FDA Submission

Your Name: Ish Girwan
Name of your Device: X-Ray Pneumonia detector

Algorithm Description

1. General Information

Intended Use Statement:
Helping Radiologists in detecting Pneumonia in Chest X-Rays

Indications for Use:
Applicable for all age groups and genders. The X-Ray image should be for Chest taken in the AP or PA position. In a clinical setting, the algorithm should be integrated into the workflow of the diagnostic clinics. The X-Ray images should be in DICOM format, respecting the HIPAA rules. The data is preprocessed and is also checked for the required conditions. If it satisfies the required checking criteria, inference is conducted. Once the prediction is complete, this data is sent to the radiologist who will further analyse the made prediction to come to any conclusion.

Device Limitations:

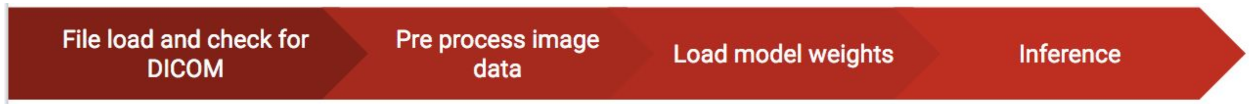
- Inference can be conducted on an average CPU. The inference time would vary depending on the processing speed of the CPU used.
- The dataset showed relatively strong correlation with Edema and Infiltration. So, the results might not be trusted in the presence of diseases.

Clinical Impact of Performance:
The model has lower precision and higher recall which signifies that there would be more false positives and less false negatives. So, this tool should be used only to assist clinicians island not as a final conclusion.

Having a false negative prediction can be life threatening for the patient as the disease won't be diagnosed. While having a false positive would lead to further diagnosis which could be expensive and also eat up the resources of the hospital which could have been used to save more lives.

2. Algorithm Design and Function

Algorithm Flowchart



DICOM Checking Steps:

- Body part is Chest
- Patient position is either PA or AP
- Modality is DX

Preprocessing Steps:

- Image is normalized
- Image is reshaped
- Image is repeated across 3 channels

CNN Architecture:

The algorithm uses a pretrained VGG16 model. The model's output is flattened and passed through dense and dropout layers as shown in the model summary.

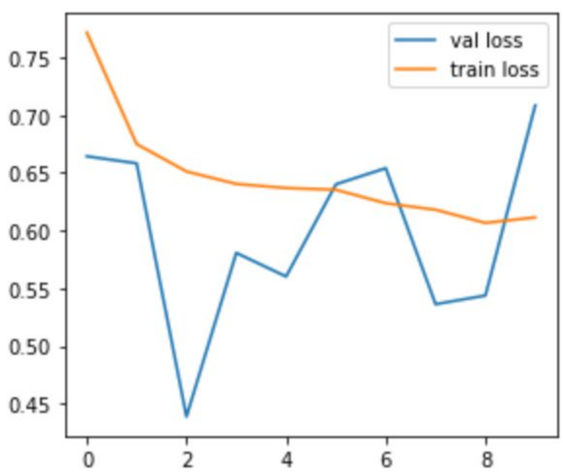
```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
model_1 (Model)              (None, 7, 7, 512)         14714688
_________________________________________________________________
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, 1)                 1025
=================================================================
Total params: 40,406,849
Trainable params: 28,051,969
Non-trainable params: 12,354,880
_________________________________________________________________
```

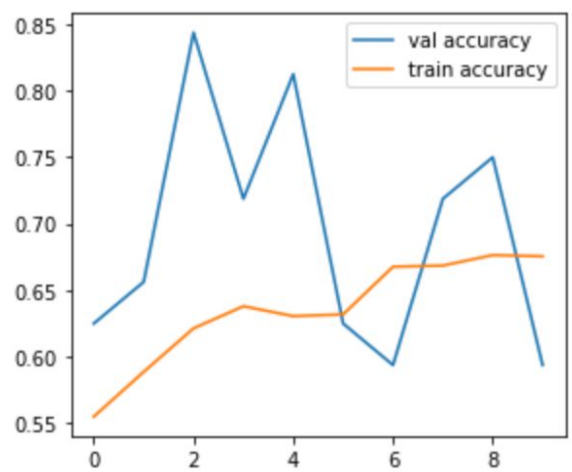
3. Algorithm Training

Parameters:

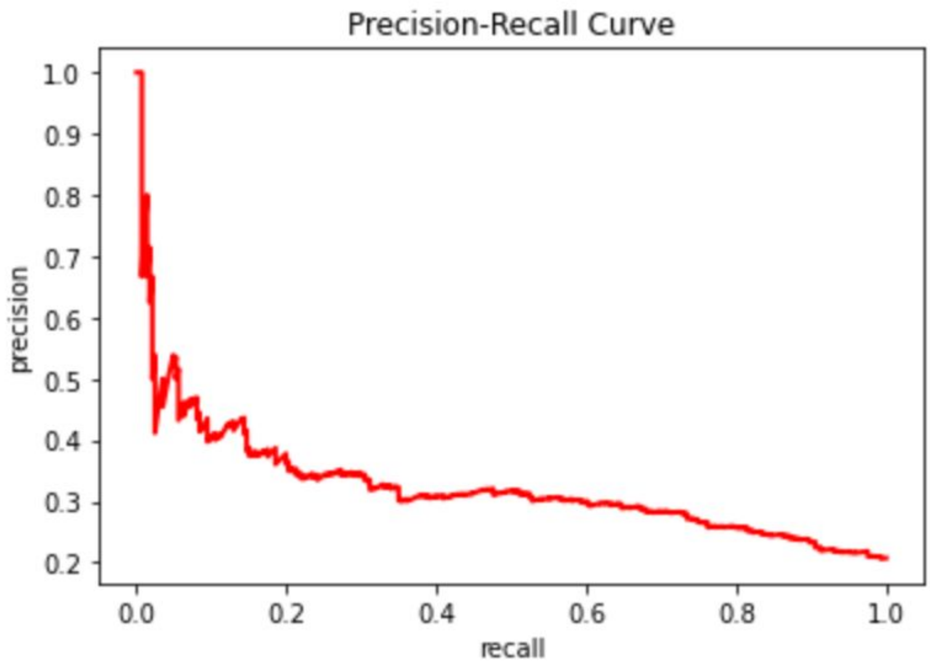
- Types of augmentation used during training
    - Horizontal flip
    - Vertical flip
    - Height shift range of 0.1,
    - Width shift range of 0.1,
    - Rotation range of 20,
    - Shear range of 0.1,
    - Zoom range of 0.1
- Batch size: 32
- Optimizer learning rate: 1e-4
- Layers of pre-existing architecture that were frozen: first 17 layers of VGG model
- Layers of pre-existing architecture that were fine-tuned: last 2 layers which are block5_conv3 anf block5_pool.
- Layers added to pre-existing architecture: flatten dense and dropout layer as shown in the architecture.

Algorithm training performance visualization



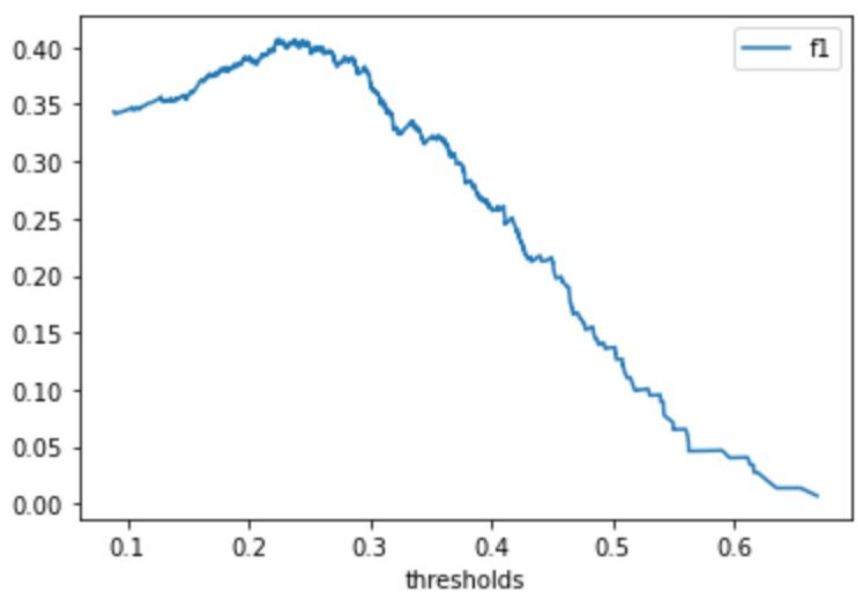


P-R curve



Final Threshold and Explanation:

Final threshold was selected where the F1 score was maximum.
Maximum F1: 0.4070450097847358
Threshold: 0.22431683540344238



The F1 Score from the paper [CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning](#).

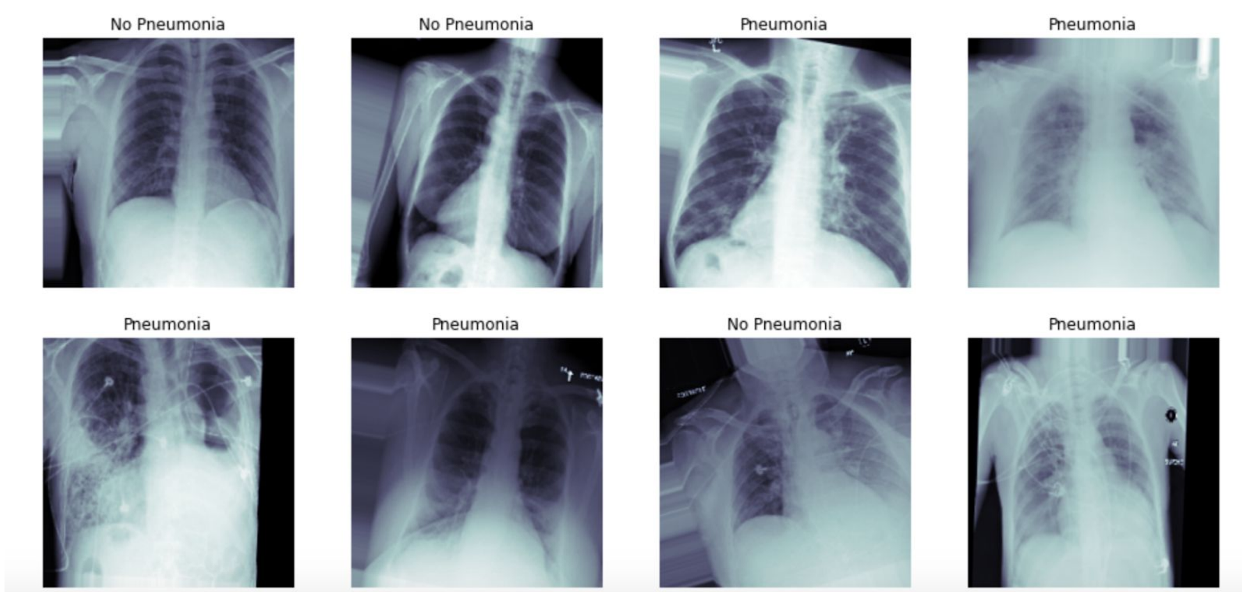
	F1 Score (95% CI)
Radiologist 1	0.383 (0.309, 0.453)
Radiologist 2	0.356 (0.282, 0.428)
Radiologist 3	0.365 (0.291, 0.435)
Radiologist 4	0.442 (0.390, 0.492)
Radiologist Avg.	0.387 (0.330, 0.442)
CheXNet	0.435 (0.387, 0.481)

The model's F1 score is better than most radiologists and comparable to that of CheXNet.

4. Databases

The training and validation set is from NIH Chest X-ray Dataset.

Description of Training Dataset:
The training dataset is equally split between Pneumonia samples and non Pneumonia samples. There are a total of 2290 samples. Some samples from the training set:



Description of Validation Dataset:
The validation dataset has 20% Pneumonia samples and 80% non Pneumonia samples to represent a real world setting. There are a total of 1430 samples.

5. Ground Truth

There are 112,120 X-ray images with disease labels from 30,805 unique patients in this dataset. The disease labels were created using Natural Language Processing (NLP) to mine the associated radiological reports. The labels include 14 common thoracic pathologies:

- Atelectasis
- Consolidation
- Infiltration
- Pneumothorax
- Edema
- Emphysema
- Fibrosis
- Effusion
- Pneumonia
- Pleural thickening
- Cardiomegaly
- Nodule
- Mass
- Hernia

The biggest limitation of this dataset is that image labels were 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:
The FDA Validation Dataset would include all age groups and genders. But the X-Ray image should be for Chest taken in the AP or PA position. The dataset showed relatively strong correlation with Edema and Infiltration so the validation set may not include data from patients who have these diseases.

Ground Truth Acquisition Methodology:
The gold standard is by obtaining sputum cultures to test for the presence of bacteria or viral bodies that cause pneumonia. Since these tests can be expensive and time taking, we can rely on a silver standard. The silver standard would be taking a weighted judgment on a sample from several Radiologists. The final label would be weighted by the experience of Radiologists.

Algorithm Performance Standard:
The algorithm’s F1 score should be more than that of average Radiologists(0.387) as mentioned in the ChesXNet paper.