FDA Submission

Name: Chinedu Chukwuelue

Name of your Device: Pneumonia Detection Assistant (Codewired-PDA)

Algorithm Description

1. General Information

Intended Use Statement:

For assisting a radiologist in detection of pneumonia in x-ray images

Indications for Use:

Screening of x-ray images

Patient population:

- Both men and women
- Age: 2 to 90

X-Ray image properties:

- Body part: Chest
- Position: AP (Anterior/Posterior) or PA (Posterior/Anterior)
- Modality: DX (Digital Radiography)

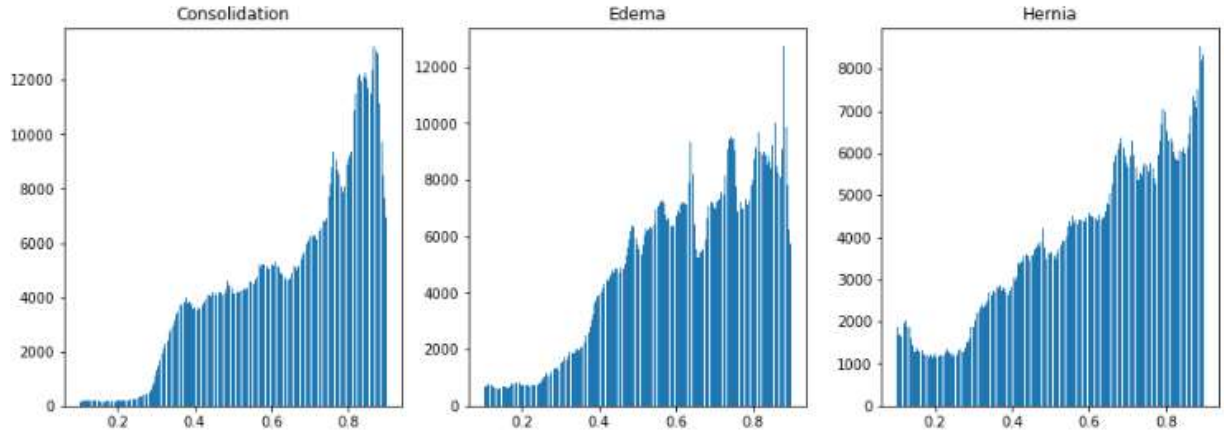
Device Limitations:

The model is recommended for use without the following comorbid thoracic pathologies:

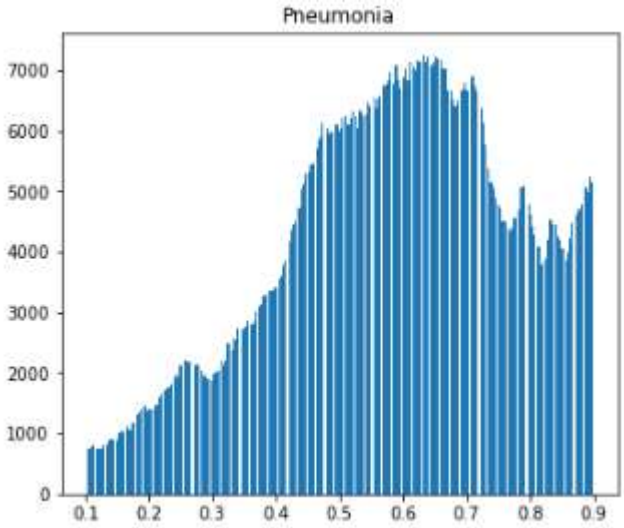
- Consolidation
- Edema
- Hernia

This can be explained with the fact that X-Ray images containing these three conditions mentioned above have similar but significantly different distribution than Pneumonia X-Rays as illustrated below

Not Recommended:



Pneumonia:



Clinical Impact of Performance:

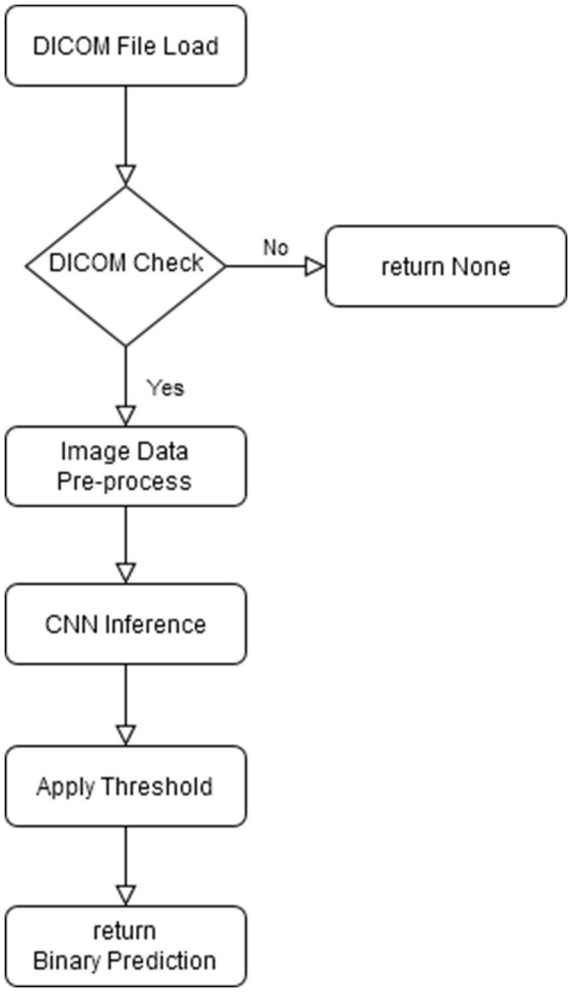
In terms of predictive value:

- If the model predicts negative, it is correct with 90.7% probability
- If the model predicts positive, it is correct with 26.3% probability

Therefore, the algorithm is recommended for assisting a radiologist to screen images that most probably do not contain Pneumonia, this will help prioritize his/her time and attention to those that potentially do. This should lead to those that require medical attention getting it much faster.

It is also worth mentioning that when algorithm predicts negative is can still be wrong with 9.3% probability. So, those cases predicted negative should still be reviewed by the radiologist.

2. Algorithm Design and Function



DICOM Checking Steps:

The algorithm performs the following checks on the DICOM image:

1. Check Patient Age is between 2 and 90 (inclusive)
2. Check Examined Body Part is 'CHEST'
3. Check Patient Position is either 'PA' (Posterior/Anterior) or 'AP' (Anterior/Posterior)
4. Check Modality is 'DX' (Digital Radiography)

Preprocessing Steps:

The algorithm performs the following preprocessing steps on x-ray images:

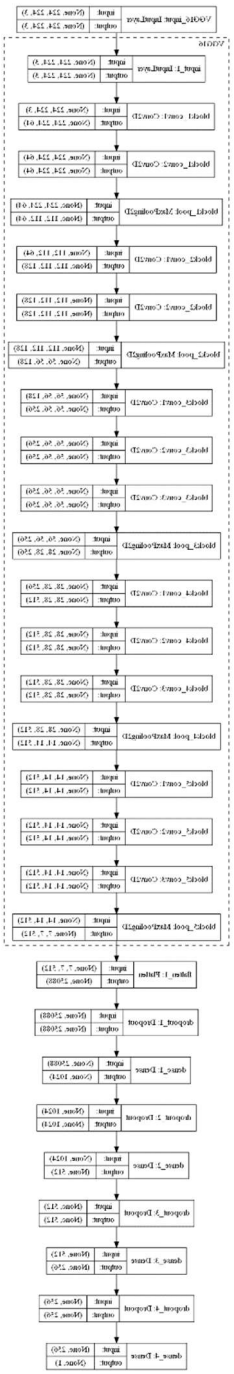
1. Converts RGB to Grayscale (if needed)
2. Re-sizes the image to 244 x 244 (as required by the CNN)
3. Normalizes the intensity to be between 0 and 1 (from original range of 0 to 255)

CNN Architecture:

The algorithm uses 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 graph: (Image also attached in Workspace for zooming)



3. Algorithm Training

Parameters:

Types of augmentation used during training:

- horizontal flip
- height shift: 0.1
- width shift: 0.1
- rotation angle range: 0 to 20 degrees
- shear: 0.1
- zoom: 0.1

Batch size: 32

Adam learning rate: le-5

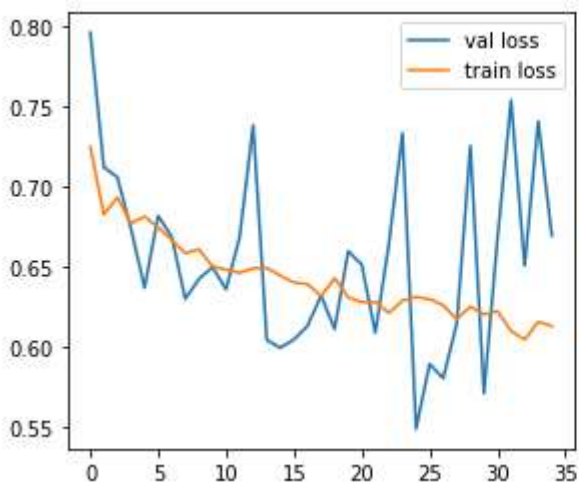
Layers of pre-existing architecture that was fine tuned

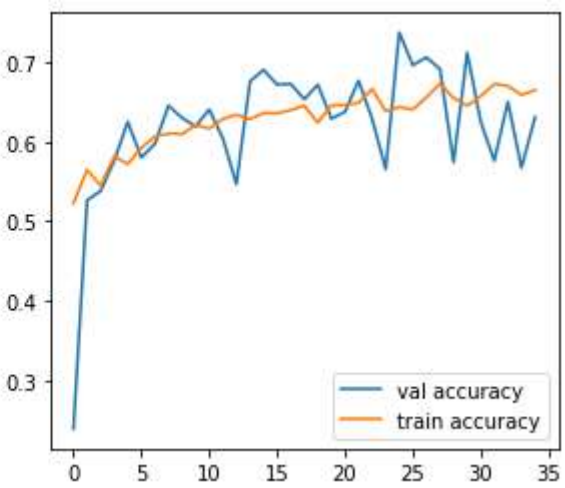
- The last 2 layers of VGG16 network: block5_conv3 + block5_pool

Layers added to pre-existing architecture

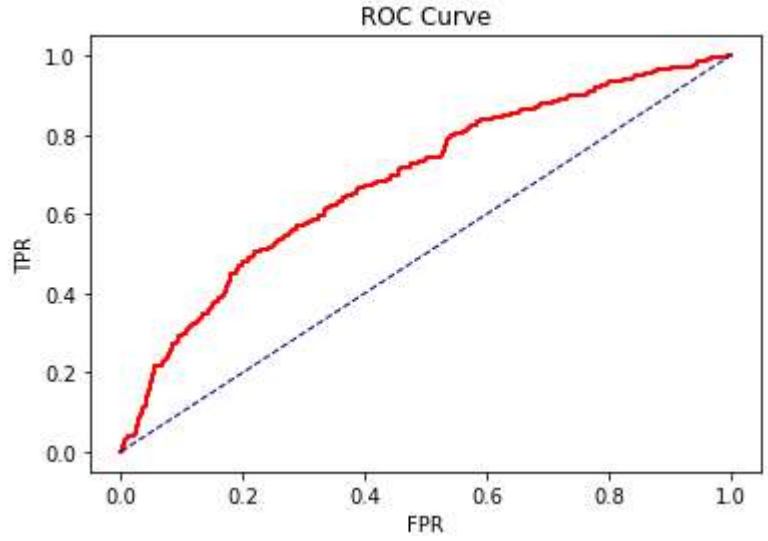
- flatten_1 (Flatten)
- dropout_1 (Dropout)
- dense_1 (Dense, 1024)
- dropout_2 (Dropout, 0.2)
- dense_2 (Dense, 512)
- dropout_3 (Dropout, 0.2)
- dense_3 (Dense, 256)
- dropout_4 (Dropout, 0.2)
- dense_4 (Dense) 1

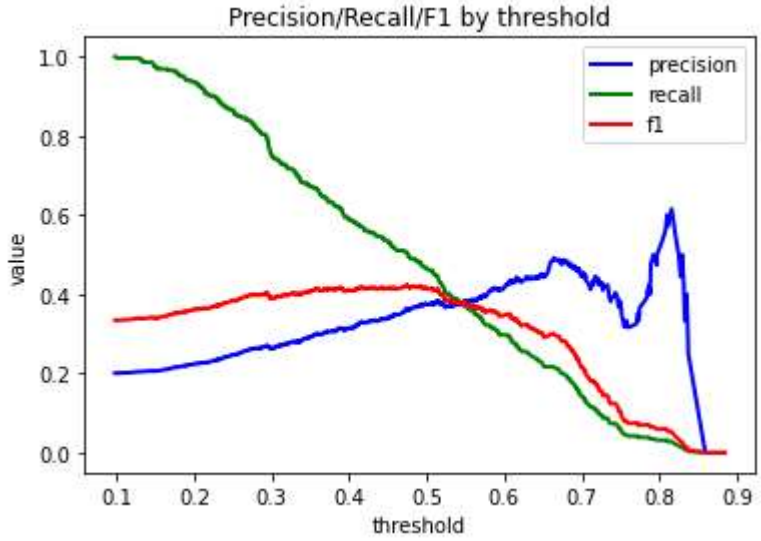
Algorithm training and performance visualization

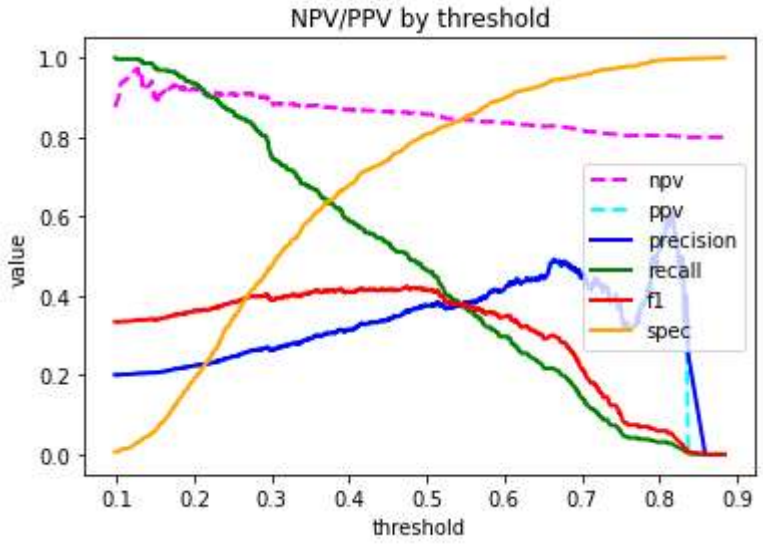




Model performance metrics based on threshold







As we can see the model has low precision, but higher recall, and maintains high negative predictive values

Final Threshold and Explanation:

The maximum F1 score for the model is 0.424 and it is achieved with a threshold value of 0.474. Below is the comparison of F1 score from the [CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning:]

Person or Device	**F1**	**95% CI 2sigma**	**68% CI 1sigma**
Radiologist 1	0.383	(0.309, 0.453)	(0.345, 0.417)
Radiologist 2	0.356	(0.282, 0.428)	(0.319, 0.392)
Radiologist 3	0.365	(0.291, 0.435)	(0.327, 0.399)
Radiologist 4	0.442	(0.390, 0.492)	(0.416, 0.467)
Radiologist Average	0.387	(0.330, 0.442)	(0.358, 0.414)
CheXNet	0.435	(0.387, 0.481)	(0.411, 0.458)
Codewired-PDA Max F1	0.424		

For simplicity and brevity, we did not calculate Confidence Interval (CI) and just assumes normal distribution. We will compare F1 score to 1-sigma Cis calculated from 2-sigma ones. Codewired-PDA F1 score is higher and outside of 68% (1sigma) CI for three radiologists out of four.

Comparing the F1 scores themselves, this model achieves higher maximum F1 score than the average score of the four radiologists in the study. State of the art neural network, as well as one radiologist from the study, do achieve higher F1 score, but the model's performance is comparable and, in many cases, exceeds the performance of human radiologists (in terms of F1 score).

Furthermore, 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, it makes sense to maximize recall and NPV even at the cost of small loss in precision. A good threshold value that achieves that is 0.3

Device	F1	Precision	Sensitivity/Recall	Specificity	NPV
Codewired-PDA Max F1	0.424	0.364	0.573	0.779	0.868
Codewired-PDA T = 0.28	0.400	0.263	0.832	0.418	0.907

If the model predicts negative, it is correct with 90.7% probability. If the model predicts positive, it is correct with 26.3% probability.

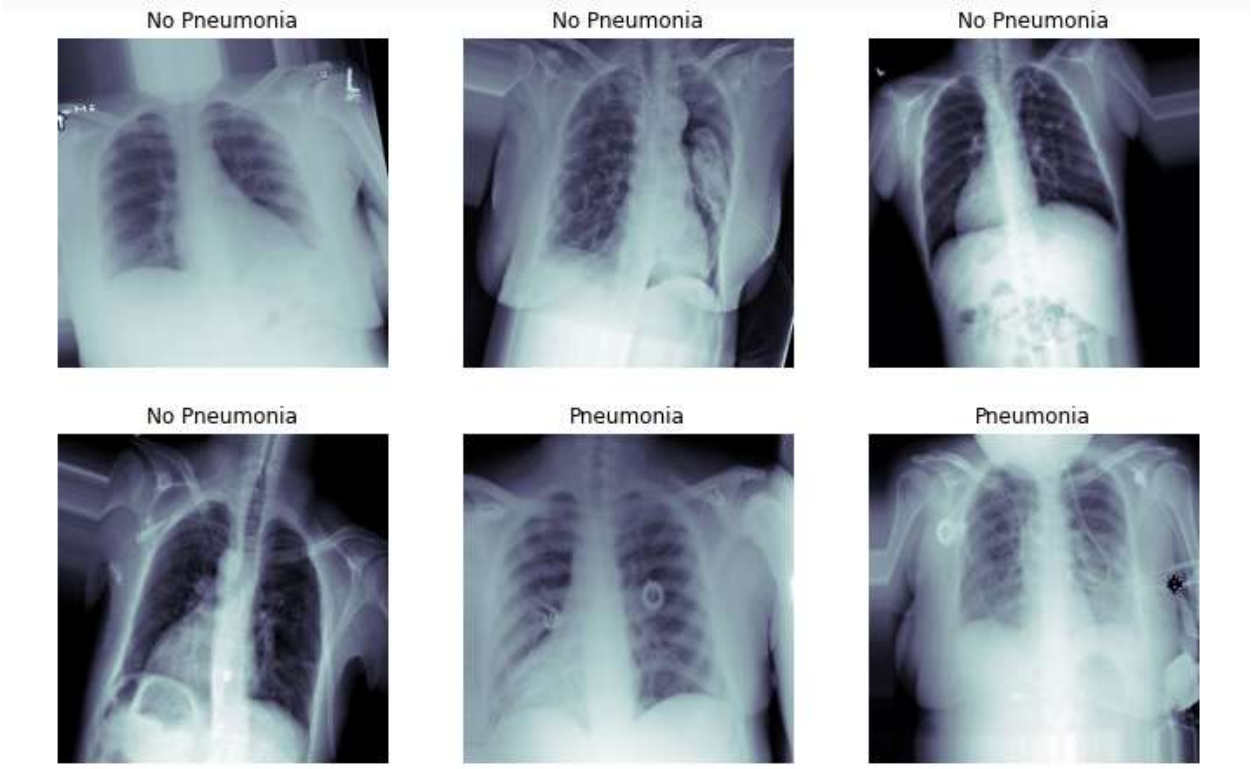
Out of all negative cases the model correctly classifies 41.8%, out of all positive cases it correctly classifies 83.2%.

4. Databases

Description of Training Dataset:

Training dataset consisted of 2290 chest x-ray images, with a 50/50 split between positive and negative cases.

Example Images



Description of Validation Dataset:

Validation dataset consisted of 1430 chest x-ray images, with 20/80 split between positive and negative cases, which more reflects the occurrence of pneumonia in the real world.

5. Ground Truth

The data is taken from a larger x-ray dataset from the NIH Chest X-ray Dataset that can be found HERE, 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:

1. Both men and women
2. Age 2 to 90
3. Without known comorbid thoracic pathologies listed in Page 1.

Ground Truth Acquisition Methodology:

The golden standard for obtaining ground truth would be to perform one of these:

- Sputum test
- Pleural fluid culture

You find out more from Mayo Clinic via this LINK

These tests are quite expensive, and in most cases, diagnosis is concluded by the physician based on radiologist's analysis/description. Since the purpose of this device is assisting the radiologist (not replacing him), the ground truth for the FDA Validation Dataset can be obtained as an average of three practicing radiologists (as a widely used 'silver standard'). The same method is used in the mentioned paper.

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 three human radiologists, as given in CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning, where a similar method is used to compare device's F1 score to average F1 score over three radiologists.

A 95% confidence interval given in the paper for average F1 score of 0.387 is (0.330, 0.442), so algorithm's 2.5% and 97.5% percentiles should also be calculated to get 95% confidence interval. This interval, when subtracted from the interval above to calculate the average, should not contain 0, which will indicate statistical significance of its improvement of the average F1 score. The same method for assessing statistical significance is presented in the above paper.