FDA Submission

Your Name: Shivank Yadav

Name of your Device: PneumoNet

Algorithm Description

1. General Information

Intended Use Statement: Assist Radiologist in classification of Pneumonia on Chest X-ray.

Indications for Use: Improve Radiologist's workflow by prioritizing chest X-rays with higher probability of Pneumonia. To be used for patients of both gender with ages between 1-90.

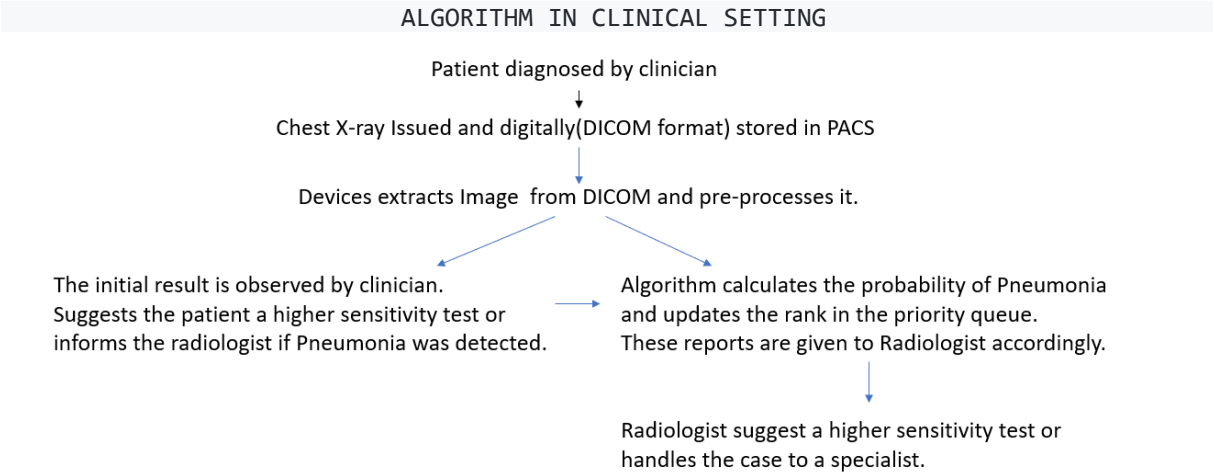
Device Limitations: GPU/Cloud Computing Instance may be required with High workload if to be used in Emergency settings. Should not use model for patients older than 90. Edema and Consolidation has a lot of similarity in their mean intensity distributions with that of Pneumonia and thus our model can misclassify here and result in False Positives.

Clinical Impact of Performance:

If we are using the device for screening and want minimum false negatives, then we need higher sensitivity(recall). That is no patient who has a disease should be marked negative.

Increasing recall will decrease precision. But if we let precision drop too much, this will result increase in false positives and thus cases of lower interest will be given more priority while being queued for the radiologist to look at.

2. Algorithm Design and Function

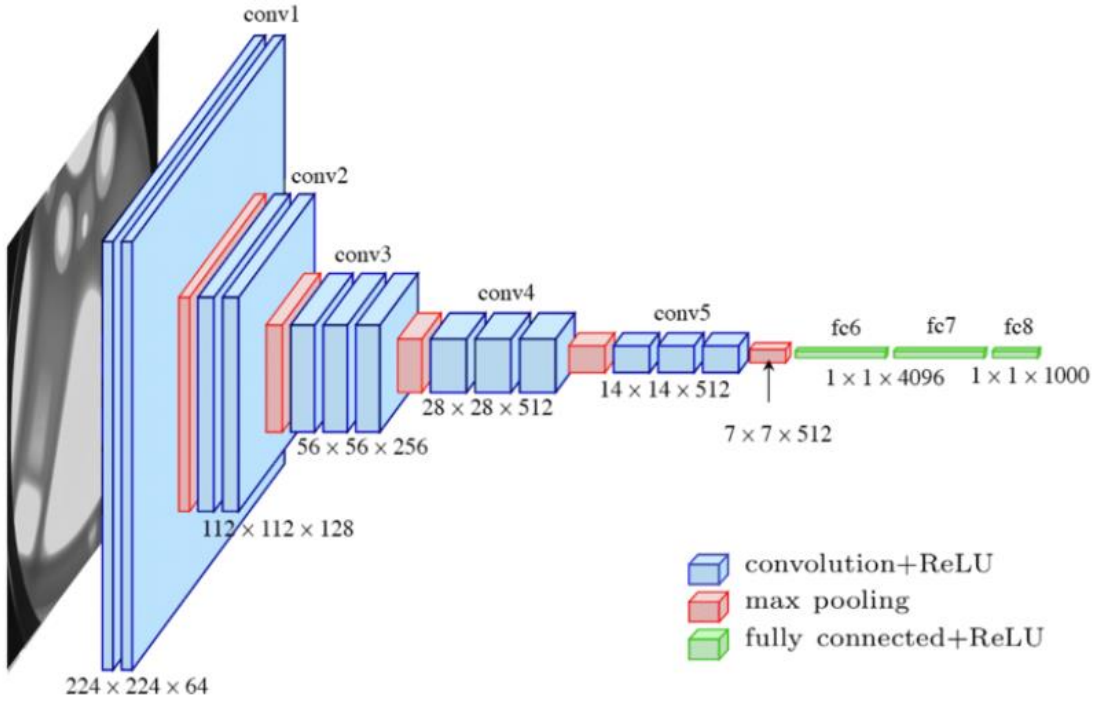


DICOM Checking Steps: Pydicom module in python library was used to extract Images as well as metadata form DICOM files.

We look at Patient Age, Patient Sex, Patient Position , Modality, Findings and Body Part Examined and check if this matches the distribution of data we trained on.

Preprocessing Steps: While Training, the Images were Standardized (using mean and deviation) or rescaled by 1/255.0. Then it was augmented using Keras ImageDataGenerator to account for variance in real world and lack of Data. Finally Resized by (1,244,244,3) and pushed into a modified VGG CNN.

CNN Architecture: Used VGG-16 as base model initialized with Imagenet weights. Replaced last layer by combination of dense and Dropout layers as follows:




```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
model (Model)                (None, 7, 7, 512)         14714688
_________________________________________________________________
flatten (Flatten)            (None, 25088)             0
_________________________________________________________________
dropout (Dropout)            (None, 25088)             0
_________________________________________________________________
dense (Dense)                (None, 1024)              25691136
_________________________________________________________________
dropout_1 (Dropout)          (None, 1024)              0
_________________________________________________________________
dense_1 (Dense)              (None, 512)               524800
_________________________________________________________________
dropout_2 (Dropout)          (None, 512)               0
_________________________________________________________________
dense_2 (Dense)              (None, 256)               131328
_________________________________________________________________
dense_3 (Dense)              (None, 1)                 257
=================================================================
Total params: 41,062,209
Trainable params: 33,426,945
Non-trainable params: 7,635,264
```

Finetuned the whole network from 15th layer of VGG Net.

3. Algorithm Training

Parameters:

- Types of augmentation used during training:

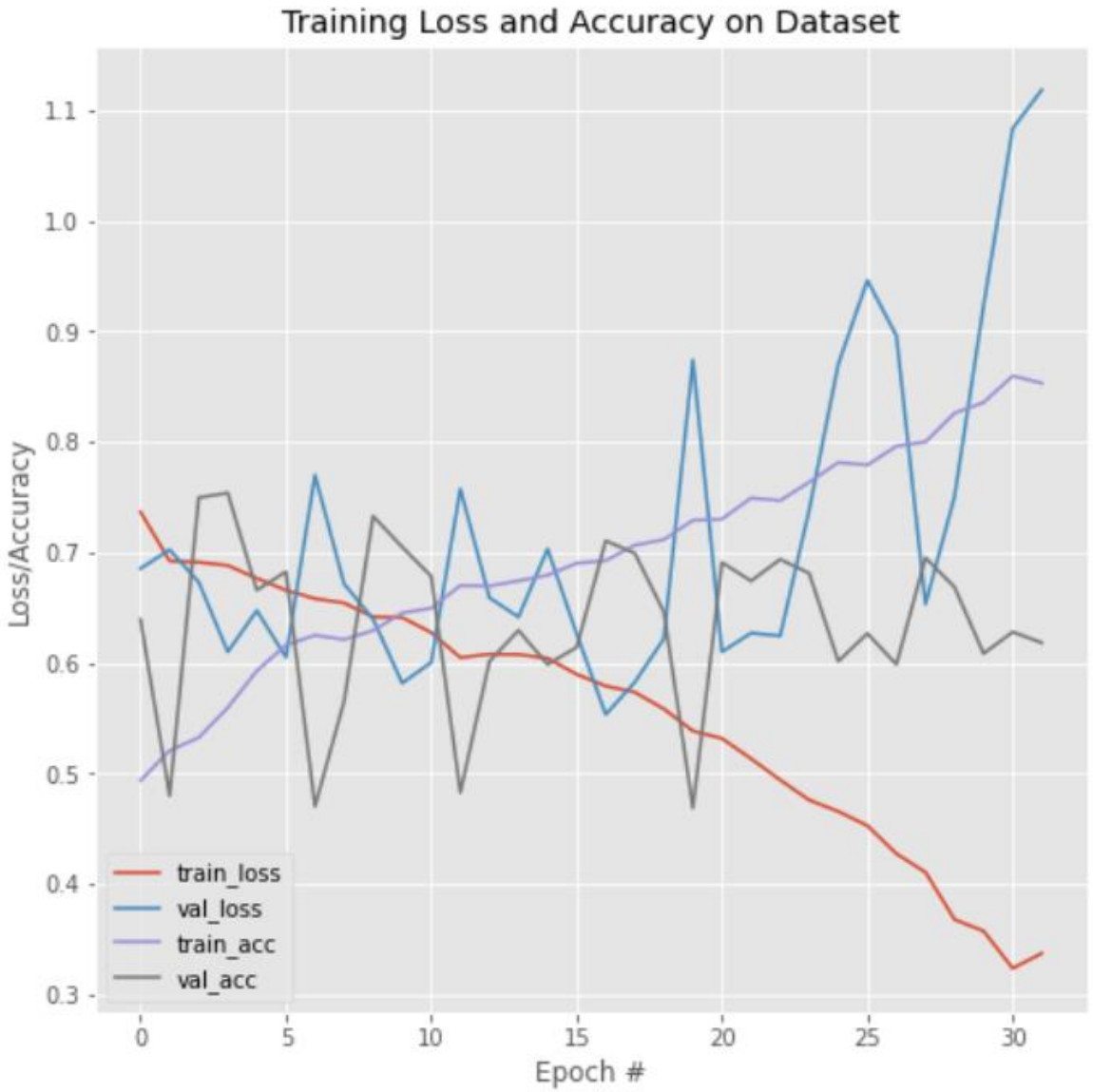
  Training Set

  ```
  * Normalization by standardization(using mean and deviation) or just
  rescalling by 1.0/255.0
  * horizontal_flip=True,
  * vertical_flip =False,
  * height_shift_range=0.1,
  * width_shift_range=0.1,
  * rotation_range=15,
  * shear_range=0.1,
  * zoom_range=0.1
  ```

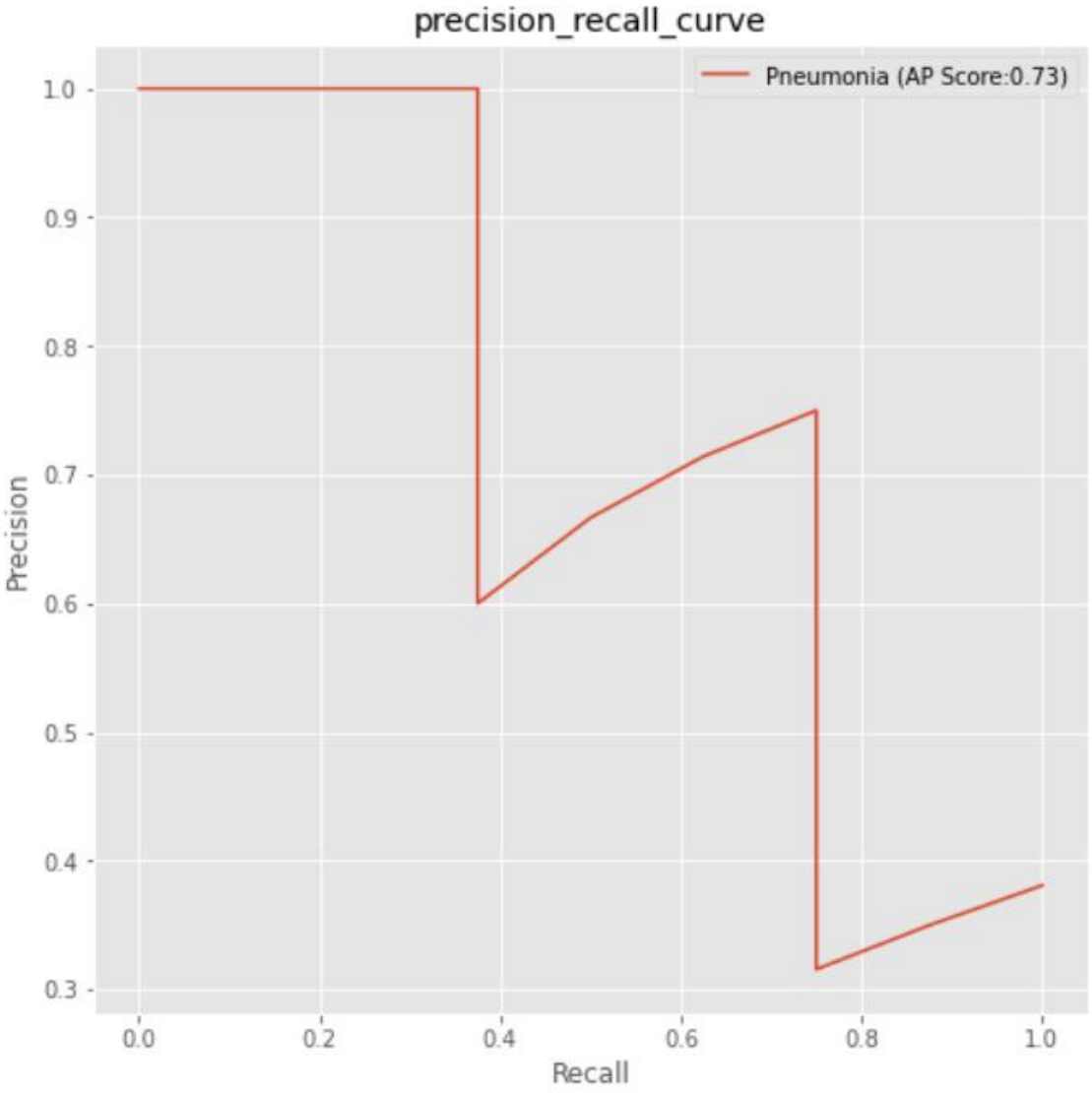
  Validation set was just Normalized using same technique as training set.

- Batch size : 32 for both validation and training set.
- Optimizer learning rate : 1e-4
- Layers of pre-existing architecture that were frozen : first 16 including input_layer
- Layers of pre-existing architecture that were fine-tuned : All the layers after first 16
- Layers added to pre-existing architecture : Dropout + dense (x3) -> Dense(256) -> Dense(1) // sigmoid layer/output layer

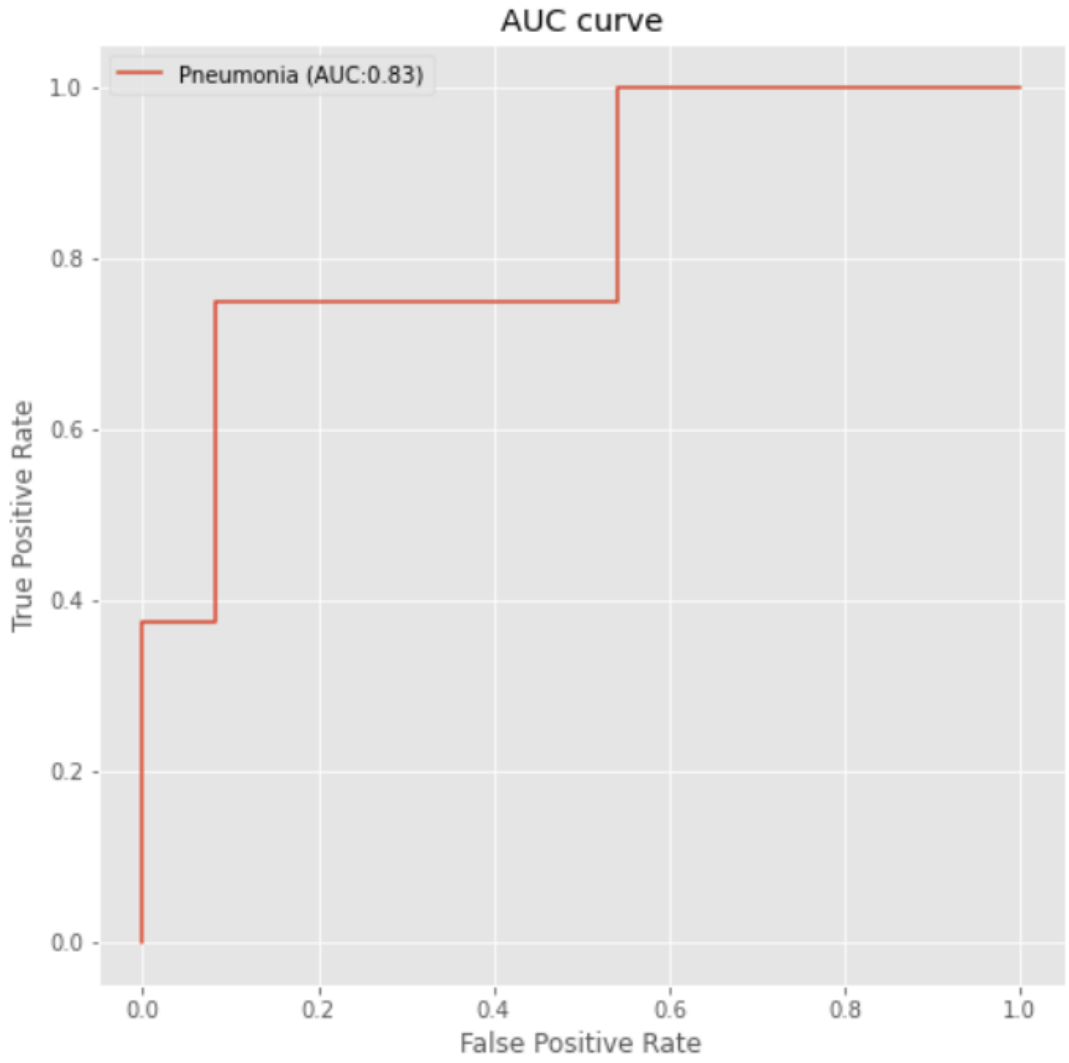
Algorithm training performance visualization



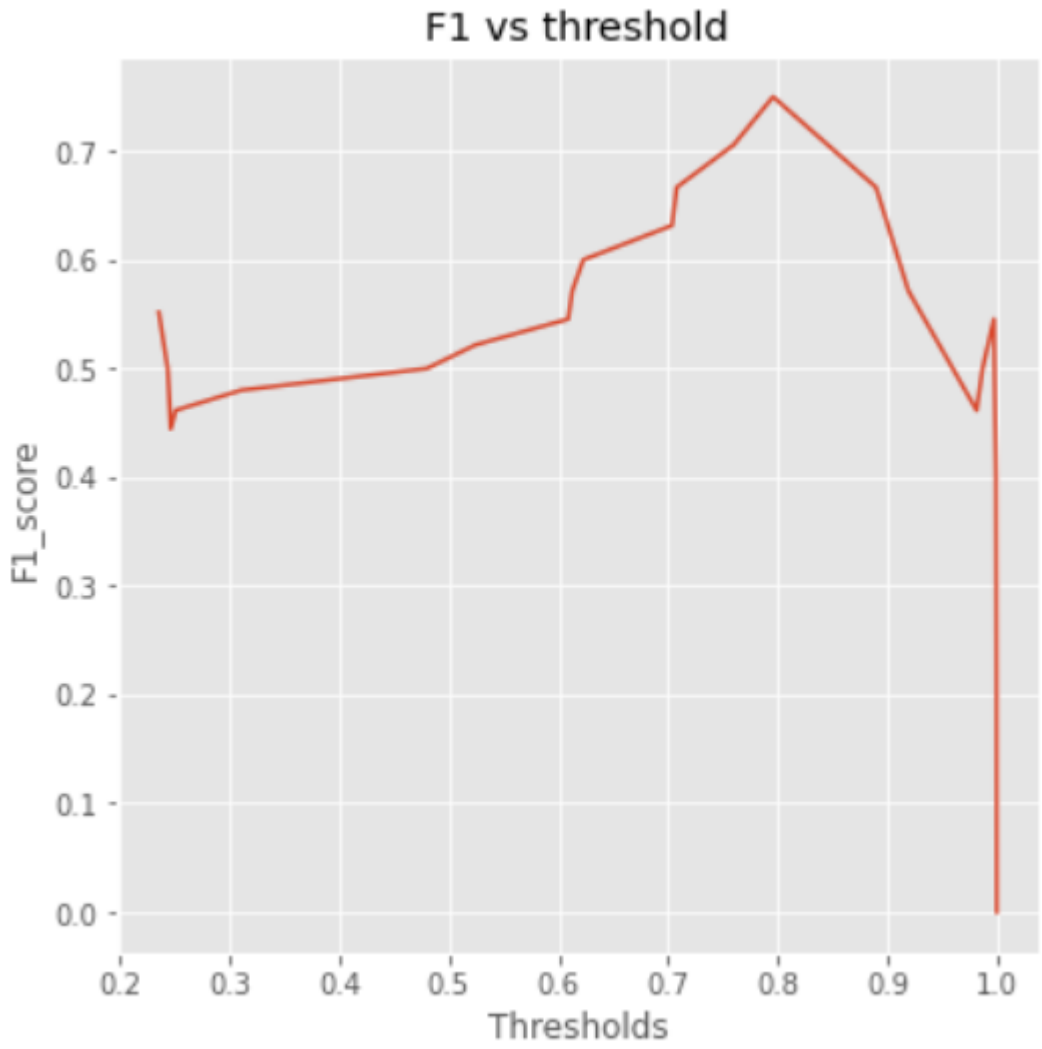
P-R Curve


precision_recall_curve
Pneumonia (AP Score:0.73)
Precision
Recall
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.0
0.2
0.4
0.6
0.8
1.0

AUC Curve


AUC curve
Pneumonia (AUC:0.83)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0

Final Threshold and Explanation:



I chose F1 score as the metrics for this problem because depending on the scenario in which the algorithm is used, both precision and recall can be important factors. Hence, to balance both of them, this metric is used.

The threshold calculated with the best F1 score is 0.7959541

4. Databases

We used a part of the National Institutes of Health Chest X-Ray Dataset. The NIH Chest X-ray Dataset is comprised of 112,120 X-ray images with disease labels from 30,805 unique patients. To create these labels, the authors used Natural Language Processing to text-mine disease classifications from the associated radiological reports. The labels are expected to be >90% accurate and suitable for weakly-supervised learning. The original radiology reports are not publicly available but you

can find more details on the labelling process in this Open Access paper: "ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases."

We have 1431 cases from 112120 cases which have Pneumonia as a finding. That is 1.2763110952550838% of whole dataset. We will distribute these into training and validation sets accordingly

Description of Training

Dataset:

In [6]:
```
print(train_data['Pneumonia'].sum()/len(train_data), len(train_data))
train_data.head()
```

```
0.5 2576
```

Out[6]:

	Image Index	Finding Labels	Follow-up #	Patient ID	Patient Age	Patient Gender	View Position	OriginalImage[Width	Height]	OriginalImagePixelSpacing[x	...	Fibros
60478	00014933_007.png	Nodule	7	14933	55	F	AP	2500	2048	0.168	...	0
87073	00021489_001.png	Infiltration	1	21489	33	M	PA	3056	2544	0.139	...	0
44691	00011488_000.png	No Finding	0	11488	50	M	PA	2500	2048	0.168	...	0
11494	00003027_001.png	Atelectasis|Infiltration	1	3027	58	M	PA	2500	2048	0.171	...	0
13549	00003523_015.png	Infiltration|Pneumonia	15	3523	23	F	AP	2500	2048	0.168	...	0

5 rows × 29 columns

We used 2576 images for training set with 50 percent of these having Pneumonia as a finding.

Description of Validation

Dataset:

In [7]:
```
print(valid_data['Pneumonia'].sum()/len(valid_data), len(valid_data))
valid_data.head()
```

```
0.2 715
```

Out[7]:

	Image Index	Finding Labels	Follow-up #	Patient ID	Patient Age	Patient Gender	View Position	OriginalImage[Width	Height]	OriginalImagePixelSpacing[x	...	F
49687	00012616_003.png	Infiltration	3	12616	73	M	PA	2874	2991	0.143000	...	
4761	00001278_000.png	No Finding	0	1278	50	F	AP	2048	2500	0.168000	...	
99287	00026259_000.png	No Finding	0	26259	57	F	PA	2770	2738	0.143000	...	
45781	00011723_008.png	Pleural_Thickening	8	11723	62	M	AP	2500	2048	0.168000	...	
79272	00019464_004.png	Effusion|Pleural_Thickening	4	19464	40	F	PA	1775	2022	0.194311	...	

5 rows × 29 columns

We used 715 validation images, 20 percent of these having Pneumonia which is similar to prevalence of Pneumonia in clinical setting.

5. Ground Truth

Here, we used NLP extracted Radiology report as labelling.

Data limitations:

The image labels are NLP extracted so there could be some erroneous labels but the NLP labelling accuracy is estimated to be >90%. Very limited numbers of disease region bounding boxes (See BBoxlist2017.csv)

Chest x-ray radiology reports are not anticipated to be publicly shared. Parties who use this public dataset are encouraged to share their "updated" image labels and/or new bounding boxes in their own studied later, maybe through manual annotation

6. FDA Validation Plan

Approach FORTIS hospital to be our clinical partner.

Patient Population Description for FDA Validation Dataset: Need access to DICOM data containing Chest X-rays of patients of age 20 to 80 in both AP and PA orientations.

Ground Truth Acquisition Methodology: Assign a team of 2 or 3 radiologists (silver standard) to label the validation data.

Algorithm Performance Standard: F1 score is the metric of choice because we can have multiple scenarios for this device where precision or recall is important respectively.

Aside from comparing with our silver standard we also want to achieve F1 score better than the average F1 score(0.387) of group of 4 radiologists as observed in literature.

Note: Increase in precision decreases recall. We may want higher precision or recall according to the scenario.

Eg. If we are using the device for screening and want minimum false negatives, then we need higher sensitivity(recall). That is no patient who has a disease should be marked negative.

Increasing recall will decrease precision. But if we let precision drop too much, this will result increase in false positives and thus cases of lower interest will be given more priority while being queued for the radiologist to look at.

Here we can choose corresponding threshold with higher precision from Precision-Recall curve or we can use F1 score that balances both precision and recall.