FDA Submission

Your Name: Qiuzhao Dong

Name of your device: NN Model for Pneumonia Detection From Chest X-Rays

Algorithm Description

1. General Information

Intended Use Statement

This algorithm is intended for use on assisting the radiologists in the screening of patients for Pneumonia using chest X-Rays with a review of a patient's medical history for diagnostic validation.

Indications for Use:

This algorithm is indicated for use for pneumonia screening for male and female patients with age from 1 to 95 year(s) old in non-emergency situations. X-ray image must be taken with PA or AP position, and the modality should be "DX" only.

Device Limitations:

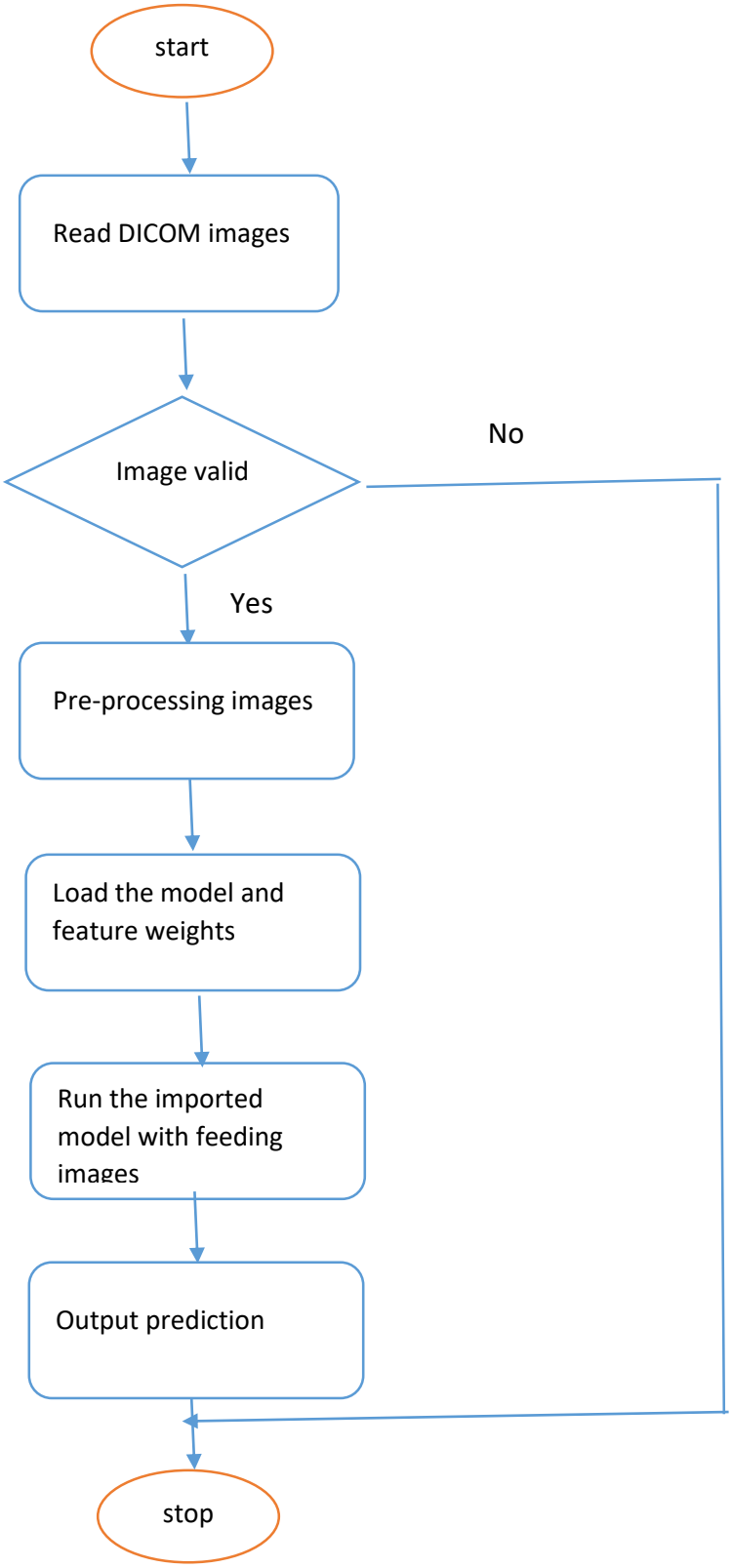
- The algorithm has to be run in a computer with GPU;
- The presence of infiltrations in a chest x-ray is a limitation of this algorithm, it causes the possible wrong classification for Pneumonia;
- The presences of Edema, Atelectasis, Effusion, Infiltration and their mixed in a chest x-ray may be a limitation of this algorithm.

Clinical Impact of Performance:

- Algorithm has a high recall but relative low precision rate;
- A high recall means a low false negative which can adversely impact a patient. So, when a patient receives a negative report, it is almost true;
- A low precision means a high false positive, in this situation, a radiologist needs to look into x-ray in detail to further confirm it and the patient may be required to do sputum culture;

- In this scenario, the x-ray marked positive could have high priority for radiologists further checking.

2. Algorithm Design and Function



DICOM Checking Steps:

The DICOMS' headers have been checked for:

1. BodyPartExamnied=='CHEST';
2. Modality == 'DX';
3. PatientPosition in 'PA' or 'AP' position.

Preprocessing Steps:

step 1: Images will be resized to 224x224;

step 2: images will be converted to RGB style;

step 3: changing image format to (1,224,224,3) to satisfy pretrained VGG16 model input;

step 4: importing VGG16 preprocessing function to preprocess the images.

CNN Architecture:

A Sequential Model was built by fine-tuning VGG16 Model with the ImageNet weights. This model takes the pretrained VGG16 model layers up to 5 blocks, the first 18 layers were frozen and not trained. Then with the output from the VGG16 model, we added several fully connected layer and a bi-classification layer. The full model architecture:

input_1 (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_4 (Flatten) (None, 25088) 0
dense_15 (Dense) (None, 1024) 25691136
dropout_12 (Dropout) (None, 1024) 0
dense_16 (Dense) (None, 512) 524800
dropout_13 (Dropout) (None, 512) 0
dense_17 (Dense) (None, 256) 131328
dense_18 (Dense) (None, 1) 257

3. Algorithm Training

Parameters:

Keras.preprocessing.image ImageDataGenerator was used to augment the images with the following parameters:

- rescale= 1./255.0,
- horizontal_flip= True,
- vertical_flip=False,
- height_shift_range=0.1,
- width_shift_range=0.1,
- rotation_range=20,
- shear_range=s0.1,
- zoom_range=0.1

Batch size: 32

Optimizer learning rate: 1e-4

Layers of pre-existing architecture that were frozen: VGG16 first 18 layers

Layers of pre-existing architecture that were fine-tuned: None

Layers added to pre-existing architecture:

(Dense(1024, activation = 'relu'))
(Dropout(0.5))
(Dense(512, activation = 'relu'))

(Dropout(0.5))
(Dense(256, activation = 'relu'))
(Dense(1, activation = 'sigmoid'))

The training performance visualization is as following:

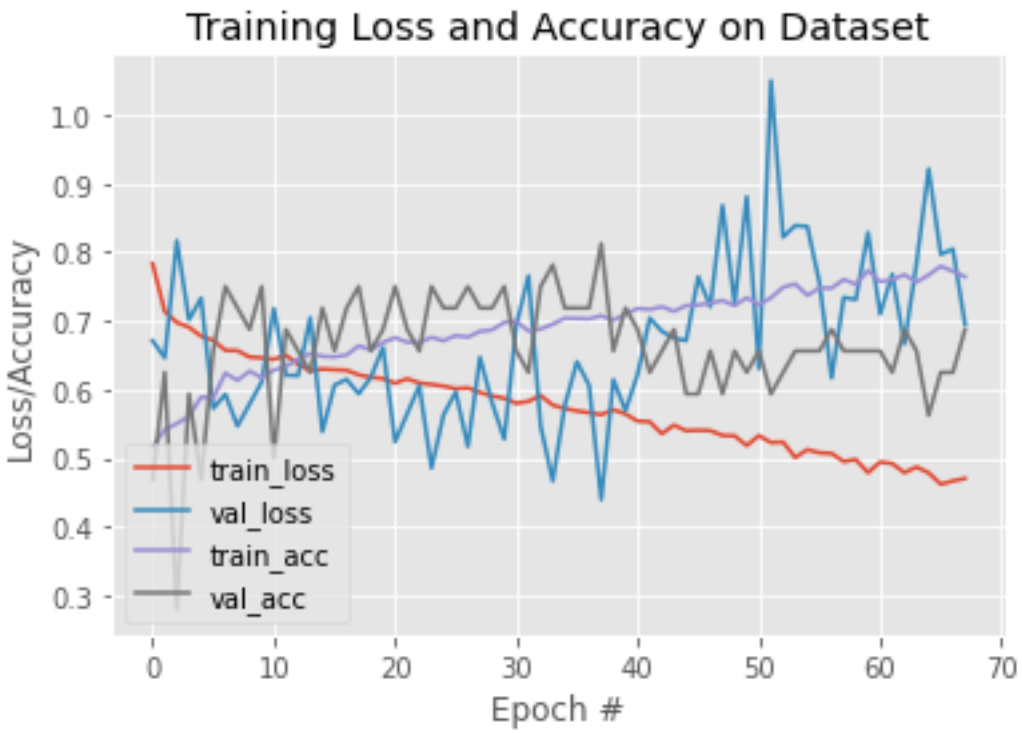


Figure 1 Model training history

From the plot, the training loss continuously decreases with more epochs, the training accuracy continuously increases with more epochs. The validation loss vibrating decreases to a minimum point, then increases vibrating. We chose the optimal model with lowest validation loss.

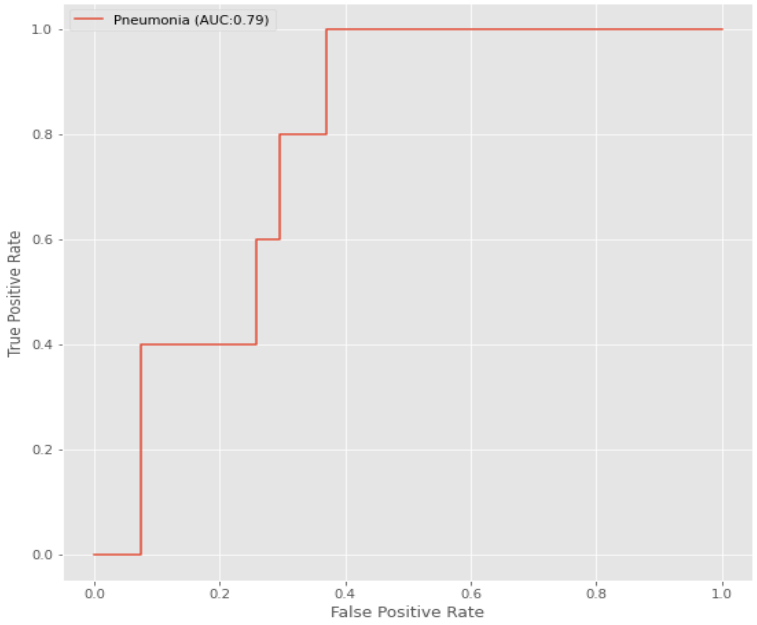


Figure 2 ROC curve

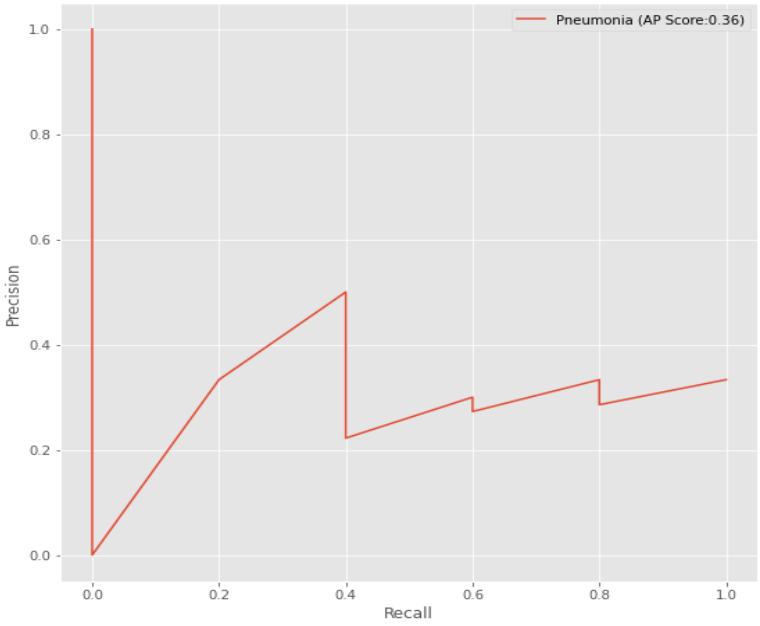


Figure 3 precision-recall curve

Final Threshold and Explanation:

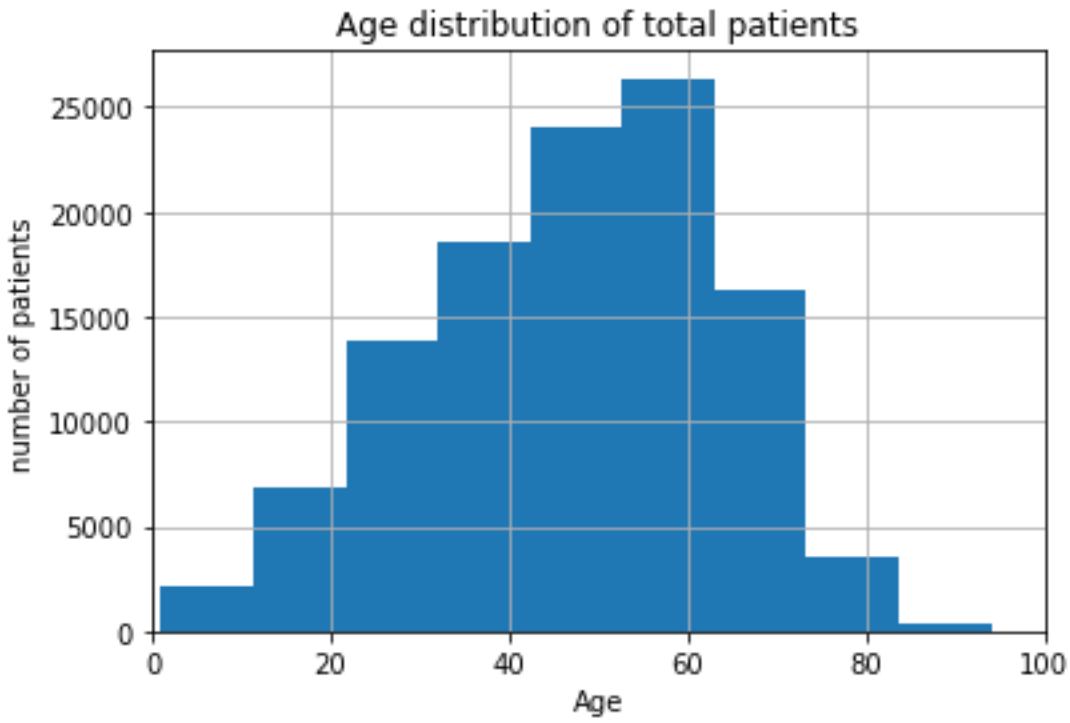
Precision	Recall	F1 score	Threshold
0.33	1.0	0.500	0.27
0.286	0.8	0.421	0.271
0.308	0.8	0.444	0.299
0.333	0.8	0.471	0.38
0.273	0.6	0.375	0.397
0.3	0.6	0.400	0.400
0.222	0.4	0.286	0.451
0.25	0.4	0.308	0.452
0.286	0.4	0.333	0.455
0.333	0.4	0.364	0.498
0.4	0.4	0.400	0.510
0.5	0.4	0.444	0.548
0.333	0.2	0.250	0.565

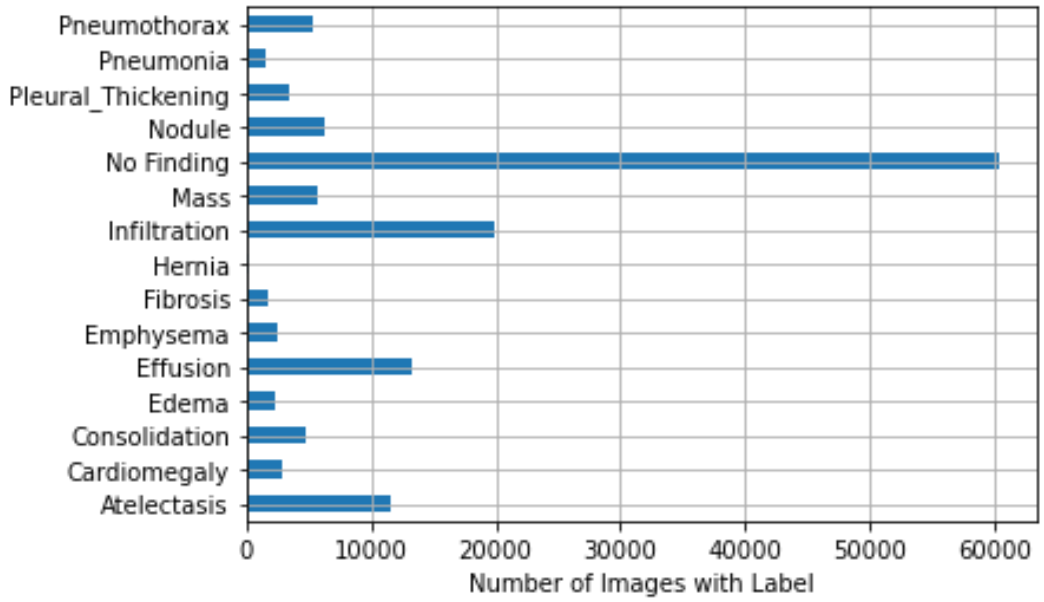
The table shows the relations of precision, recall, F1 score and threshold. F1 score is the harmonic mean between precision and recall. Typically, F1 score closer to 1 is good and desirable. Low f1 score implies that one of precision/recall is very low. From my model, Recall has been given more importance since in the clinical setting, we don't want to miss a positive case. Based on this criteria, I chose threshold 0.27 with Recall close to 1 and maximum F1 score 0.5.

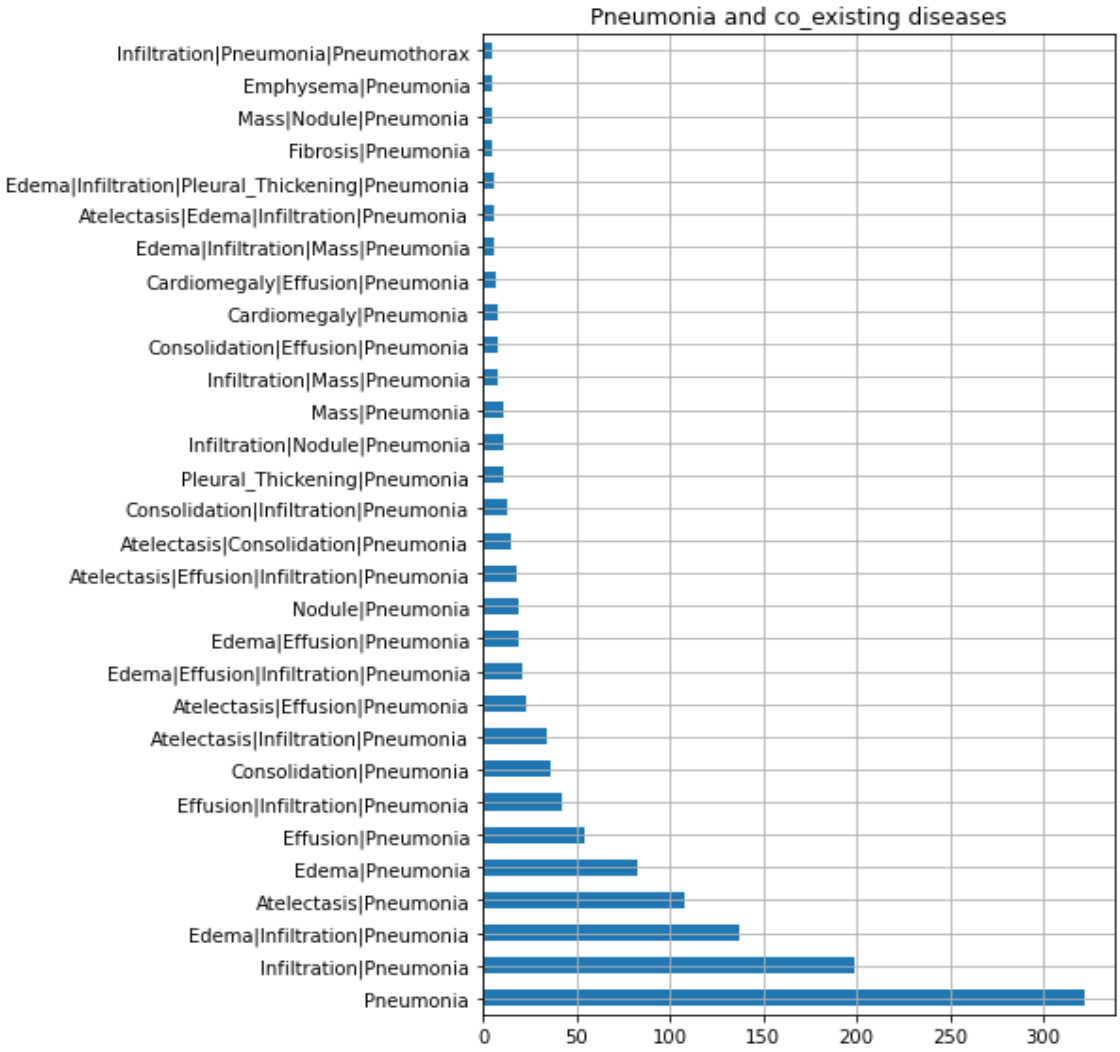
4. Databases

The original dataset used for designing training dataset and validation dataset has the following parameters:

- total x-rays data: 112120;
- Age distribution: 2 to 412 years (must have some errors)
- Genders: Male (56.5%) and Female (43.5%)
- Number of patients with no disease: 16403
- Number of patients with one or more diseases: 14402
- Number of patients with pneumonia: 1431(1.27%)


Age distribution of total patients
number of patients
25000
20000
15000
10000
5000
0
0
20
40
60
80
100
Age


Pneumothorax
Pneumonia
Pleural_Thickening
Nodule
No Finding
Mass
Infiltration
Hernia
Fibrosis
Emphysema
Effusion
Edema
Consolidation
Cardiomegaly
Atelectasis
0
10000
20000
30000
40000
50000
60000
Number of Images with Label



Description of Training Dataset

The training dataset has been carefully chosen to balance the positive patients and negative patients from a highly imbalanced dataset. It is important to keep the positive cases the same as the negative cases. Number of x-rays images with and without Pneumonia are both 1145.

The age and gender distribution in training dataset match the age and gender distribution in the original dataset.

Description of Validation Dataset

The validation dataset has been chosen to follow the real world scenario. The imbalanced ratio is 1:4 for the pneumonia cases to the non-pneumonia cases. The

number of cases with pneumonia is 286 counts, and the number of cases without pneumonia is 1144 counts. The validation dataset has no overlap with the training dataset.

5. Ground Truth

The original dataset used in this project was curated by the NIH. There are total 112,120 X-Ray images with disease labels from 30,805 unique patients. The disease labels for each image were created using Natural Language Processing (NLP) to process associated radiological reports. The limitation is that the accuracy of the NLP labels is estimated to be >90%. It maybe obtains higher ground truth if radiologists can review these x-ray images.

6. FDA Validation Plan

Patient Population Description for FDA Validation Dataset:

- Chest x-rays of male and female patients with age up to 95 years old;
- View position of x-ray image should be PA or AP;
- Image modality is DX;
- Presence of other diseases such as Infiltration, Edema, Effusion, Atelectasis, Cardiomegaly, Consolidation, Emphysema, Fibrosis, Hernia, Mass, Nodule, Pleural Thickening, and Pneumothorax is acceptable;
- The dataset should include 20% Pneumonia cases.

Ground Truth Acquisition Methodology:

The x-ray diagnosis of a radiologist could be seen as the ground truth, although the Gold truth should be via a biopsy procedure which are expensive and time consuming.

Algorithm Performance Standard:

Based on the provided paper – “CheXnet –Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning “--by P. Rajpurkarpar, et al., a deep learning algorithm is capable of achieving the higher performance, comparable to that of a radiologist.

Rajpurkarpar's CheXNet algorithm achieve an F1 score of 0.435, while Radiologists averaged an F1 score of 0.387.

This NN model with a F1 score of 0.5, recall ~1, AUC=0.79 is designed to assist radiologists for pneumonia screening.