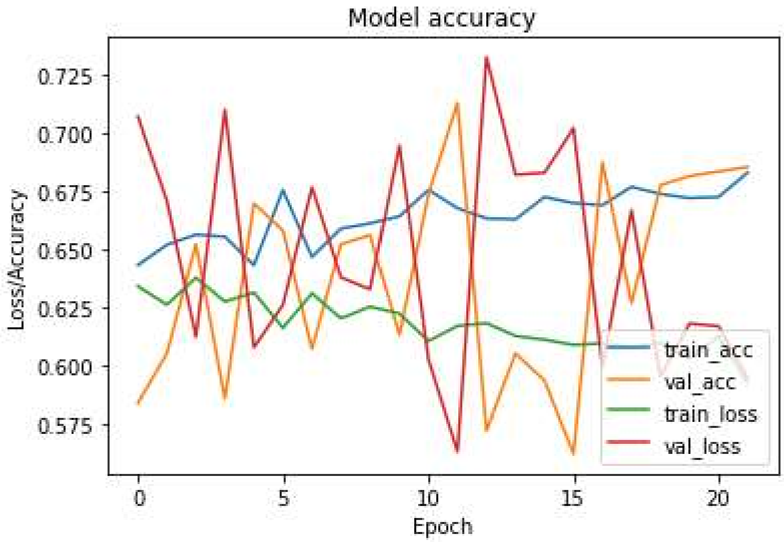
  - Height Shift Range = 0.1
  - Width Shift Range = 0.1
  - Rotation Range = 20
  - Shear Range = 0.1
  - Zoom Range = 0.1

- Batch size = 32
- Optimizer learning rate = 3e-4
- Layers of pre-existing architecture that were frozen: First 20 layers
- Layers of pre-existing architecture that were fine-tuned: None
- Layers added to pre-existing architecture:
  - Flatten
  - Dropout 0.5
  - Dense 1024, Activation = ReLU
  - Dropout 0.5
  - Dense 512, Activation = ReLU
  - Dropout 0.5
  - Dense 256, Activation = ReLU
  - Dense 1, Activation = Sigmoid

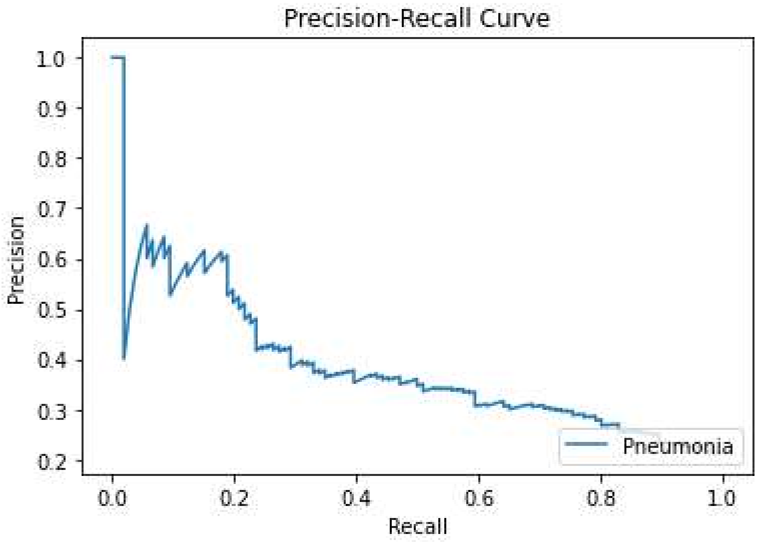
AUC Curve



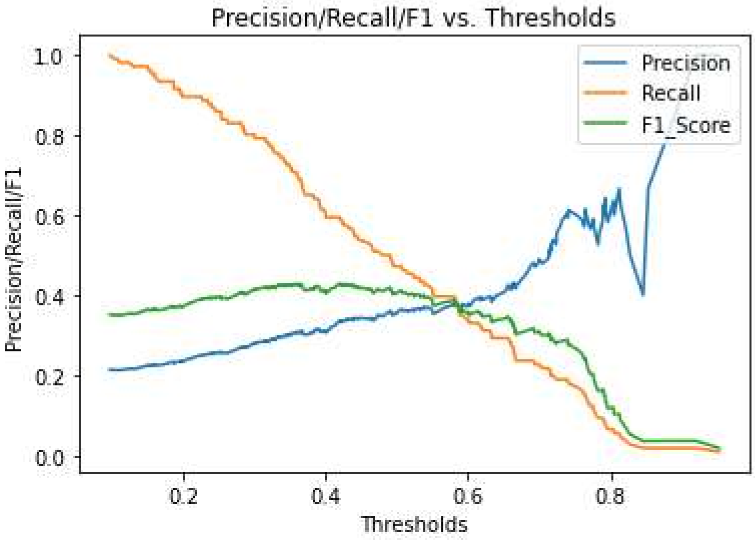
Model Accuracy and loss



Precision Recall Curve



Final Threshold and Explanation:



- `Max F1 Score: 0.42979942693409734`
- `Threshold corresponding to max F1: 0.35952678322792053`
- Based on the F1-Score vs Threshold Chart, to balance between the Precision and Recall, the threshold of 0.359 will give the max value of F1-Score.

Training LIMITATION: Based on the above curve and analysis the training accuracy and the validation **accuracy is at a lower end.** If the accuracy is at 95+ then the Device could be much more powerful tool.

4. Databases

- The databases contains 112,120 X-Ray images. The number of Pneumonia Positive images is only 1430 (1.27%).
- Therefore, to split the databases for training, I will have to:
  - Obtain all the postive cases of Pneumonia.
  - Divide the positive cases into 80%-20% for the Training and Validation Dataset.

5. Ground Truth

- The ground truth is NLP-derived labels. NLP at this stage is not complex enough to capture all the existing information of the reports. Hence, the accuracy is roughly 90%.

6. FDA Validation Plan

Patient Population Description for FDA Validation Dataset:

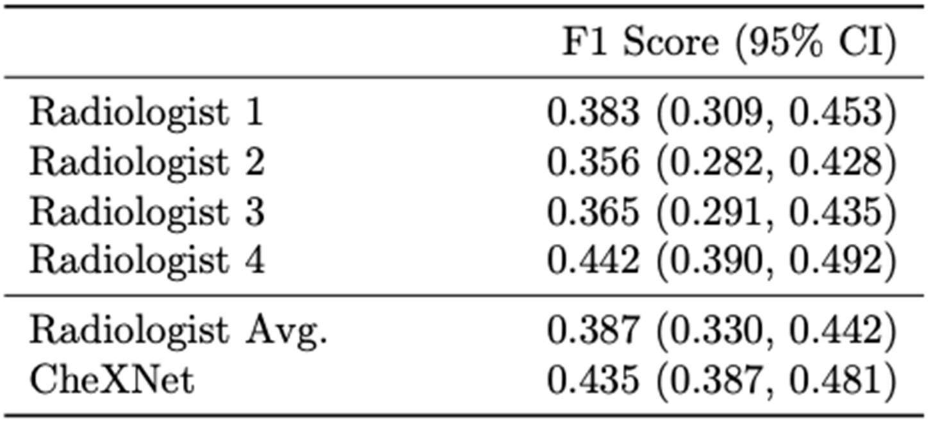
- Male and female patients in the age of 1 to 100. The gender distribution is slightly toward Male patient, the male to female ratio is approximately 1.2

- The patient may exihibit the following comorbid with Pneumonia: Atelectasis, Cardiomegaly, Consolidation, Edema, Effusion, Emphysema, Fibrosis, Hernia, Infiltration, Mass, Nodule, Pleural_Thickening, Pneumonia, Pneumothorax -
- The X-Ray Dicom file should has the following properties:
  - Patient Postition: AP or PA
  - Image Type: DX
  - Body Part Examined: CHEST

Ground Truth Acquisition Methodology:

- Establish a silver standard of radiologist reading

Algorithm Performance Standard:



	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 F1-Score should be approximately 0.435 to out-perform state-of-the-art method (CheXNet) [https://arxiv.org/pdf/1711.05225.pdf]