FDA SUBMISSION

Your Name: Sulagna S

Name of your Device: NN-Pneumonia

ALGORITHM DESCRIPTION

- **General Information**

Intended Use Statement: Intended to assist the radiologists in the screening of Pneumonia patients using chest x-rays.

Indications for Use: Indicated for use for screening pneumonia studies in children and adults up to the age of 90yrs in non- emergency situations. The chest x-rays must be in AP or PA position and the modality should be "DX" only.

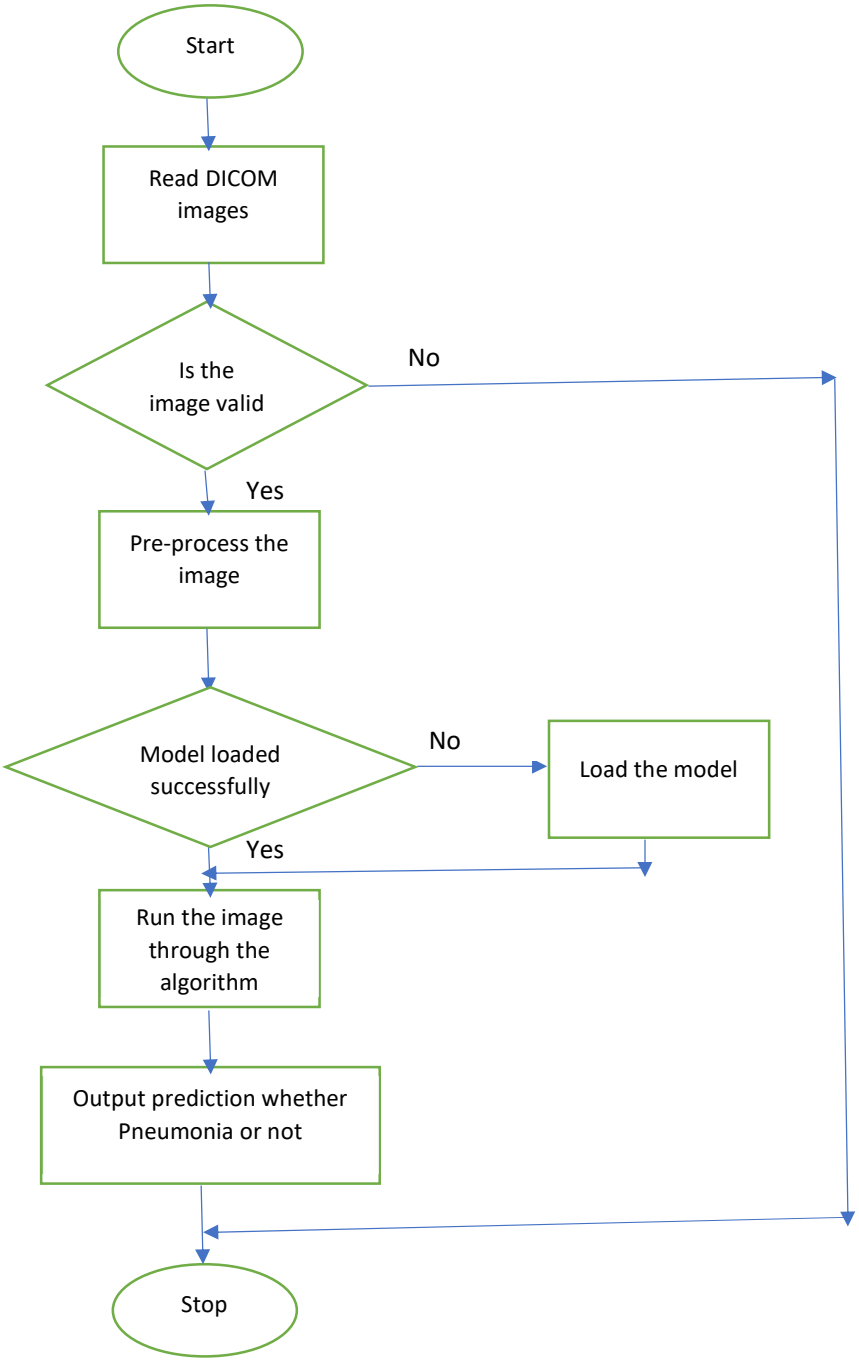
Device Limitations:

- Pneumonia can co-exist with one or more of thoracic pathologies like - Infiltration, Edema, Atelectasis, Effusion, Consolidation, Nodules, Pleural thickening, Mass, Cardiomegaly.
- It therefore becomes difficult to detect & diagnose Pneumonia if more than one of the conditions overlap.
- The algorithm may thus diagnose a condition as Pneumonia even if in reality it isn't Pneumonia but one of the conditions mentioned above.
- Or the algorithm may detect no Pneumonia if Pneumonia co-exists with one or more of the conditions mentioned above.

Clinical Impact of Performance:

- Model has a high recall but low precision rate.
- A high recall means that the algorithm has low a low False Negative (FN) rate. So, when a patient receives a negative report, it is most likely true.
- Low precision means that the algorithm has a high true false positive (FP) rate, and a radiologist will need to look into the x-ray to confirm the presence or absence of Pneumonia.
- The x-rays marked as negative can take a lower priority, which in turn would help in a quicker detection of positive cases.

- **Algorithm Design and Function**



DICOM Checking Steps: The DICOMs are checked for the following before running them through the algorithm

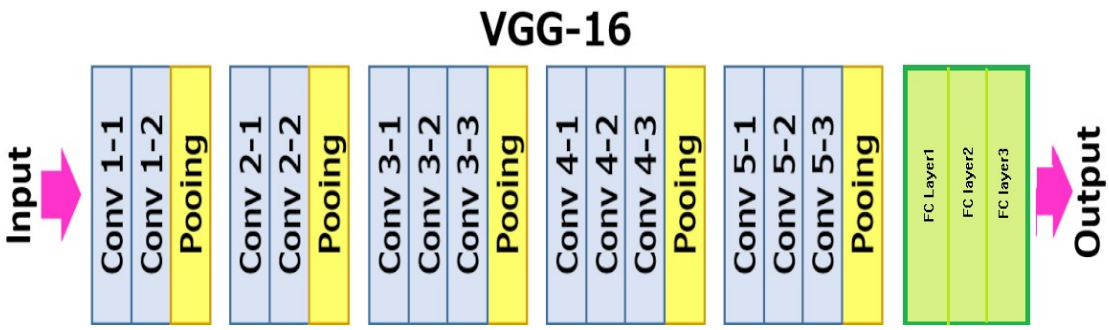
1. The body part - the body part should be “CHEST”
2. The patient position - the position should be either “PA” or “AP”
3. The modality - it should be “DX” i.e. X-ray images only

Pre-processing Steps: The following steps are taken before running the image through the model:

1. Images are downscaled from the original image of size 1024x1024 to 224x224.
2. Convert from grey scale to RGB format.
3. Convert to format that the algorithm accepts i.e. (1,224,224,3)
4. Since a pretrained VGG16 model was used to design this algorithm, the same pre-processing as was done for the images the VGG16 network was trained on, as been done for the training set.

CNN Architecture: A pretrained VGG16 network has been used with changes to the last classification layers. The original VGG16 has classification layers to predict one of 1000 categories, but we only need 2 – Pneumonia and No Pneumonia.

We thus take the initial 18 layers – all the convolution and max pooling layers and then add a classification layer that predicts one of 2 categories.



- **Algorithm Training**

PARAMETERS

Types of augmentation used during training:

The following image augmentation was done

- Horizontal flip (and no vertical flip)
- Random rotation with max value 20 degrees
- Height shift range: 0.1
- width shift range: 0.1
- Sheer range: 0.1
- Zoom range: 0.1
- Standardisation as per VGG 16 pre-processing

Batch size: 32

Optimizer learning rate: 0.00005

Layers of pre-existing architecture that were frozen: VGG16 all layers till the Dense layers (first 18 layers)

Layers of pre-existing architecture that were fine-tuned: None. The classification layers were dropped.

Layers added to pre-existing architecture: 3 fully connected linear layers have been added to the pre-existing architecture.

Layer 1 hidden dim = 4096, and a RELU activation function

Layer 2 hidden dim = 128, and a RELU activation function

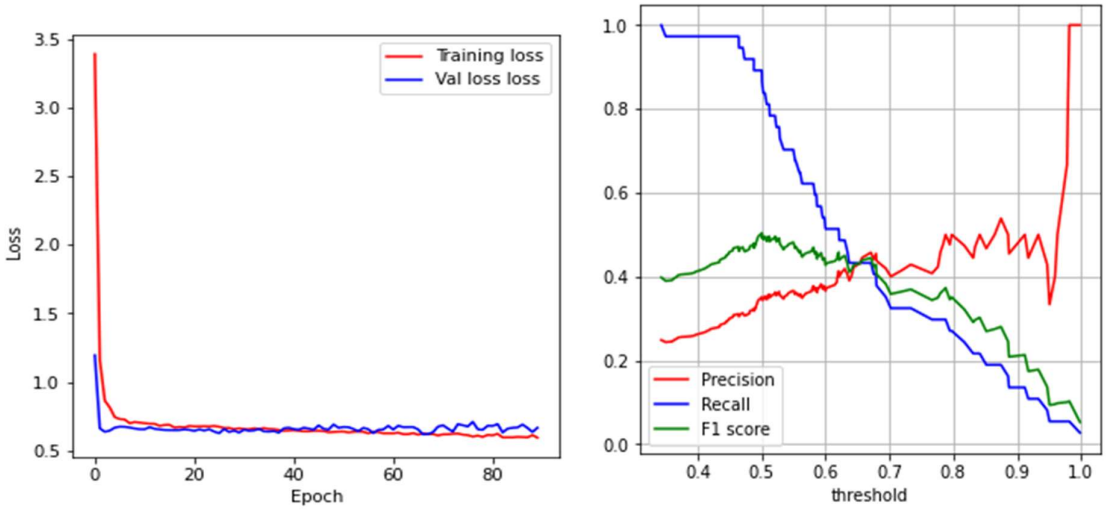
Layer 3 is the final layer with output size =1 & a sigmoid activation function

There is a dropout layer between 1 and 2 fully connected layers and 2 and 3 fully connected layers.

FINAL THRESHOLD AND EXPLANATION:

The training and validation loss are as in the graph below. Further training with other combinations of batch size, learning rate, hidden dimensions did give us a lower training loss

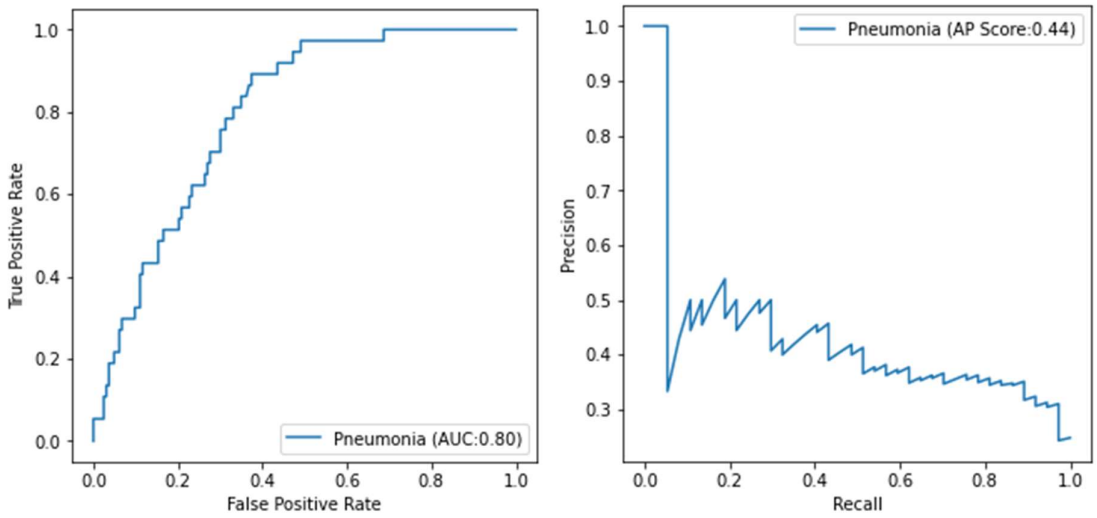
but also gave a high validation loss. Stopping at about 80-90 epochs seemed like a good balance.



By looking at the precision, recall and f1-score variation with threshold value, the threshold has been decided to be 0.5223 which gives precision, recall and f1score as below:

```
Precision is: 0.35443037974683544
Recall is: 0.7567567567567568
Threshold is: 0.52280563
F1 Score is: 0.4827586206896552
```

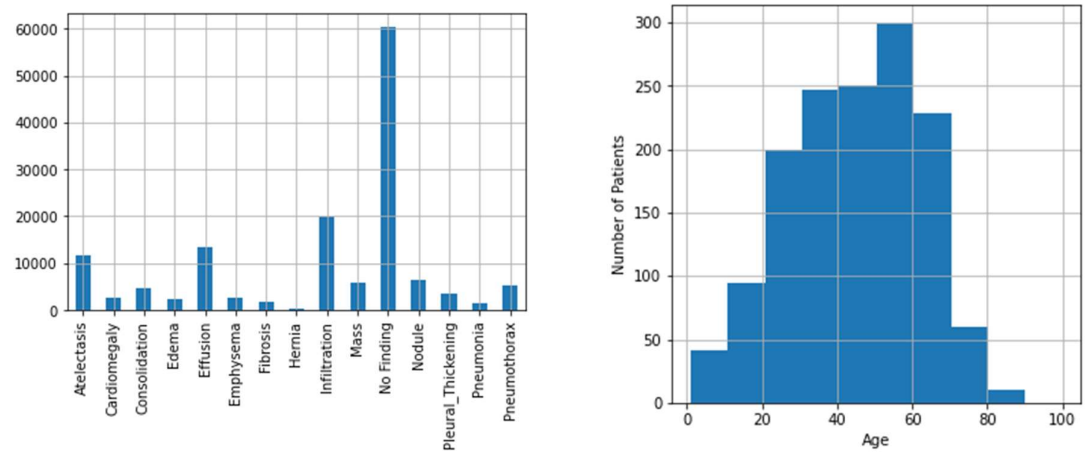
- The graph above on the right, shows how precision, recall and f1-score vary with the different threshold values. Precision and recall are inversely related and f1 is their harmonic mean.
- F1score takes into false positive and false negative rates. Thus, an f1score closer to 1 is good and desirable, while a f1score closer to 0 isn't. Low f1 score basically indicates that one of precision/recall is very low.
- For my model, I want a threshold that will not optimise one (precision or recall) at the cost of the other. I want a good balance between the 2.
- From the graph above, I chose a threshold of 0.522 - which is nearly the models max f1 score, with a relatively high recall and a not too bad precision.
- Recall has been given more importance since in the clinical setting, we don't want to miss a positive case (i.e., recall is more important)

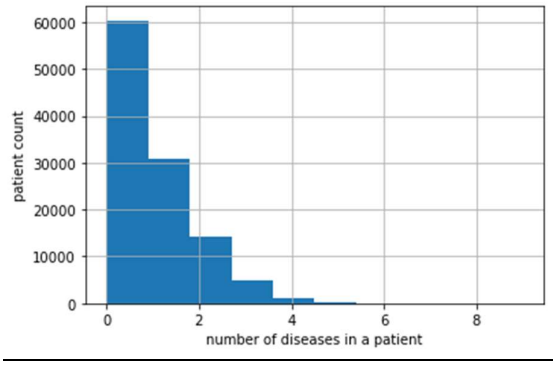


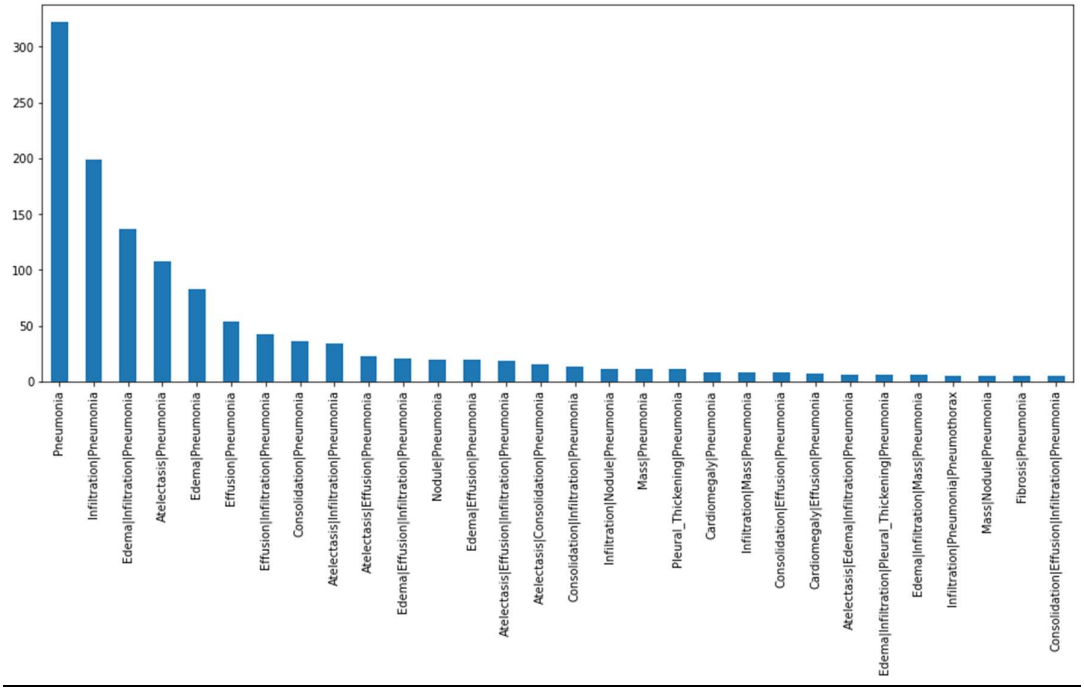
- **<u>Databases</u>**

The original dataset used for designing and training the algorithm to detect Pneumonia has the following characteristics:

- Total x-rays: 112120
- Age distribution: 2 to 412 years.
- Males and Females
- Number of x-rays with no disease: 60361 (53.83%)
- Patients with one or more diseases: 51758 (46.16%)
- Number of x-rays indicating Pneumonia: 1431 (1.27%)
- Number of x-rays indicating no Pneumonia: 50328 (44.88%)







Description of Training Dataset: The rows with incorrect ages of >100 have been dropped.

- The training dataset has been carefully chosen so as to keep it balanced as compared to the highly imbalanced original dataset. The ratio of x-rays with Pneumonia (positive cases) to those without Pneumonia (negative cases) has been set to 1:1
- Number of x-rays with Pneumonia in the Training set:1144
- Number of x-rays with no Pneumonia in the Training set: 1144

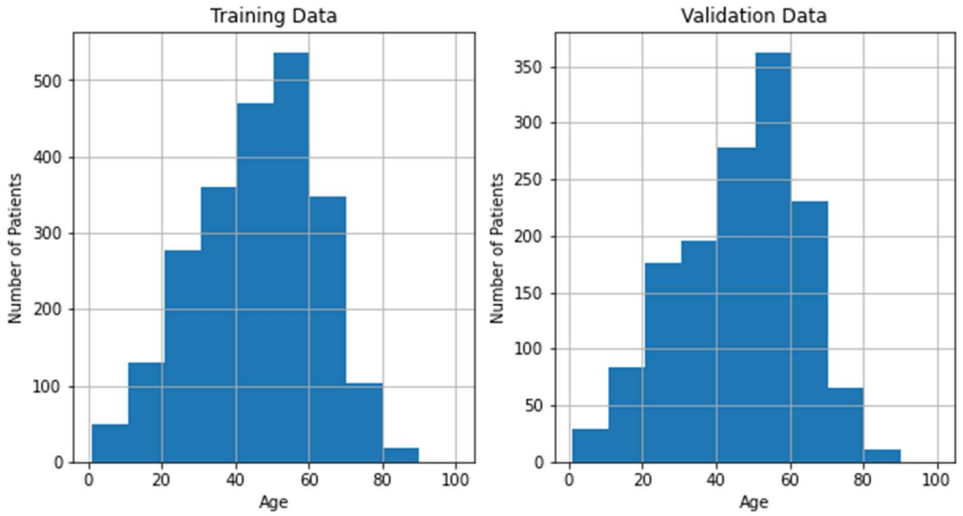
Description of Validation Dataset:

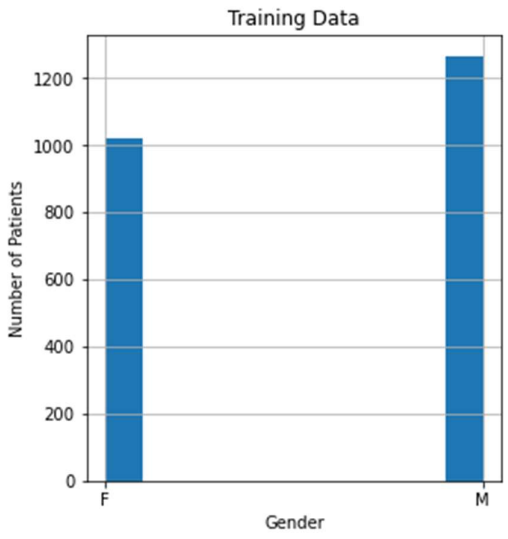
- The validation set continues to be imbalanced to represent as closely as possible the real-world scenario. However, the imbalance ratio has been adjusted. The Ratio of

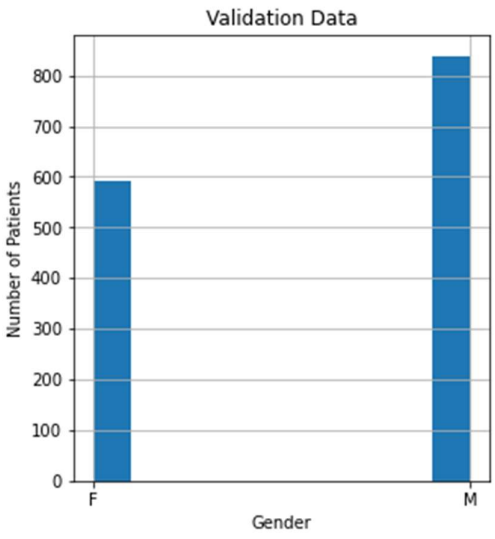
Number of Pneumonia x-rays to number of non-pneumonia x-rays has been set to 1:4

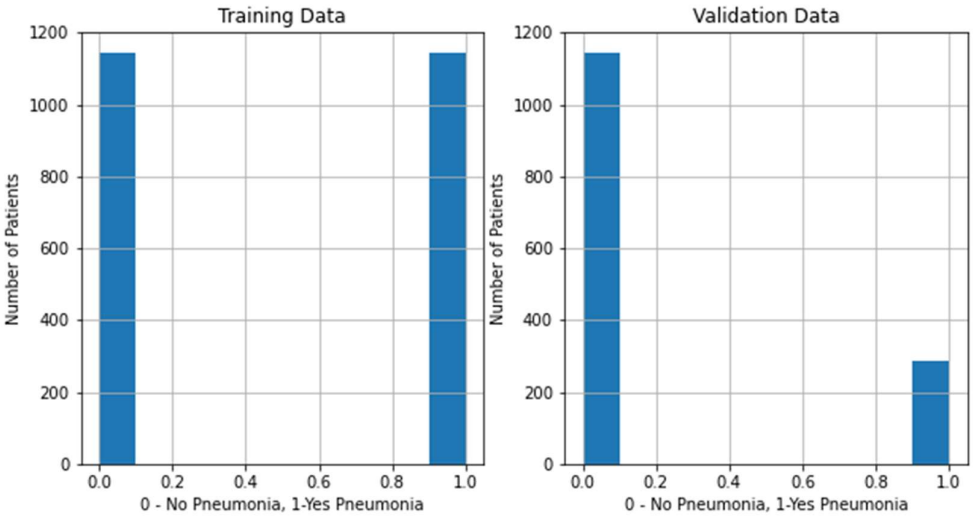
- There is no patient overlap between the training dataset and validation dataset.

The age, gender and pneumonia distribution in the training and validation set are as shown in the graphs below.









- **Ground Truth**

The 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. 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 biggest limitation of this dataset is that image labels were NLP-extracted so there could be some erroneous labels but the NLP labelling accuracy is estimated to be >90%.

Another way of obtaining ground truths would be to get the x-rays examined by radiologists. Radiologists are experts at diagnosing diseases by looking at x-rays.

- **FDA Validation Plan**

Patient Population Description for FDA Validation Dataset:

- Chest X-rays of children and adults up-to the age of 90 years.
- The positions for x-rays can be PA or AP.
- Presence of other diseases like Atelectasis, Cardiomegaly, Consolidation, Edema, Effusion, Emphysema, Fibrosis, Hernia, Infiltration, Mass, Nodule, Pleural Thickening, and Pneumothorax is acceptable.
- The validation set should contain at-least 20% Pneumonia cases.

Ground Truth Acquisition Methodology: X-rays analysed by a radiologist is acceptable as ground truth, though the Gold standard would be via a sputum test, pleural fluid culture test or a bronchoscopy which are expensive and time consuming.

Algorithm Performance Standard:

As mentioned in this paper https://arxiv.org/abs/1711.05225 – “CheXNet -Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning”, a deep learning algorithm is capable of achieving high performances, comparable to that of a Radiologist.

The paper concluded that the CheXNet with a F1score of 0.435 (95% CI range: 0.387,0.481) performs better than radiologists which means it can replace radiologists. The radiologists have an average f1score of 0.387(95% CI range: 0.330,0.442)

This NN-Pneumonia with a F1score of 0.4827 (95% CI range: 0.4492, 0.5185), AUC =0.8, recall =0.7567 is designed only to assist radiologists.