FDA Submission

Your Name: Saddam Chowdhury

Name of your Device: AI X-Ray Pneumonia Detection Assistant

Algorithm Description

1. General Information

Intended Use Statement:

For assisting radiologists in detecting the presence of Pneumonia in Chest X-Rays.

Indications for Use:

Applicable for all age groups and genders. The X-Ray image should be for the chest taken in 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, the data is sent to the radiologist who will further analyze the predictions generated by the model to reach any conclusion.

Device Limitations:

The dataset showed relatively strong correlation with Edema and Infiltration. In the table below, it is shown how the model performs in the absence of the comorbid thoracic pathologies indicated in the left-most column.

absent	auc	f1	thresh	precision	recall
None	0.717	0.446	0.406	0.322	0.724
Emphysema	0.716	0.45	0.406	0.326	0.722
Atelectasis	0.735	0.449	0.406	0.322	0.737
Nodule	0.72	0.456	0.406	0.331	0.73
Hernia	0.716	0.445	0.406	0.321	0.723
Consolidation	0.729	0.446	0.406	0.323	0.721
Pleural_Thickening	0.717	0.444	0.406	0.32	0.719
Cardiomegaly	0.721	0.457	0.406	0.334	0.723
Edema	0.671	0.36	0.406	0.249	0.647
Mass	0.724	0.448	0.42	0.329	0.701
Fibrosis	0.718	0.45	0.406	0.325	0.728
Effusion	0.72	0.431	0.409	0.315	0.678
Pneumothorax	0.721	0.457	0.406	0.334	0.718
Infiltration	0.687	0.341	0.408	0.233	0.634

The above analysis suggests that the results might not be trusted in the presence of Edema and Infiltration.Therefore, the model is recommended for use when these conditions are not present.

It is noteworthy that chest x-rays are used for initial diagnosis of pneumonia, and the location and extent of inflammation in the lungs caused by the disease. However, chest x-rays cannot reliably differentiate bacterial from a non-bacterial cause.

Clinical Impact of Performance:

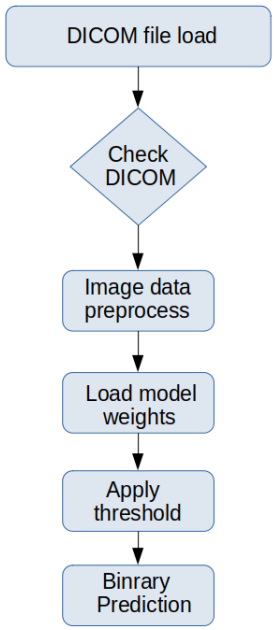
The model has lower precision (29%) and higher recall (80%). The corresponding negative prediuctive value (NPV) is 90%. This implies that

- If the model predicts a negative, it is correct with 90% probability
- If the model predicts a positive, it is correct with 29% probability

Therefore, the algorithm is recommended for assisting a radiologist to screen images with negative results. It is noteworthy that the algorithm can still incorrectly predict a negative with 10% probability. So, the negative cases predicted by the model should still be reviewed by the radiologist.

2. Algorithm Design and Function

Algorithm Flowchart:



DICOM Checking Steps:

The algorithm performs the following checks on the DICOM image:

- Patient Age is between 2 and 90
- Body part is Chest
- Patient position is either PA or AP
- Modality is DX

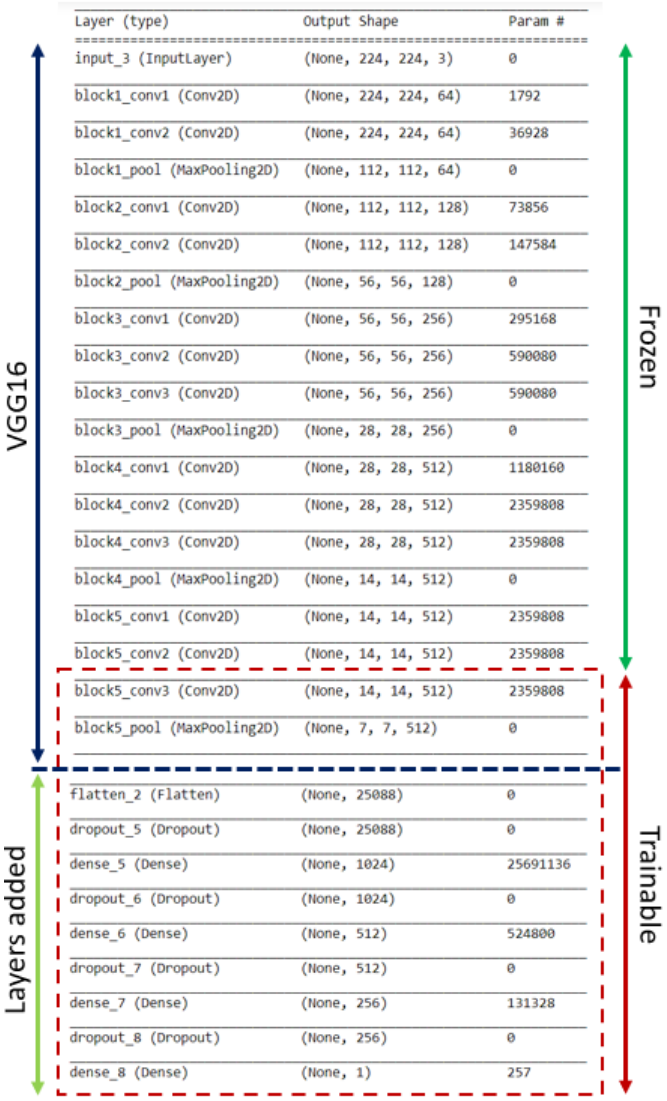
Preprocessing Steps:

The algorithm performs the following preprocessing steps on an image data:

- Converts RGB to Grayscale (if needed)
- Re-sizes the image to 224 x 224 (as required by the CNN model)
- Normalizes the intensity to be between 0 and 1 (from original range of 0 to 255)
- Repeats image across 3 channels

CNN Architecture:

The algorithm uses a pre-trained VGG16 Neural Network (except the last block of Convolution + Pooling layers that was re-trained), with additional 4 blocks of 'Fully Connected + Dropout' layers. The network output is a single probability value for binary classification. Below is the CNN architecture:



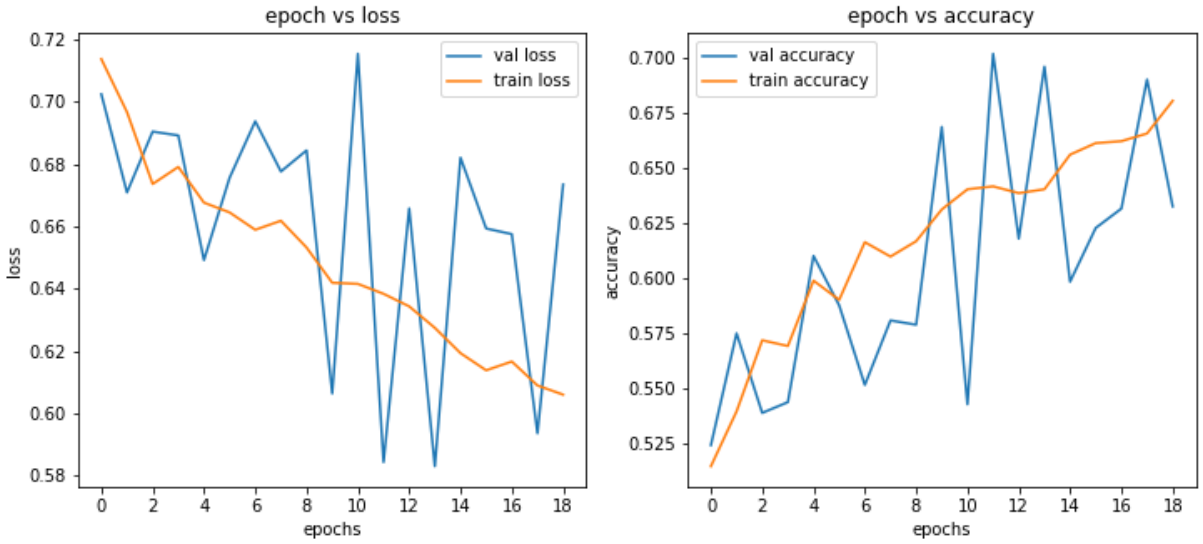
Layer (type)	Output Shape	Param #
input_3 (InputLayer)	(None, 224, 224, 3)	0
block1_conv1 (Conv2D)	(None, 224, 224, 64)	1792
block1_conv2 (Conv2D)	(None, 224, 224, 64)	36928
block1_pool (MaxPooling2D)	(None, 112, 112, 64)	0
block2_conv1 (Conv2D)	(None, 112, 112, 128)	73856
block2_conv2 (Conv2D)	(None, 112, 112, 128)	147584
block2_pool (MaxPooling2D)	(None, 56, 56, 128)	0
block3_conv1 (Conv2D)	(None, 56, 56, 256)	295168
block3_conv2 (Conv2D)	(None, 56, 56, 256)	590080
block3_conv3 (Conv2D)	(None, 56, 56, 256)	590080
block3_pool (MaxPooling2D)	(None, 28, 28, 256)	0
block4_conv1 (Conv2D)	(None, 28, 28, 512)	1180160
block4_conv2 (Conv2D)	(None, 28, 28, 512)	2359808
block4_conv3 (Conv2D)	(None, 28, 28, 512)	2359808
block4_pool (MaxPooling2D)	(None, 14, 14, 512)	0
block5_conv1 (Conv2D)	(None, 14, 14, 512)	2359808
block5_conv2 (Conv2D)	(None, 14, 14, 512)	2359808
block5_conv3 (Conv2D)	(None, 14, 14, 512)	2359808
block5_pool (MaxPooling2D)	(None, 7, 7, 512)	0
flatten_2 (Flatten)	(None, 25088)	0
dropout_5 (Dropout)	(None, 25088)	0
dense_5 (Dense)	(None, 1024)	25691136
dropout_6 (Dropout)	(None, 1024)	0
dense_6 (Dense)	(None, 512)	524800
dropout_7 (Dropout)	(None, 512)	0
dense_7 (Dense)	(None, 256)	131328
dropout_8 (Dropout)	(None, 256)	0
dense_8 (Dense)	(None, 1)	257

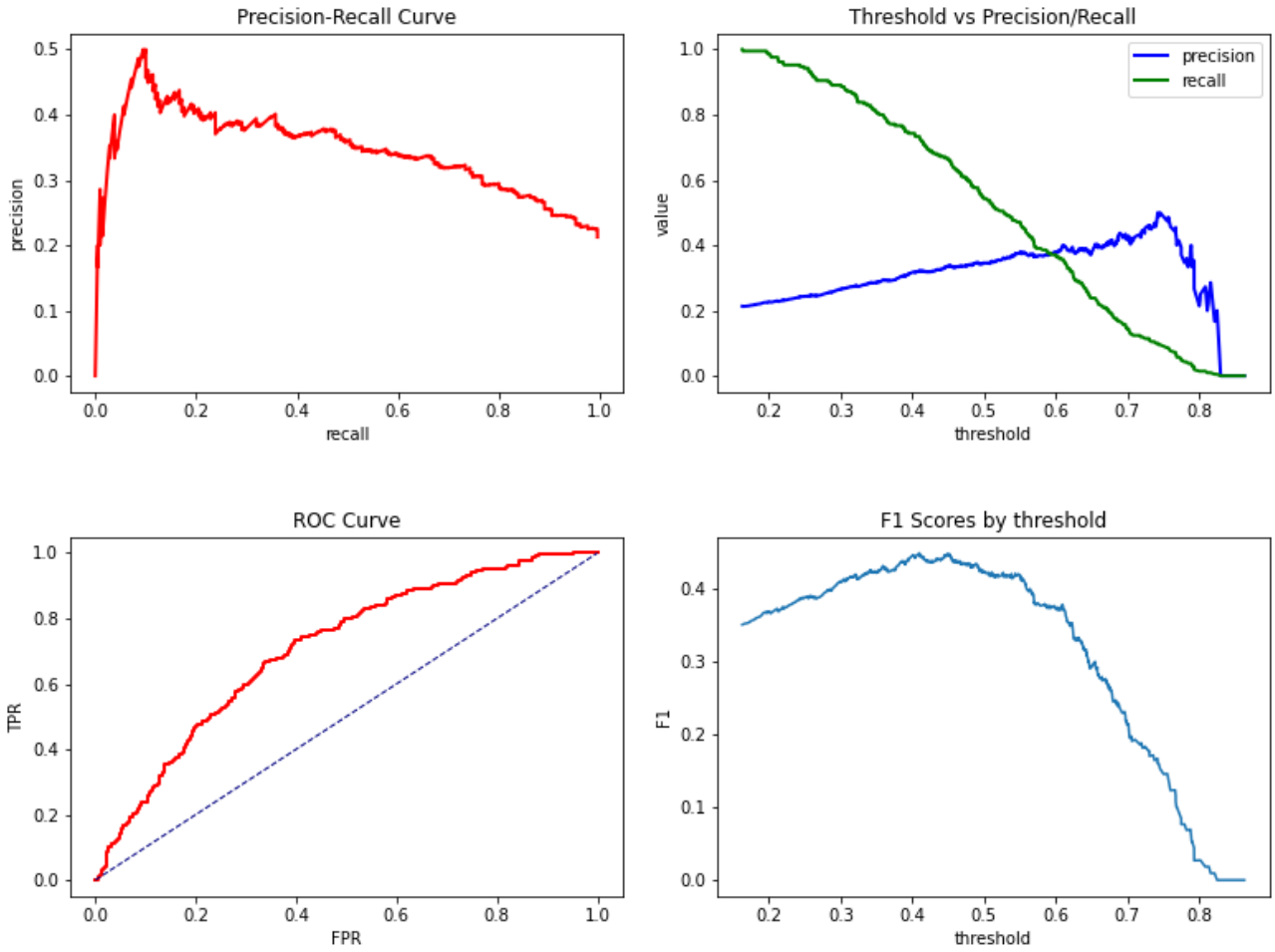
3. Algorithm Training

Parameters:

- Types of augmentation used during training:
  - horizontal flip
  - rotation range of 5
  - shear range of 0.05
  - zoom range of 0.05
- Batch size: 32
- Optimizer learning rate: 1e-5
- 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 as shown in the above diagram.
- Layers added to pre-existing architecture: flatten, dense and dropout layers as shown in the above diagram.

Algorithm training performance visualization:





Final Threshold and Explanation:

The maximum F1 score for the model is 0.448 and it is achieved with a threshold value of 0.41. Since the model does not have a high precision with any meaningful recall value, its usefulness tends to lie in its recall (and negative predictive value). Therefore, a threshold value of 0.356 (with correspoding F1 value of 0.425) for the model is selected such that it maximizes recall and NPV with a minimal loss in precision. In the table below (also in the threshold vs precision/recall curve provided above), it is shown how precision/recall changes with various threshold values.

```
Recall: 0.871, precision: 0.270, F1: 0.413, threshold: 0.310
Recall: 0.838, precision: 0.279, F1: 0.419, threshold: 0.332
Recall: 0.800, precision: 0.289, F1: 0.425, threshold: 0.356
Recall: 0.767, precision: 0.294, F1: 0.426, threshold: 0.377
Recall: 0.743, precision: 0.316, F1: 0.444, threshold: 0.399
Recall: 0.733, precision: 0.323, F1: 0.446, threshold: 0.410
Recall: 0.695, precision: 0.319, F1: 0.438, threshold: 0.422
Recall: 0.671, precision: 0.329, F1: 0.442, threshold: 0.444
Recall: 0.624, precision: 0.336, F1: 0.437, threshold: 0.466
Recall: 0.576, precision: 0.346, F1: 0.433, threshold: 0.488
Recall: 0.529, precision: 0.348, F1: 0.420, threshold: 0.510
```

The F1 Scores reported in the paper CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning are provided in the following table:

	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)

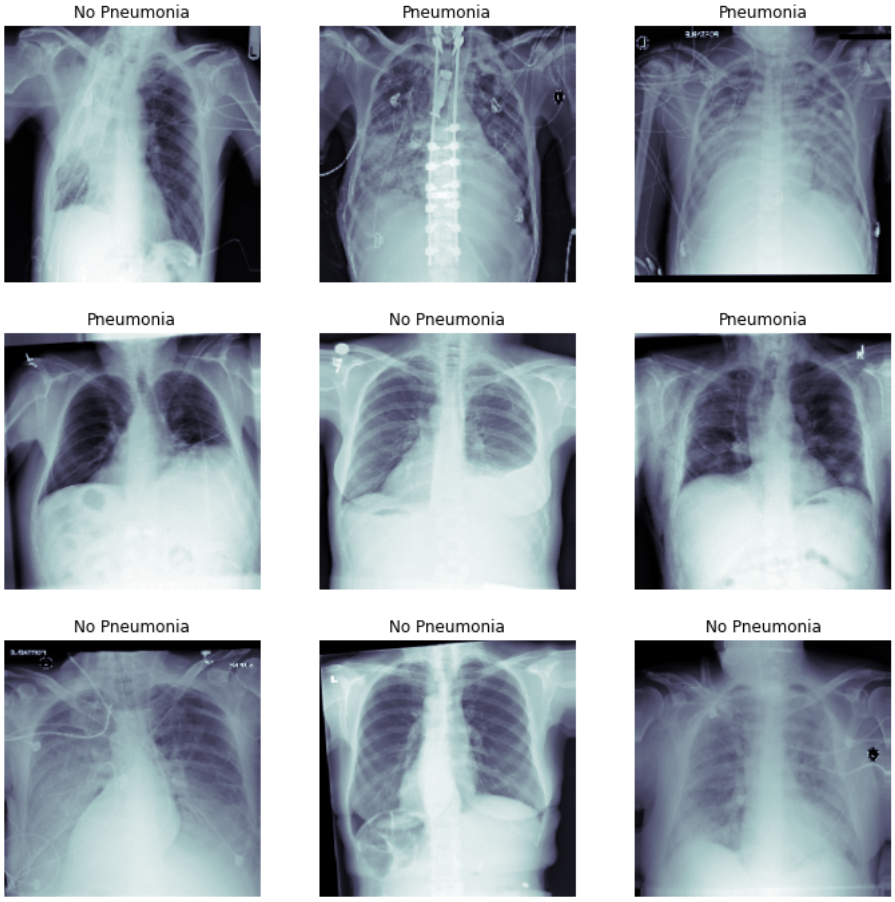
It is evident from the above comparison that the model's F1 score is better than most radiologists and comparable to that of CheXNet.

4. Databases

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

Description of Training Dataset:

Training dataset consisted of 2290 chest X-ray images which are equally split between Pneumonia and non-Pneumonia cases. Some training samples are shown below:



Description of Validation Dataset:

Validation dataset consisted of 1430 chest X-ray images, with 20/80 split between Pneumonia and non-Pneumonia cases, which more reflects the real world scenario.

5. Ground Truth

The data is taken from a larger ChestX-ray dataset which comprises 112,120 frontal-view X-ray images of 30,805 unique patients. The image labels are created using natural language processing (NLP), mining the associated radiological reports. The labels include fourteen common thoracic pathologies: Atelectasis, Consolidation, Infiltration, Pneumothorax, Edema, Emphysema, Fibrosis, Effusion, Pneumonia, Pleural_thickening, Cardiomegaly, Nodule, Mass and Hernia. Since the labels were NLP-extracted, 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. 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, defined for a specific disease, is the method that detects the disease in question with the highest sensitivity and accuracy. The gold standard is used to assess the performance of a machine learning/AI algorithm by providing the ground truth labels. The gold standard for pneumonia would be to perform one or a combination of the following tests: sputum culture, blood tests, pleural biopsy, bronchoscopy, CT scan, etc. Since these tests are expensive and time consuming, we can rely on a silver standard. A silver standard makes a diagnosis by having a voting system that aggregates the diagnoses from multiple experts (e.g. radiologists). Some diseases may be very hard to detect, so this strategy can minimize the effects of human errors. Using the silver standard, the ground truth for pneumonia can be obtained by a weighted judgment on a sample from several radiologists (similar to the method used in the paper CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning). The final label would be weighted by the experience of the radiologists.

Algorithm Performance Standard:

In terms of Clinical performance, the algorithm's performance can be measured by calculating F1 score against 'silver standard' ground truth as described above. The algorithm's F1 score should exceed 0.387 which is an average F1 score taken over four experienced radiologists, as given in the above mentioned paper, where a similar method is used to compare CheXNet's F1 score with the average F1 score over four radiologists.