FDA Submission

Name: Tingyi Dong
Name of your Device: Pneumonia Binary Classifier from Chest X-ray

Algorithm Description

1. General Information

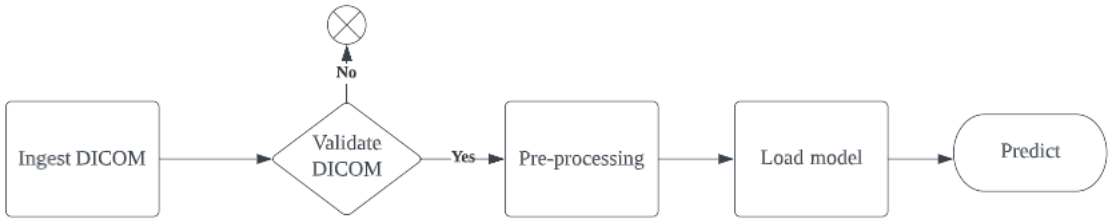
Intended Use Statement
For assisting the radiological diagnosis of the presence or absence of pneumonia from chest X-rays.

Indications for Use
To classify and detect the presence or absence of pneumonia, and to reduce the time of radiological diagnosis in digital chest X-ray for both male and female between the age of 5 to 120 years old, who might exhibit other diseases comorbid with pneumonia.

Device Limitations
This device requires digitalized chest X-ray scans and high computing power computer (GPU) or cloud-based service to run the algorithm.

Clinical Impact of Performance
False negative may lead to mistreatment when mis-diagnosing a patience with pneumonia as healthy. False positive may lead to unnecessary checks of radiologist on healthy patience predicted as having pneumonia. As a result, in clinical setting, false negatives can be worse false positive. This algorithm is can help radiologist prioritize their time and attention to chest x-rays that potentially contain pneumonia. Yet, those cases predicted negative should still be reviewed by radiologists.

2. Algorithm Design and Function



DICOM Checking Steps
The algorithm performs the following checks on the DICOM image meta-data using pydicom library:
- Checks “Patient Age” is between 5 and 120 (inclusive)
- Checks “Examined Body Part” is 'CHEST'
- Checks “Patient Position” is either 'PA' (Posterior/Anterior) or 'AP' (Anterior/Posterior)
- Checks “Modality” is 'DX' (Digital Radiography)

DICOM Xray that does not meet all these criteria will not be assessed.

Preprocessing Steps
The algorithm performs the following preprocessing steps on the image pixel data:
- Converts RGB to Grayscale (if applicable)
- Normalizes the intensity to be between 0 and 1 (from original range of 0 to 255)
- Stacks and resizes to fit the input shape of the VGG16 model (1,244,244,3)

CNN Architecture

```
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
model_1 (Model)              (None, 7, 7, 512)         14714688
_________________________________________________________________
flatten_1 (Flatten)          (None, 25088)             0
_________________________________________________________________
dense_1 (Dense)              (None, 512)               12845568
_________________________________________________________________
dropout_1 (Dropout)          (None, 512)               0
_________________________________________________________________
dense_2 (Dense)              (None, 128)               65664
_________________________________________________________________
dropout_2 (Dropout)          (None, 128)               0
_________________________________________________________________
dense_3 (Dense)              (None, 1)                 129
=================================================================
Total params: 27,626,049
Trainable params: 15,271,169
Non-trainable params: 12,354,880
_________________________________________________________________
```

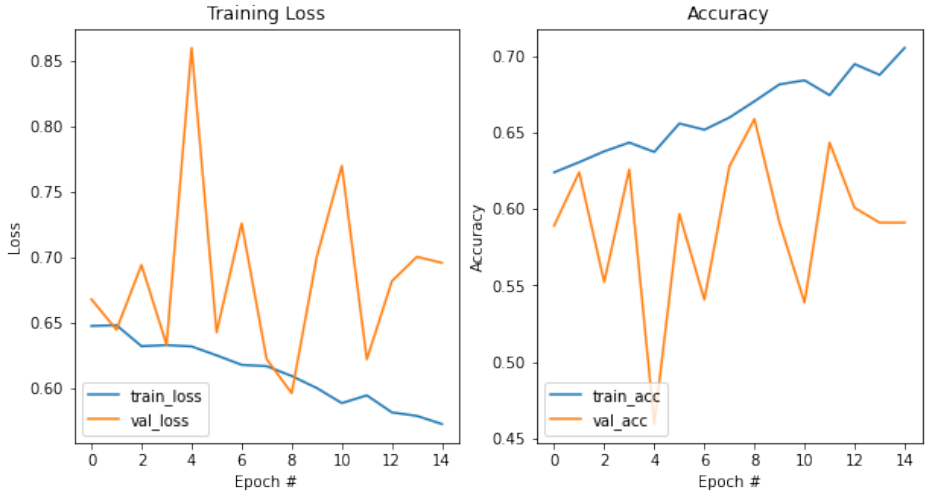
The algorithm uses pre-trained VGG16 Neural Network (except the last block of Convolution + Pooling layers that was re-trained), with additional two blocks of 'Fully Connected + Dropout' layers. The network output is a single probability value for binary classification.

3. Algorithm Training

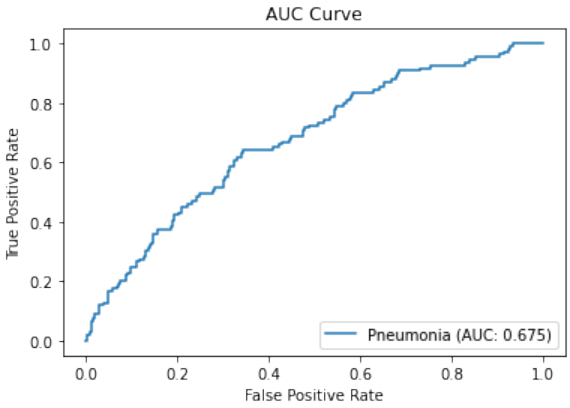
Parameters:

- Types of image augmentation used during training:
  - horizontal_flip
  - height_shift_range: 0.1
  - width_shift_range: 0.1
  - rotation_range: 0 to 20 degrees
  - shear: 0.1
  - zoom: 0.1
- Batch size: 32
- Optimizer: Adam with learning rate at 1e-4 and drop 50% after every 10 epochs
- Layers of pre-existing architecture that were frozen
  - All except last convolution + pooling block
- Layers of pre-existing architecture that were fine-tuned
  - The last 2 layers of VGG16 network: block5_conv3 + block5_pool
- Layers added to pre-existing architecture:
  - flatten_1 (Flatten)
  - dense_2 (Dense, 512)
  - dropout_3 (Dropout, 0.2)
  - dense_3 (Dense, 128)
  - dropout_4 (Dropout, 0.5)

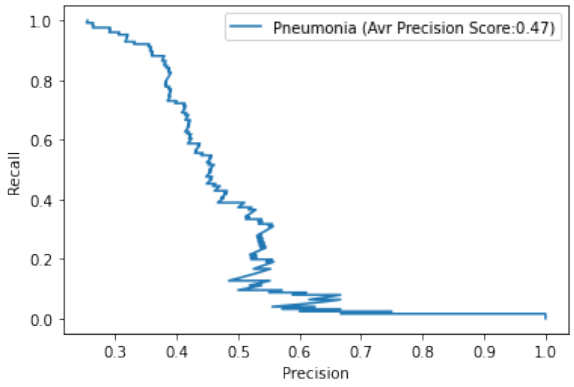
Algorithm training performance visualization



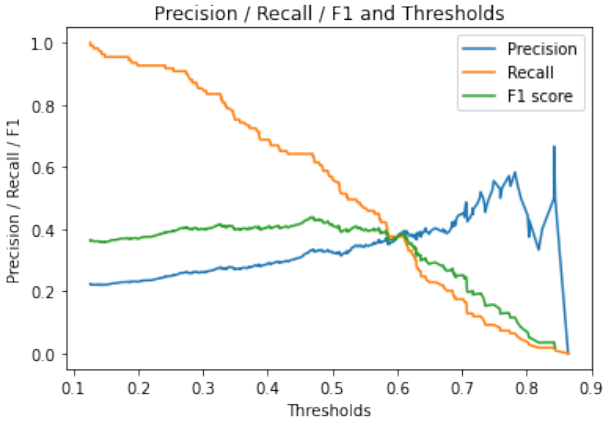
AUC (Area Under The Curve) ROC (Receiver Operating Characteristics) Curve



Precision-Recall curve



Final Threshold and Explanation:
As the dataset is imbalanced for pneumonia disease, I chose F1-score to achieve balance between precision and recall. A threshold at 0.411 was chosen to reach maximum F1-score at 0.531.

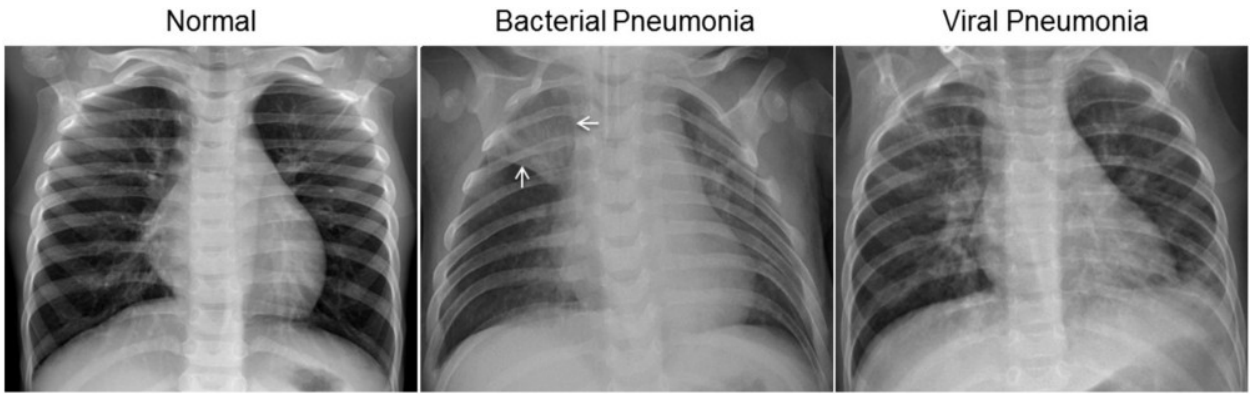


4. Databases

The dataset used consists of 112,120 frontal-view chest X-ray PNG images in 1024x1024 resolution, with disease labels from 30,805 unique patients, as well meta data for all images. This dataset was curated by the NIH specifically to address the problem of a lack of large x-ray datasets with ground truth labels to be used in the creation of disease detection algorithms.

The disease labels were created using Natural Language Processing (NLP) to mine the associated radiological reports. The labels also include 14 common thoracic pathologies, including Pneumonia.

Example of chest X-ray for normal versus pneumonia is shown below.



For this model, a 80-20 % split was used to create training and validation datasets. For training dataset, a 50-50% balance was maintained between positive and negative cases; while for validation dataset, class distribution reflected clinical settings with 24.8% (~25%) positive cases [source].

5. Ground Truth

The data is taken from a larger Xray dataset, with disease labels created using Natural Language Processing (NLP) mining the associated radiological reports. The labels include 14 common thoracic pathologies (Pneumonia being one of them):

- 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%.
The original radiology reports are not publicly available but more details on the labeling process can be found here.

6. FDA Validation Plan

Patient Population Description for FDA Validation Dataset:
The following population subset is to be used for the FDA Validation Dataset:

- Both men and women
- Between age of 5 and 120 years of
- Patient may exhibit the following diseases comorbid with Pneumonia: Atelectasis, Cardiomegaly, Consolidation, Edema, Effusion, Emphysema, Fibrosis, Hernia, Infiltration, Mass, Nodule, Pleural_Thickening, Pneumonia, Pneumothorax

The XRay DICOM file should have the following properties:

- Patient Position: AP or PA
- Image Type: DX
- Body Part Examined: CHEST

Ground Truth Acquisition Methodology:
The ground truth for the FDA validation dataset can be obtained as an average of three practicing radiologists (as a widely used 'silver standard') since the purpose of this device is to assist the radiologist.

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 **0.435** as indicated below to outperform the current state-of-the-art method, CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rayswith Deep Learning, where a similar method is used to compare device's F1 score to average F1 score over three radiologists.

	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)