

# FDA Submission

**Your Name:**

Jingxian Lin

**Name of your Device:**

Pneumonia Classifier

## Algorithm Description

### 1. General Information

**Intended Use Statement:**

This device is intended to help radiologists examine the presence of pneumonia in X-ray image based on Convolutional Neural Network

**Indications for Use:**

It is targeted to aid radiologist to classify X-ray image for patients aged between 1 and 100, and there could be certain comorbidities with pneumonia like Atelectasis, Consolidation, Infiltration, Pneumothorax, Edema, Emphysema, Fibrosis, Effusion, Pleural thickening, Cardiomegaly, Nodule, Mass, Hernia

**Device Limitations:**

CPU with at least 8G RAM

**Clinical Impact of Performance:**

The algorithm is trained to get a high F1 score and recall value, so the false positive rate maybe high, which should be considered when applied: False positives classify a patient with no pneumonia as positive to get unnecessary treatment; false negatives classify a patient with pneumonia as negative to miss required treatment, which is worse

### 2. Algorithm Design and Function

**Algorithm Flowchart:**

Exploratory Data Analysis -> Data Preprocessing -> Feature Engineering -> Model Training -> Prediction and Evaluation

**DICOM Checking Steps:**

To make sure Modulaity == 'DX' and PatientPosition in ['AP', 'PA'] and BodyPartExamined == 'CHEST'

**Preprocessing Steps:**

Include resizing and normalization

**CNN Architecture:**

Model: "sequential"		
Layer (type)	Output Shape	Param #
=====		
model (Model)	(None, 7, 7, 512)	14714688
flatten (Flatten)	(None, 25088)	0
dropout (Dropout)	(None, 25088)	0
dense (Dense)	(None, 1024)	25691136
dropout_1 (Dropout)	(None, 1024)	0
dense_1 (Dense)	(None, 512)	524800
dropout_2 (Dropout)	(None, 512)	0
dense_2 (Dense)	(None, 256)	131328
dense_3 (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

Implement a package like Keras' ImageDataGenerator with some of the built-in augmentations: A horizontal flip Random height shift within 10% range of X-ray image height Random width shift within 10% range of X-ray image width Random rotation within 20-degree range Random shear transformation within 0.1 range Random zooming within one tenth range

- Batch size

Set batch\_size as 32

- Optimizer learning rate

Use the learning rate  $1e-4$  for Adam optimizer

- Layers of pre-existing architecture that were frozen

First 17 layers of pre-existing architecture are frozen

- Layers of pre-existing architecture that were fine-tuned

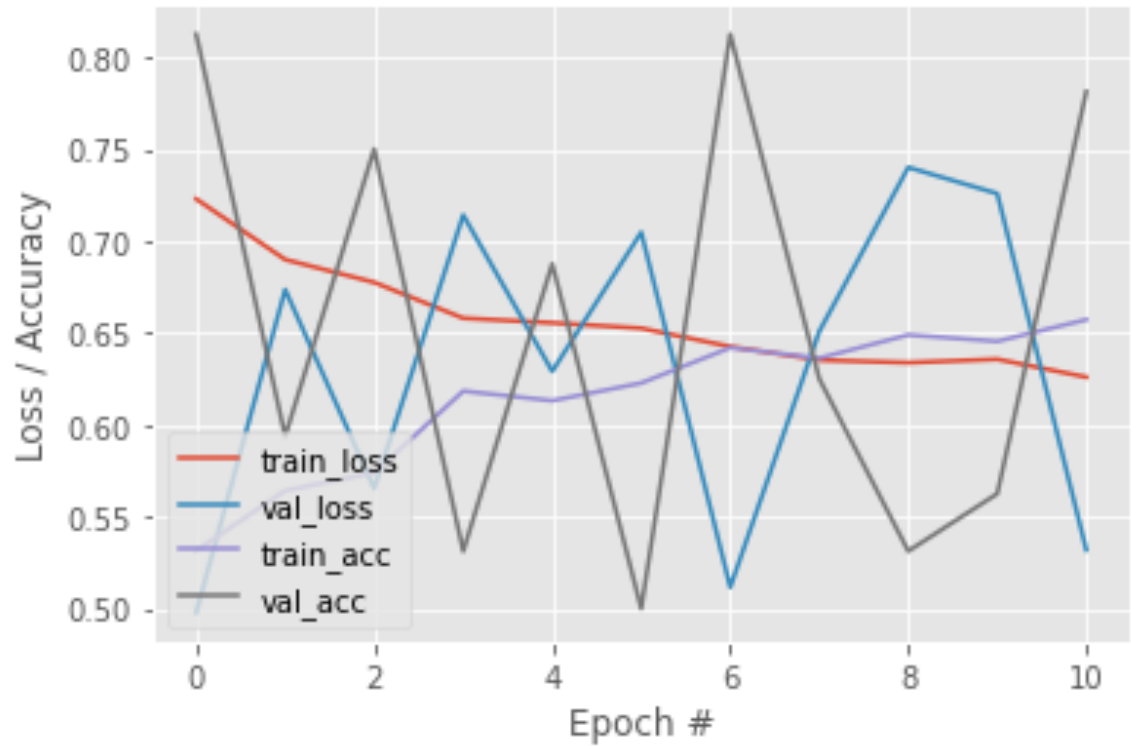
The last layer of pre-existing architecture is fine-tuned

- Layers added to pre-existing architecture

Flatten, Dropout, and Dense layers are added to pre-existing architecture, as shown in CNN Architecture above

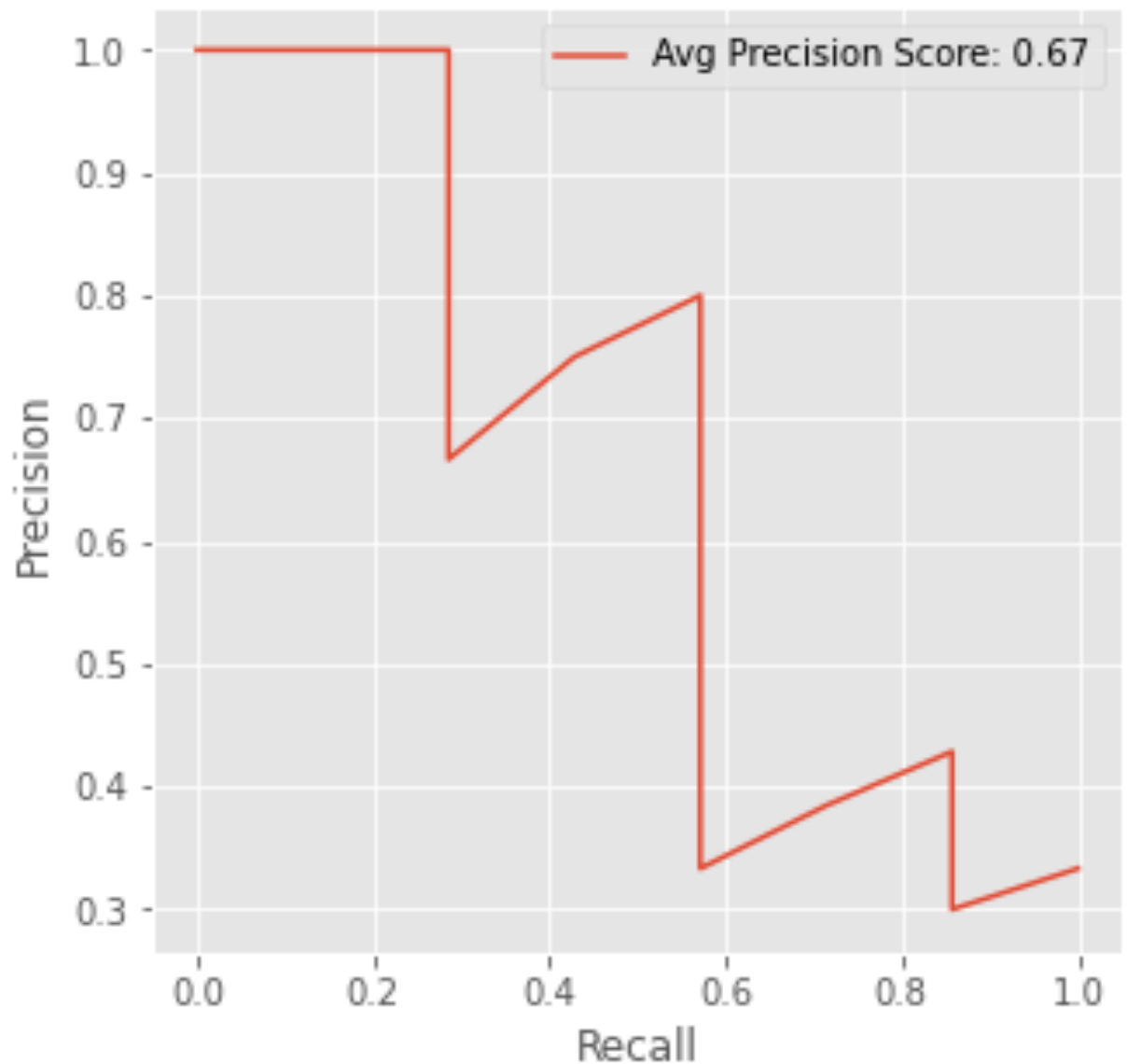
#### Algorithm Training Performance Visualization:

Loss and Accuracy on Training and Validation Datasets



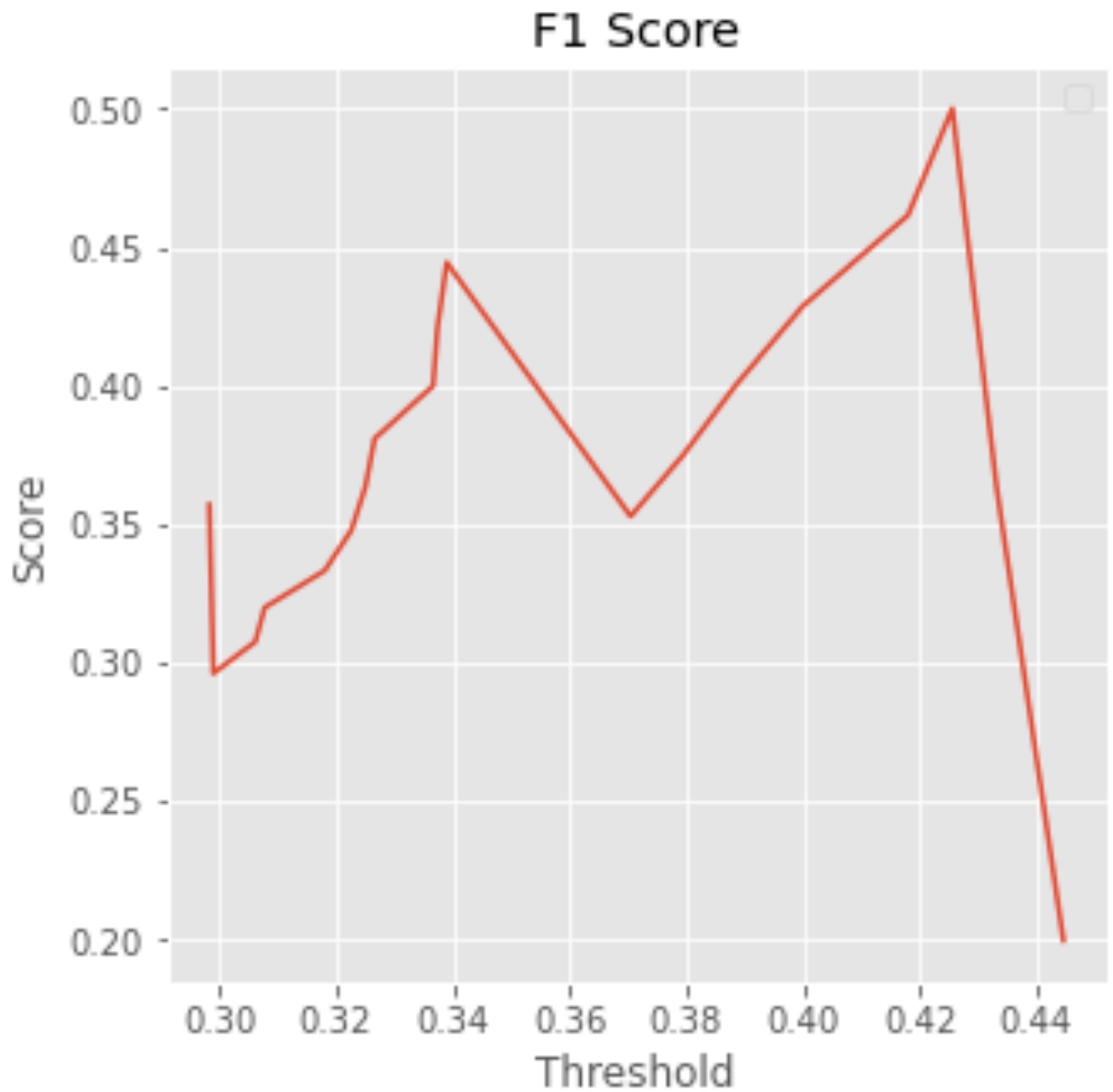
P-R Curve:

## Precision Recall Curve



Precision is  $TP / (TP + FP)$ , the percentage of true positives over all positive results, the higher precision, the more confident about positive results; recall is  $TN / (TN + FP)$ , the percentage of true negatives over all patients with no disease, the higher recall, the more confident about negative results. Consider the trade-off between precision and recall, F1 score =  $2 \times \text{precision} \times \text{recall} / (\text{precision} + \text{recall})$  is selected as criterion

**Final Threshold and Explanation:**



Final threshold is set as 0.4255593 to get a higher F1 score 0.5

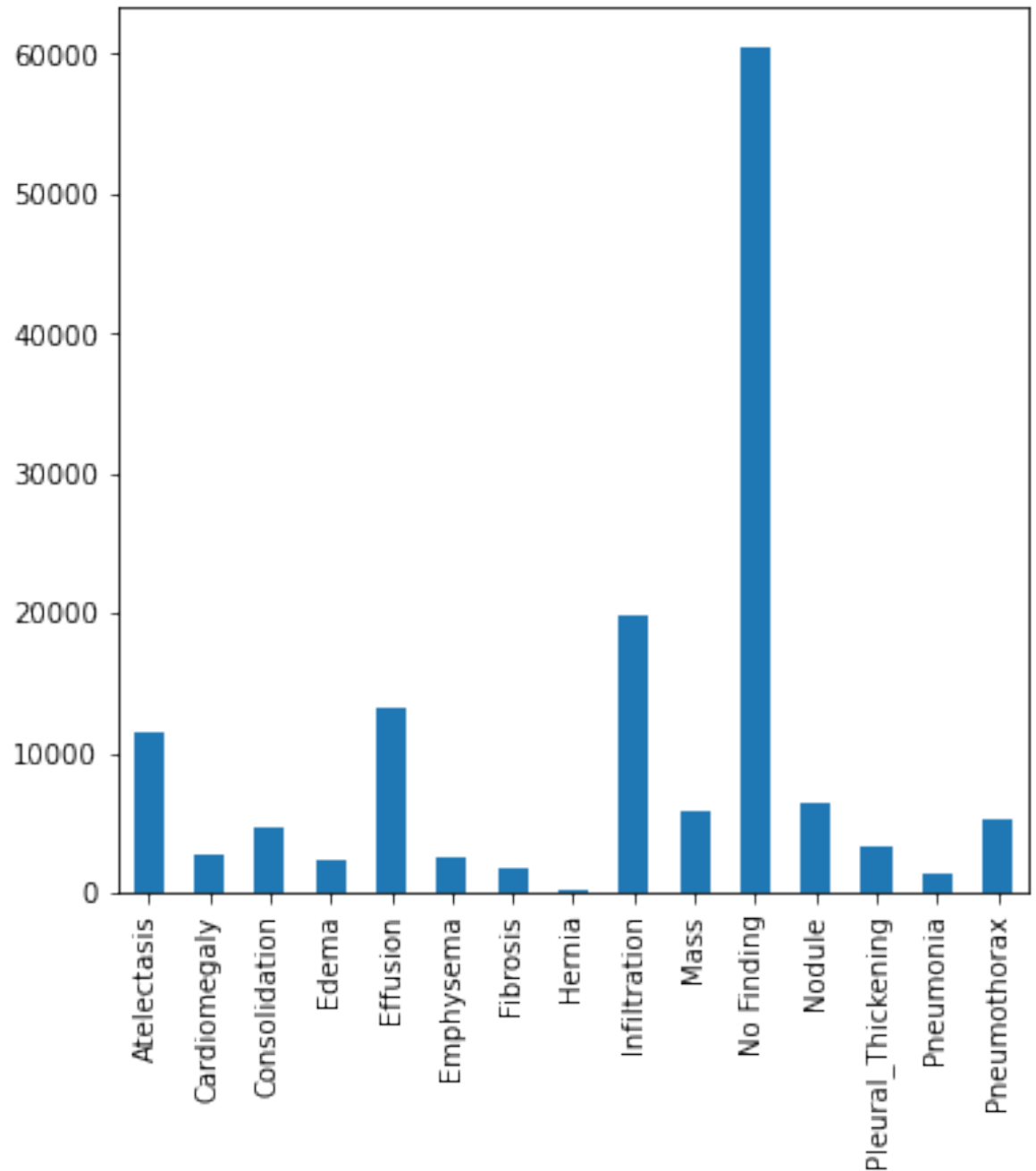
## 4. Databases

(For the below, include visualizations as they are useful and relevant)

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



From the disease distribution, Infiltration, Effusion, and Atelectasis are top 3 except No Finding

### Description of Training Dataset:

The percentage of the presence of pneumonia in the original training dataset is 0.013, and this imbalance is adjusted according to the lesson in the course to make the new percentage become 0.5

### **Description of Validation Dataset:**

The percentage of the presence of pneumonia in the original validation dataset is also 0.013, and this unbalance is adjusted following the lesson in the course to become 0.2

## **5. Ground Truth**

The disease labels were created using Natural Language Processing (NLP) to mine the associated radiological reports. 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:**

For FDA Validation Dataset, the sampling should be based on equal number of male and female patients, the age range between 1 and 100, and including both the presence and absence of any of 14 common thoracic pathologies

### **Ground Truth Acquisition Methodology:**

To acquire Ground Truth, make sure Modality == 'DX' and PatientPosition in ['AP', 'PA'] and BodyPartExamined == 'CHEST', and prefer biopsy data

### **Algorithm Performance Standard:**

Metrics chosen to monitor for Performance could be F1 score, Precision, or Recall, and F1 score is a better selection. According to the paper "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning" (<https://arxiv.org/pdf/1711.05225.pdf>), CheXNet's F1 score is 0.435. After applying the current algorithm to FDA Validation Dataset, this model would output a F1 score, then the statistical significance of the new score compared to the CheXNet's score should be checked. To support that this algorithm is a better one, the new score is not necessary to be equal or higher, it could be lower but inside a tolerance window, in this case, the calculated p-value should be compared to the confidence level or 95% confident interval should be examined, all conclusion should be based on statistical inference!