FDA SUBMISSION

NAME : Tenzin Gurme
Device Name : Pneumonia Detector

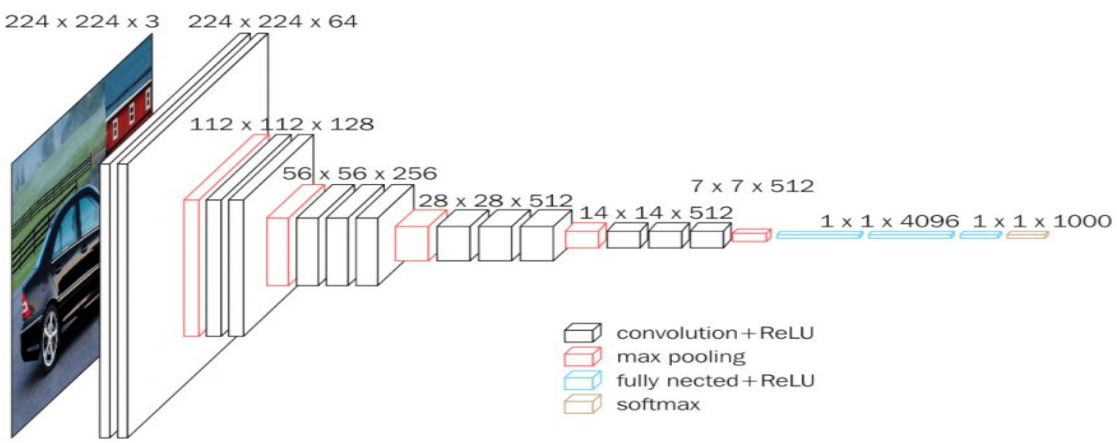
General Information:

Intended Use Statement :
This algorithm is intended to be used for Patients diagnosed with Pneumonia, aged between 5-100 who have no abnormality in Chest X-ray before.
Indication of Use Statement:
This algorithm could be used for prior (early) detection of Pneumonia in Patients through Chest X-ray and for screening purposes.

Device Limitation:
The result of algorithm may lead to false-negative in patients suffering from diseases like Atelectasis, Cardiomegaly, Consolidation, Edema, Effusion, Emphysema, Fibrosis, Hernia, Infiltration, Mass, Nodule, Pleural Thickening, or Pneumothorax in a chest x-ray as prior for Pneumonia detection may lead increase sensitivity and if a chest x-ray may lead to false-positive pneumonia classifications as specificity is 1, however, the presence of Pneumonia can be accurately detected from this algorithm
The model is trained specifically for frontal X-rays. So, X-rays from other angles might not be accurate .

Clinical Impact of Performance:
The algorithm has a quite good accuracy in detecting Pneumonia which can be highly beneficial in automating the process of detecting Pneumonia cases directly from the chest X-ray which does not save a lot of time in diagnosing the patient and also preparing the patient mentally as well. And reduce Burn out.

Algorithm Design and Function:



The algorithm is designed based on the pre-trained vgg 16 model. After the last layer of the vgg16 model. Following layers are added:

```
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
model_1 (Model)              (None, 7, 7, 512)         14714688
_________________________________________________________________
batch_normalization_1 (Batch (None, 7, 7, 512)         2048
_________________________________________________________________
conv2d_1 (Conv2D)            (None, 7, 7, 1024)        525312
_________________________________________________________________
dropout_1 (Dropout)          (None, 7, 7, 1024)        0
_________________________________________________________________
batch_normalization_2 (Batch (None, 7, 7, 1024)        4096
_________________________________________________________________
conv2d_2 (Conv2D)            (None, 7, 7, 256)         262400
_________________________________________________________________
dropout_2 (Dropout)          (None, 7, 7, 256)         0
_________________________________________________________________
average_pooling2d_1 (Average (None, 1, 1, 256)         0
_________________________________________________________________
batch_normalization_3 (Batch (None, 1, 1, 256)         1024
_________________________________________________________________
conv2d_3 (Conv2D)            (None, 1, 1, 1)           257
_________________________________________________________________
reshape_1 (Reshape)          (None, 1)                 0
=================================================================
Total params: 15,509,825
Trainable params: 3,151,361
Non-trainable params: 12,358,464
```

The model takes the input of X-ray images with (224, 224, 3) and gives output probability of Pneumonia / Non-pneumonia.

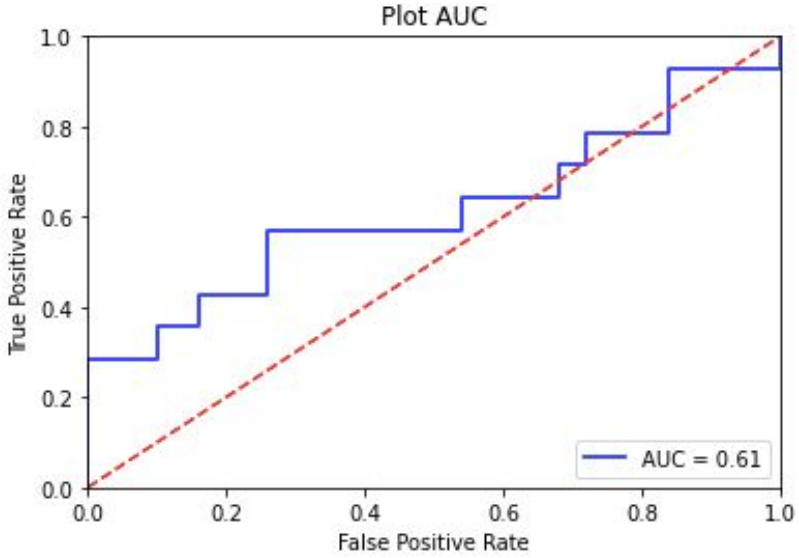
- VGG16 model with non-trainable layers with pre-trained ImageNet weights.
- There 3 Conv2D layers with different filters works as a Dense layer to fine tune the weights dropout layers are used after the convolution layer with 0.5 probability.
- Average pooling and batch normalization used to normalize the inputs from the before layer.
- Last convolution layer gives output in the range 0-1.

Dicom checks:

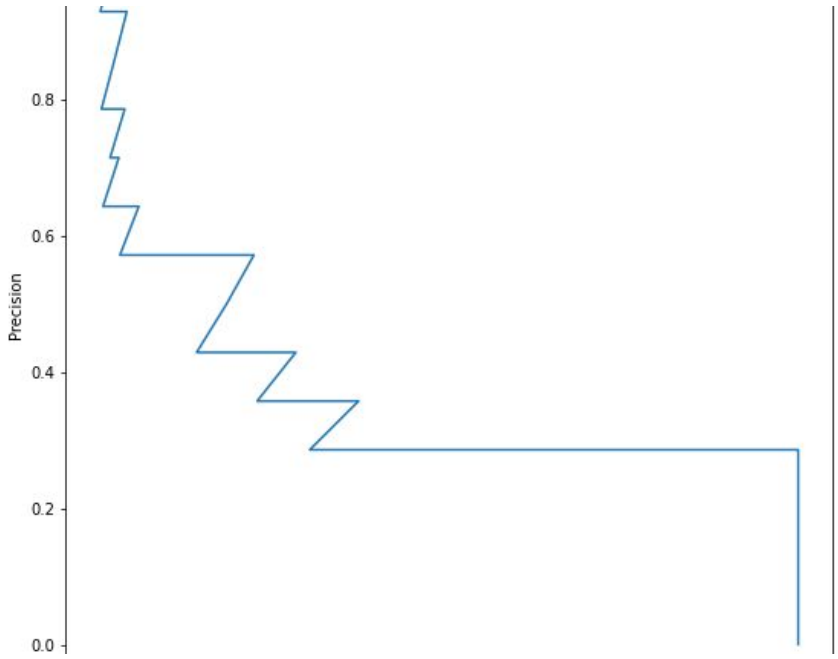
- Input X-ray image must be of ‘CHEST’.
- Modality must be ‘DX’ Patients position must be ‘Anteroposterior’ or ‘Poster-anterior.’

For training's data augmentation, the image is rescaled to 1/255, horizontal flip, height shift range of 0.05, width shift range of 0.1, rotation range of 10, shear range of 0.1, the zoom range of 0.05, so that model will consider those image as well which are slightly rotated or zoomed. For the training image generator, a batch size of 64 images is used.

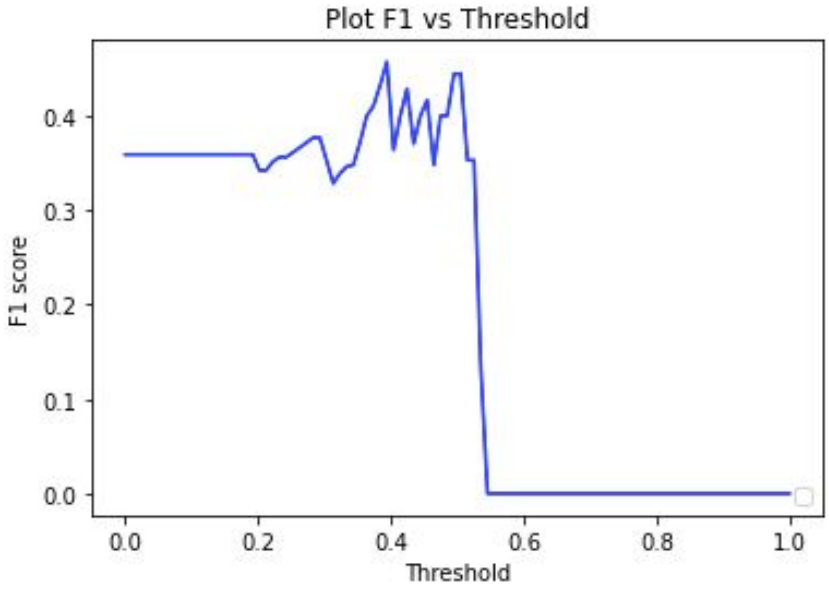
The algorithm was trained based on an RMSprop optimizer and with a learning rate of 10^-4. Then the model remaining is fine-tuned for this case for about more than 4 epochs until the model loss is stabilized.
The current model was trained up to 0.6659 training accuracy and training loss of 0.6412 which has a final validation loss of 0.5393



The algorithm has an area under the curve for True positive rate and false positive rate of 0.5.



- Precision-Recall curves summarize the trade-off between the true positive rate and the positive predictive value for a predictive model using probability thresholds.
- In medicine, better precision-recall is needed to reduce False positives. Curves help to identify better thresholds for algorithms.

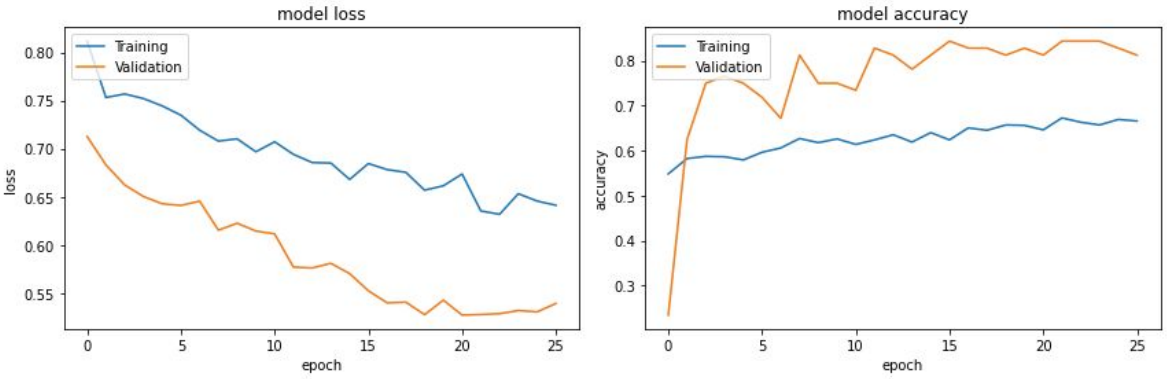



```
Threshold of 0.50 gives best accuracy at 0.8438

0.4444444444444445


CONFUSION MATRIX **************************
[[ 4 10]
 [ 0 50]]

TEST METRICS ******************************
Accuracy: 84.375%
Precision: 83.33333333333334%
Recall: 100.0%
F1-score: 90.9090909090909
```



Behavior of training and validating loss.

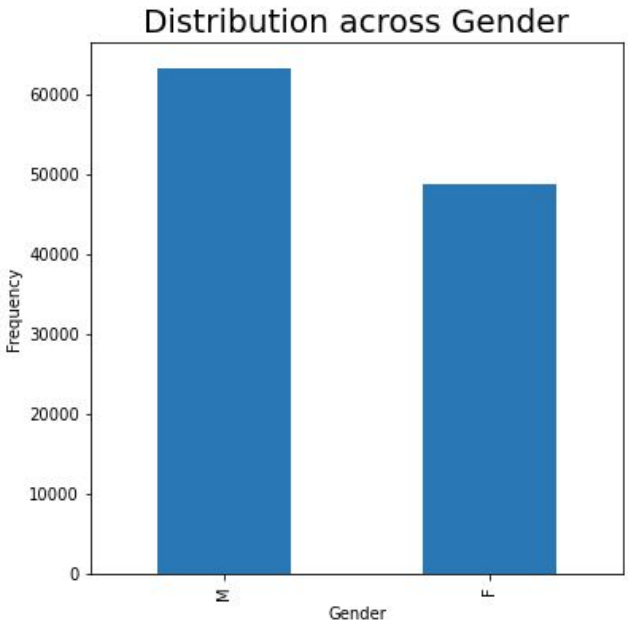
Database:

The dataset was curated by the NIH. 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:

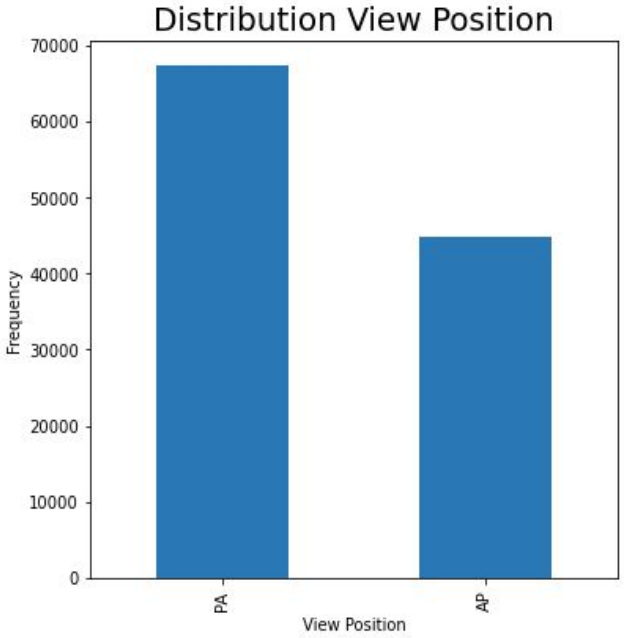
- Atelectasis
- Effusion
- Consolidation
- Pneumonia
- Infiltration
- Pleural thickening
- Pneumothorax
- Cardiomegaly
- Edema
- Nodule
- Emphysema
- Mass
- Fibrosis
- Hernia

The dataset consists of 63340 male participants and 48780 female participants. The positioning of the participant during the chest X-ray is AP in 44810 cases and PA during 67310 cases.

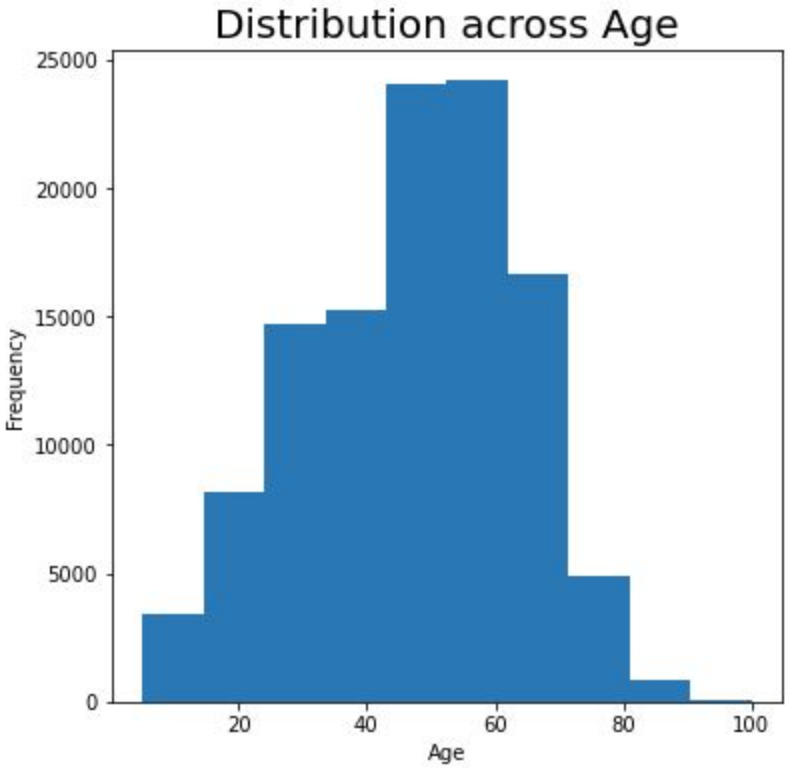
```
M    63340
F    48780
Name: Patient Gender, dtype: int64
```



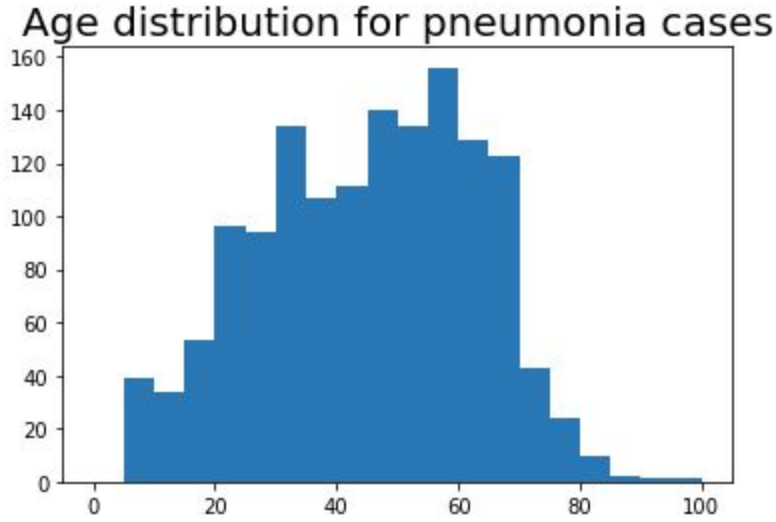

```
PA    67310
AP    44810
Name: View Position, dtype: int64
```

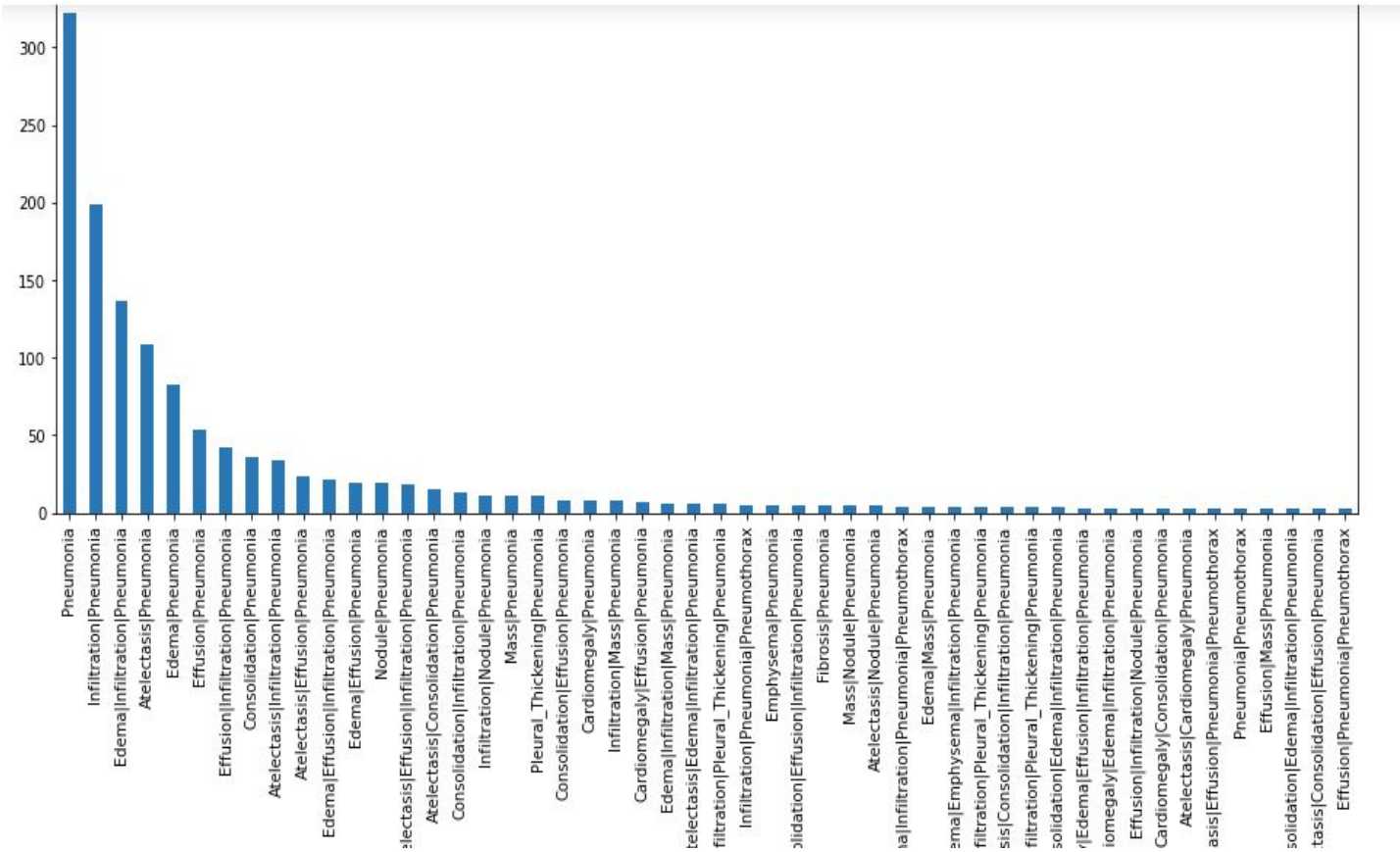


Age distribution of Patients(5-100: remove outlier) :

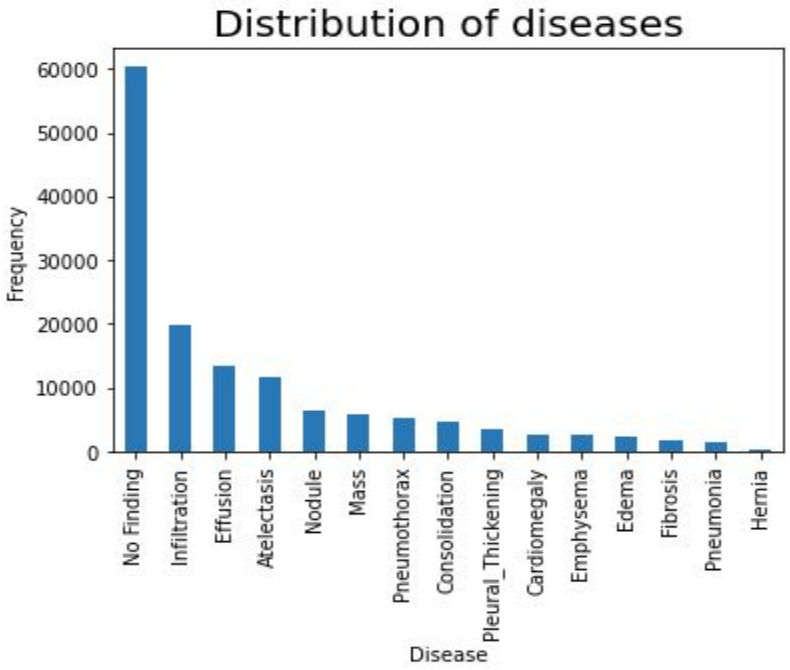


There are 1413 pneumonia cases in the given dataset which can be in the presence of different 15 images. Age distribution shows that there are more chances of Pneumonia in age 20-70.

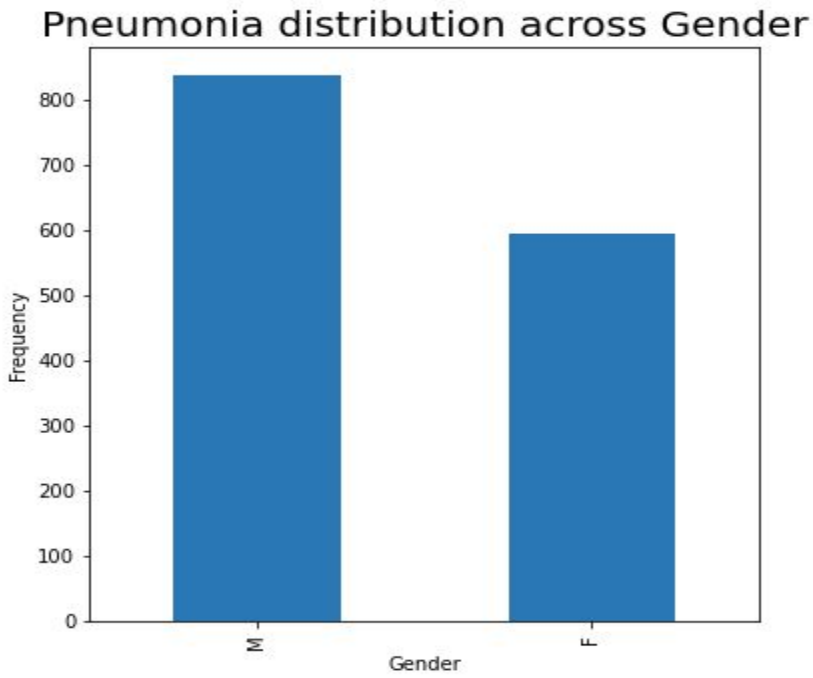




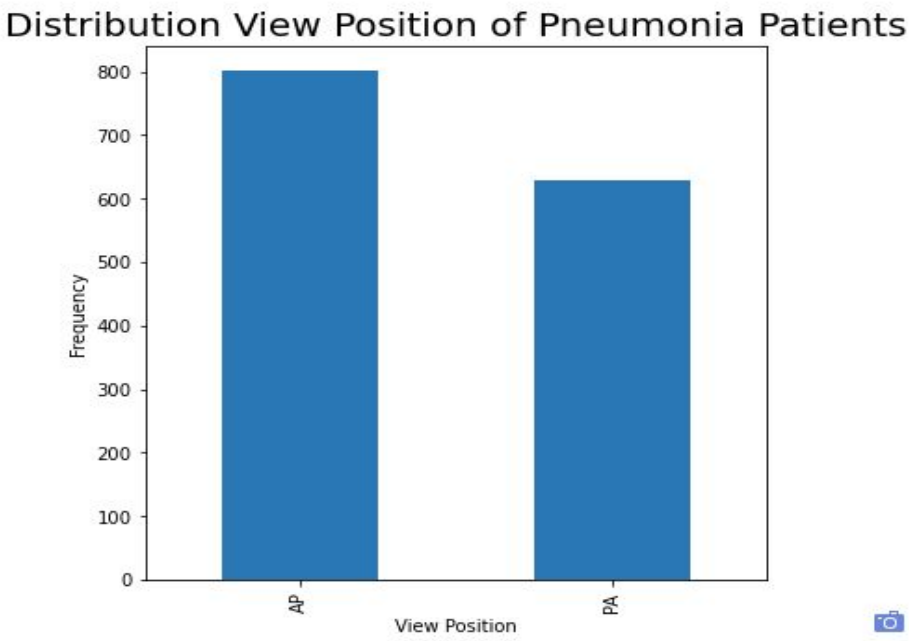
```
Pneumonia case: 1431.0
```




```
M    838
F    593
Name: Patient Gender, dtype: int64
```



```
AP     801
PA     630
Name: View Position, dtype: int64
```



For training the model, we have split the data in 80-20%.with 50% cases of pneumonia in training and 25% of cases in validation.

```
    ## Either build your own or use a built-in library to split your original d
    ## that can be used for training and testing your model
    ## It's important to consider here how balanced or imbalanced you want each
    ## for the presence of pneumonia

    train_df, valid_df = train_test_split(data,
                                          test_size = 0.2,
                                          stratify = all_xray_df['pneumonia_class'])

    p = train_df[train_df.Pneumonia==1].index.tolist()
    np = train_df[train_df.Pneumonia==0].index.tolist()
    np_sample = sample(np,len(p))
    train_df = train_df.loc[p+np_sample]

    p2 = valid_df[valid_df.Pneumonia==1].index.tolist()
    np2 = valid_df[valid_df.Pneumonia==0].index.tolist()
    np_sample2 = sample(np2,len(p2)*3)
    valid_df = valid_df.loc[p2+np_sample2]

    return train_df, valid_df

train_df, valid_df = create_splits(all_xray_df)
```

In [8]:
```
train_df['Pneumonia'].sum()/len(train_df)
```

```
: train_df['Pneumonia'].sum()/len(train_df)
: 0.5

: valid_df['Pneumonia'].sum()/len(valid_df)
: 0.25
```

Ground Truth:

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 details about the labeling process can be found here.

FDA Validation Plan:

Patient Population Description for FDA Validation Dataset:

For validation of the algorithm, the collected dataset should be made up of chest X-rays between the ages 5-100 for both male and female. However, it should be made sure that the validation set did not contain a patient with prior history of Pneumonia and the patient with other diseases like Atelectasis, Cardiomegaly, Consolidation, Edema, Effusion, Emphysema, Fibrosis, Hernia, Infiltration, Mass, Nodule, Pleural Thickening, or Pneumothorax should be checked again with Radiologists for False Positive.

Ground Truth Acquisition Methodology:

For the acquisition of the chest X-ray, the patient can be in any position anterior posterior or posterior-anterior. The patient should also be checked for other diseases mentioned above, as it may result in False positive. In other cases for labeling, try to obtain biopsy data or having radiologists creating labels that should then be weighted across the radiologists to create a single label (a.k.a. “Silver standard”) for each image in the set. This process is better than NLP labeling but also has some conflicts. So, weights should be given properly.

Algorithm Performance Standard:

	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)

My algorithm ends with the F1 score as 0.4444 which is far greater than Radiologist average score 0.387. Thus proving my model is pretty accurate in detection.