FDA Submission

Your Name: Joao Diogo de Oliveira

Name of your Device: Auxiliary Pneumonia Detection

Algorithm Description

1. General Information

Intended Use Statement: The device is intended to help Radiologists screen Patients for Pneumonia disease by analysing chest x-rays;

Indications for Use: The device is intended to be used as a tool to help Radiologists , but never as a ground truth decider. It is meant to be used with DICOM files, tested with patients between 1 and 95 years old, of Chest X-Rays in 2 view-points (AP or PA), for a distribution of slightly more men than woman (54% vs 46%).

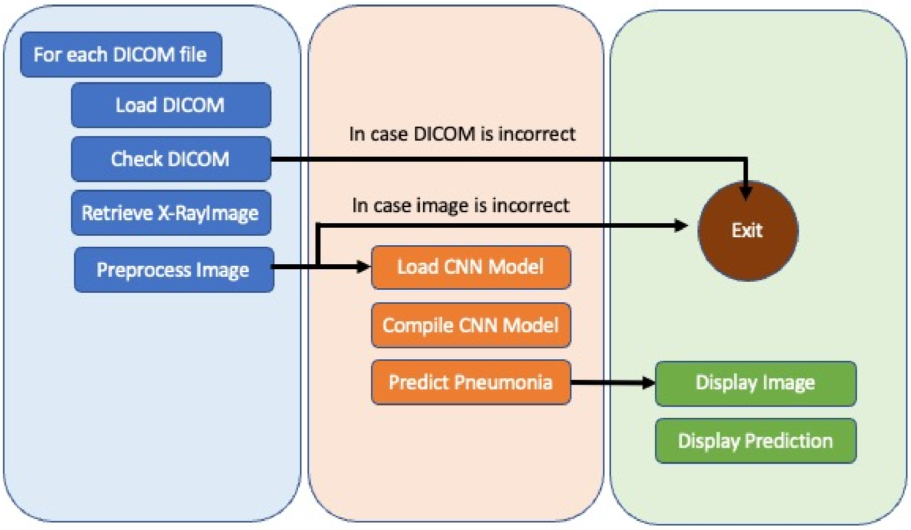
Device Limitations: It is known that Pneumonia can exist together with other diseases, namely Edema and Consolidation are the ones with the highest correlation (from our test). So that is a clear problem, which makes it very hard for the model to distinguish between Pneumonia and other diseases, has the diseases have similar distribution. Other Diseases linked with Pneumonia are Effusion, Atelectasis, Infiltration and Nodule, meaning that they will make it hard to distinguish between them.
Meaning that we could see False Positives in Patients with above diseases as well as False Negatives.

Clinical Impact of Performance:

- Our Precision for No-Pneumonia is: 86%
- Our Recall for No-Pneumonia is: 53%
- Our Precision for Pneumonia is: 19%
- Our Recall for Pneumonia is: 57%

As we can see our Recall is higher than our Precision in general, so it means less False Negative, which means that when the model predicts that the X-Ray doesn't have Pneumonia, then there is a high change that indeed it doesn't have (but not impossible, so all X-rays still need to be revised).

2. Algorithm Design and Function



DICOM Checking Steps: There are few checks:

1. Check if Modality of X-Ray is DX in Dicom, otherwise it stops;
2. Check if Age is between 1-95 years old. In case it isn't it throws a warning but still displays image and predicts;
3. Check if body part is Chest otherwise it stops;
4. Check if Position is either PA or AP, otherwise it stops;

Preprocessing Steps: The original image values, are:

1. First normalized, so from pixels between 0-255 we divide by 255 them in order to get a normalized distribution of 0-1
2. We resize the image from the original file size to the 224x224 pixels;

CNN Architecture:

It was tested in the VGG16 and ResNet50. The model which performed best was based on

the VGG16

We freezed the first 17 layers for training (given we pushed the pretrained model VGG16 with the imagenet weights), and added 3 extra Linear layers followed by Dropout layers and 1 Output Layer:

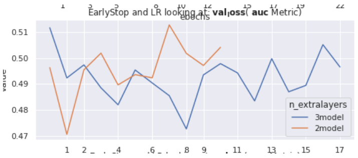
- Linear Layer of 1024 tensors with ReLu activation
- Followed by a Dropout layer of 0.5
- Linear Layer of 512 tensors with ReLu activation
- Followed by a Dropout layer of 0.5
- Linear Layer of 256 tensors with ReLu activation
- Followed by a Dropout layer of 0.5
- Output Layer

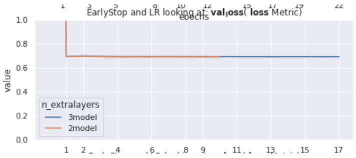
3. Algorithm Training

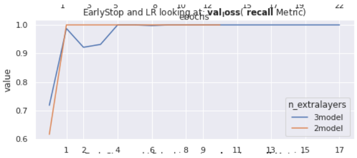
We used 100 epochs to train our model, however using a callback with a early stop of patience 10. On top of that, we experienced with multiple combinations of pretrained models and layer setup.

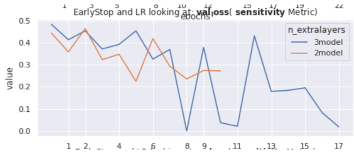
Parameters:

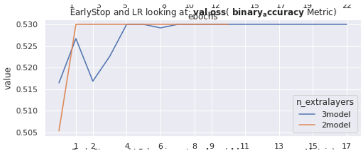
- Types of augmentation used during training
  - horizontal_flip = hor_flip
  - height_shift_range = 0.1
  - width_shift_range = 0.1
  - rotation_range = 20
  - shear_range = 0.1
  - zoom_range = 0.1
- Batch size: 32
- Optimizer learning rate: start with a high value (0.01) and going down to 0.000001 (with a linear scheduler)
- Layers of pre-existing architecture that were frozen: all initial 17 layers frozen (as mentioned above)
- Layers of pre-existing architecture that were fine-tuned: None
- Layers added to pre-existing architecture; added 3 extra Linear layers followed by Dropout layers and 1 Output Layer:
  - Linear Layer of 1024 tensors with ReLu activation
  - Followed by a Dropout layer of 0.5
  - Linear Layer of 512 tensors with ReLu activation
  - Followed by a Dropout layer of 0.5
  - Linear Layer of 256 tensors with ReLu activation
  - Followed by a Dropout layer of 0.5
  - Output Layer

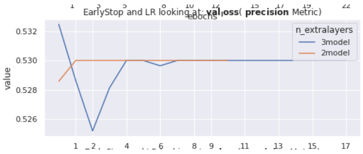


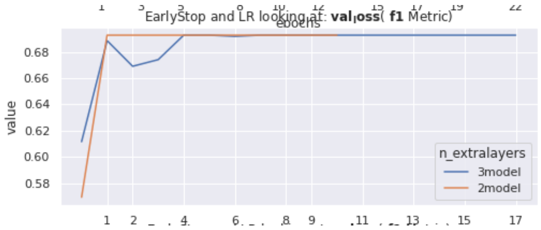


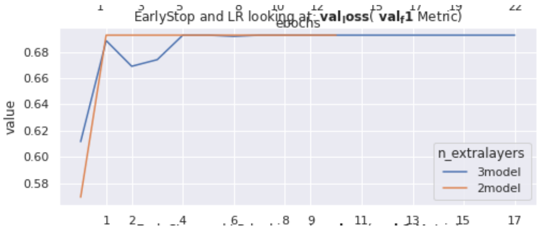


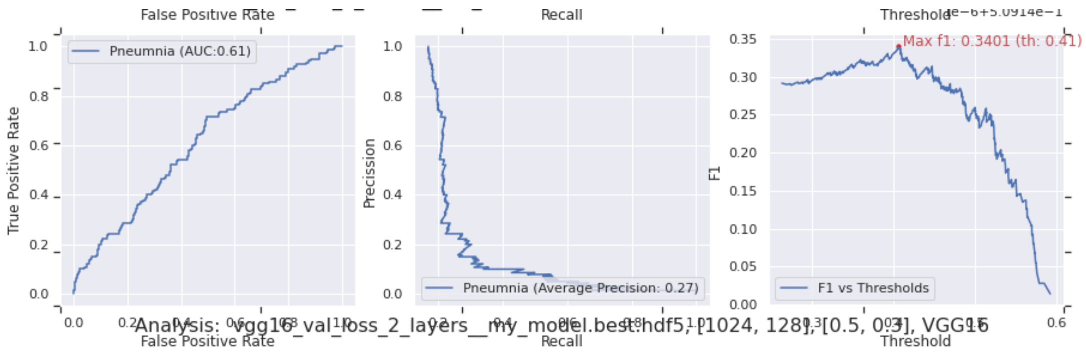












Final Threshold and Explanation: As you can see from the image above, the threshold choosen was 0.41 given it yelds the best F1 score.

This was the best threshold which gives the balance between Recall and Precision. So a more complete model.

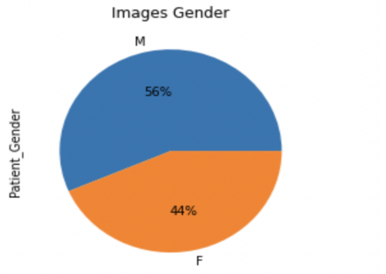
4. Databases

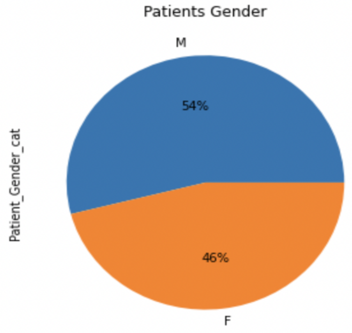
(For the below, include visualizations as they are useful and relevant)

Description of General Dataset:

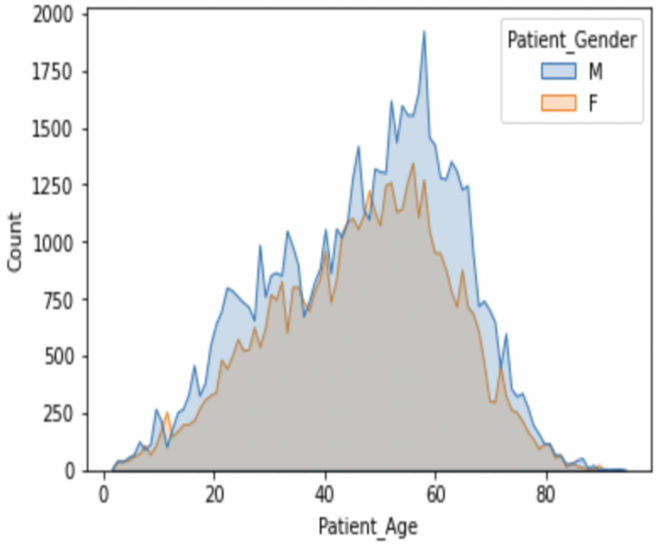
The general Dataset where data was taken from is composed by:

- Total X-rays: 111,863
- Total number of patients: 30,805
- Male vs Female Ratio: 56% vs 44% (x-rays); 54% vs 46% (patients);

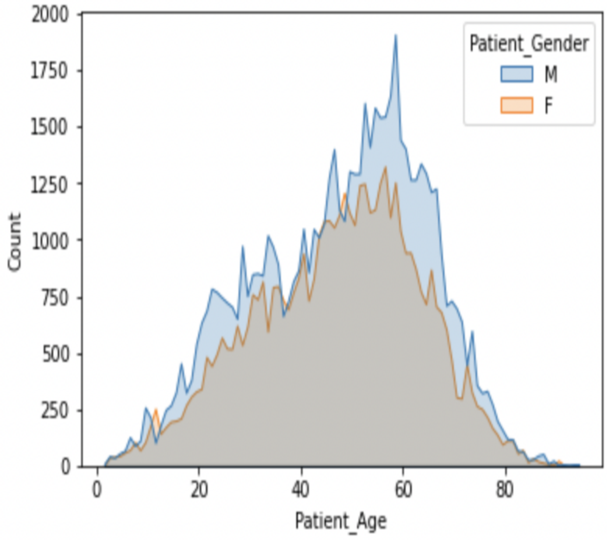




- The distribution between Gender and Age is as follows:

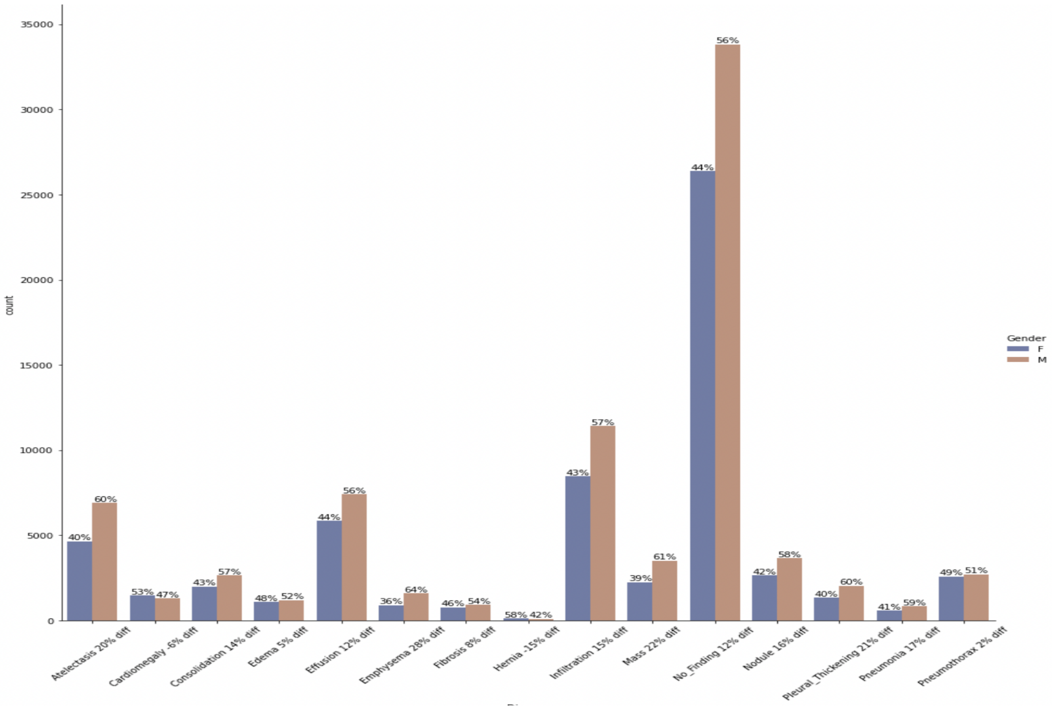


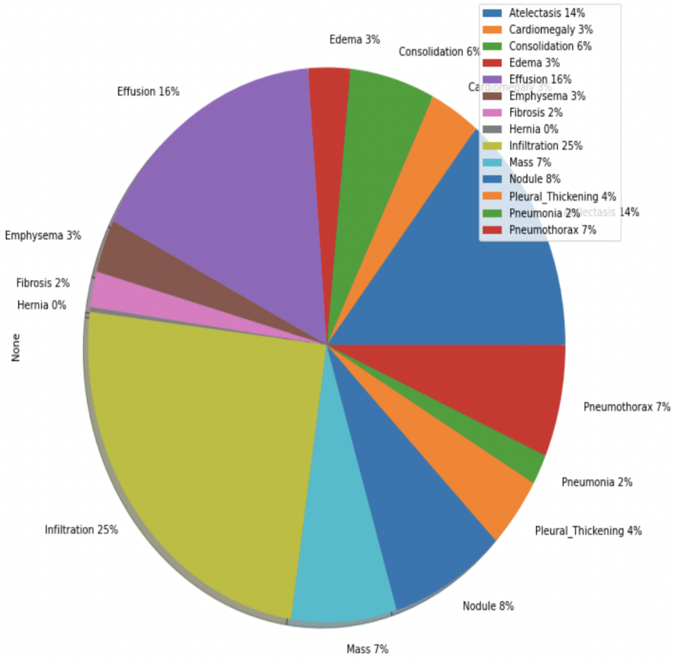
Historigram of Age according to Sex (no-Pneumonia)

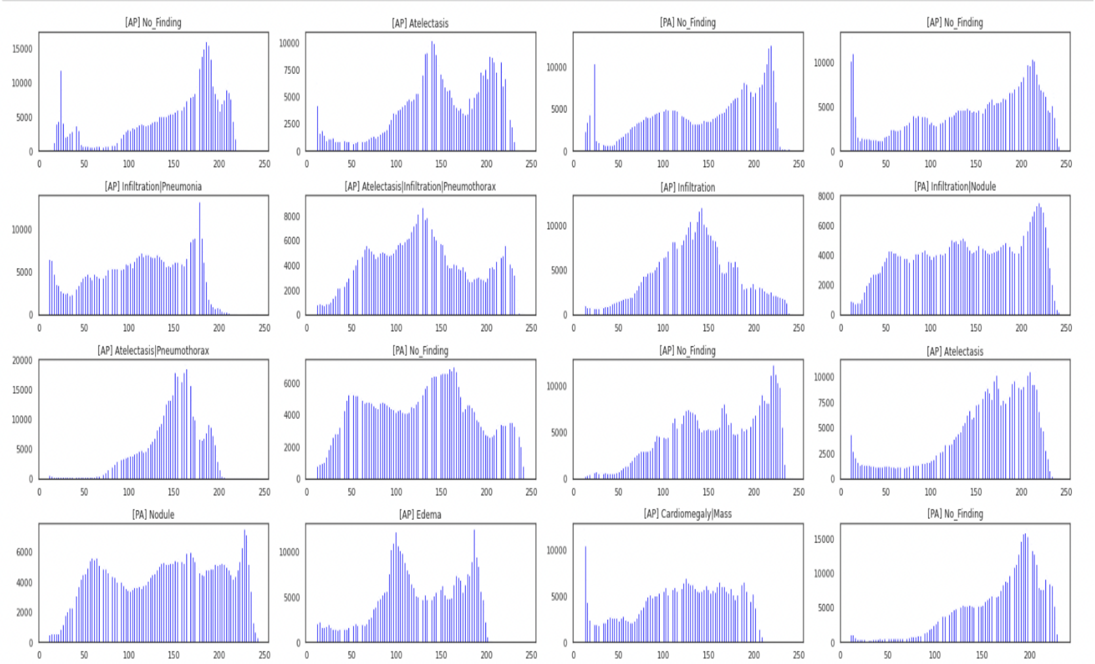


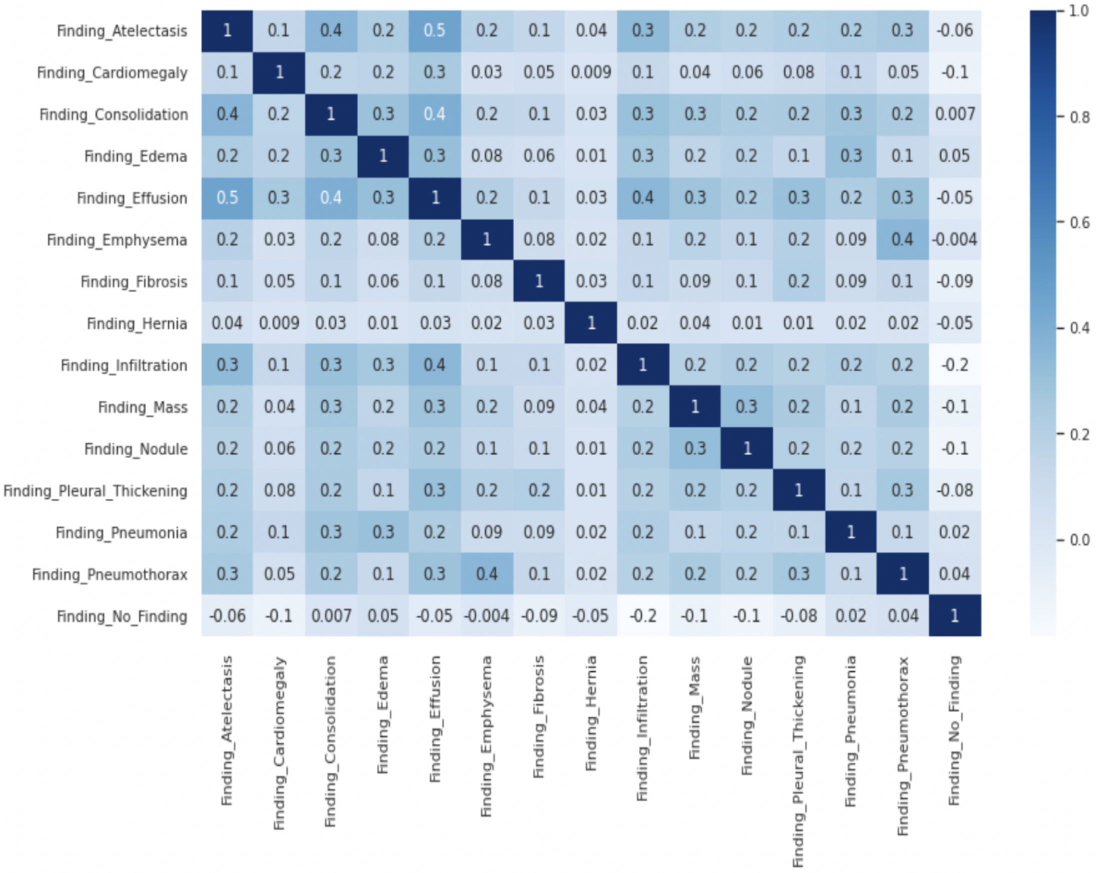
Historigram of Age according to Sex (Pneumonia)

- Regarding the Diseases: Dataset has 14 diseases with the following distribution and correlation


35000
30000
25000
20000
15000
10000
5000
0
count
Gender
F
M
40%
60%
53%
47%
43%
57%
48%
52%
44%
56%
36%
64%
46%
54%
58%
42%
43%
57%
39%
61%
44%
56%
42%
58%
40%
60%
41%
59%
49%
51%
Atelectasis 20% diff
Cardiomegaly -6% diff
Consolidation 14% diff
Edema 5% diff
Effusion 12% diff
Emphysema 28% diff
Fibrosis 8% diff
Hernia -15% diff
Infiltration 15% diff
Mass 22% diff
No_Finding 12% diff
Nodule 16% diff
Pleural_Thickening 21% diff
Pneumonia 17% diff
Pneumothorax 2% diff


Atelectasis 14%
Cardiomegaly 3%
Consolidation 6%
Edema 3%
Effusion 16%
Emphysema 3%
Fibrosis 2%
Hernia 0%
Infiltration 25%
Mass 7%
Nodule 8%
Pleural_Thickening 4%
Pneumonia 2%
Pneumothorax 7%
Edema 3%
Consolidation 6%
Cardiomegaly 3%
Effusion 16%
Atelectasis 14%
Emphysema 3%
Fibrosis 2%
Hernia 0%
None
Pneumothorax 7%
Pneumonia 2%
Pleural_Thickening 4%
Infiltration 25%
Nodule 8%
Mass 7%



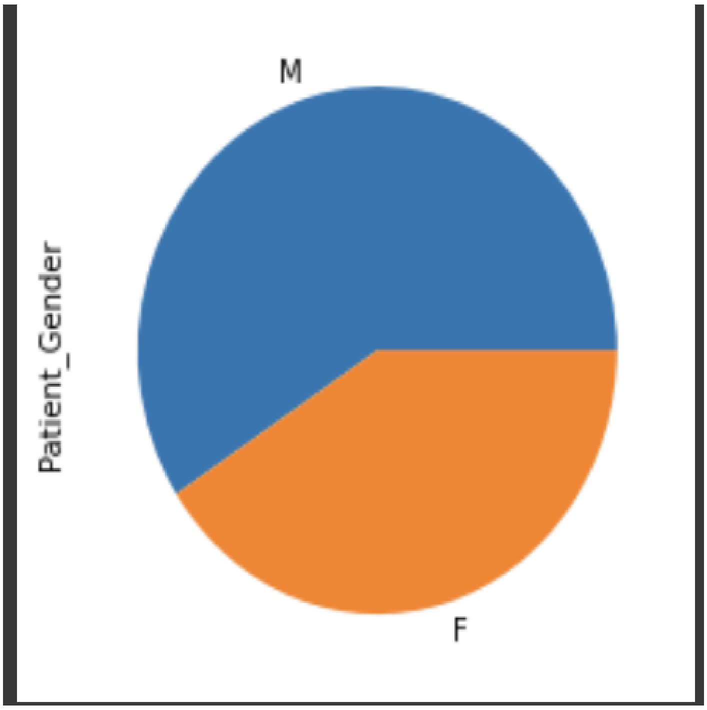


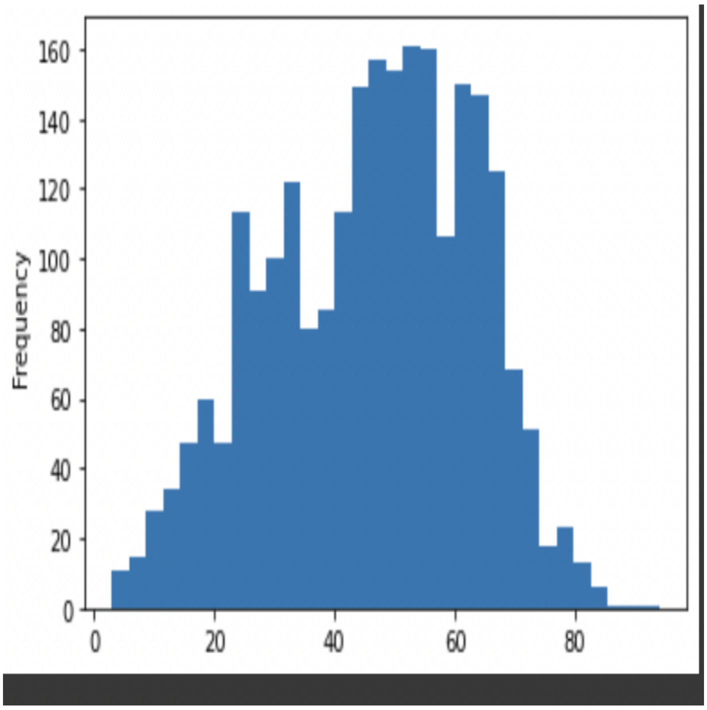
	Finding_Atelectasis	Finding_Cardiomegaly	Finding_Consolidation	Finding_Edema	Finding_Effusion	Finding_Emphysema	Finding_Fibrosis	Finding_Hernia	Finding_Infiltration	Finding_Mass	Finding_Nodule	Finding_Pleural_Thickening	Finding_Pneumonia	Finding_Pneumothorax	Finding_No_Finding
Finding_Atelectasis	1	0.1	0.4	0.2	0.5	0.2	0.1	0.04	0.3	0.2	0.2	0.2	0.2	0.3	-0.06
Finding_Cardiomegaly	0.1	1	0.2	0.2	0.3	0.03	0.05	0.009	0.1	0.04	0.06	0.08	0.1	0.05	-0.1
Finding_Consolidation	0.4	0.2	1	0.3	0.4	0.2	0.1	0.03	0.3	0.3	0.2	0.2	0.3	0.2	0.007
Finding_Edema	0.2	0.2	0.3	1	0.3	0.08	0.06	0.01	0.3	0.2	0.2	0.1	0.3	0.1	0.05
Finding_Effusion	0.5	0.3	0.4	0.3	1	0.2	0.1	0.03	0.4	0.3	0.2	0.3	0.2	0.3	-0.05
Finding_Emphysema	0.2	0.03	0.2	0.08	0.2	1	0.08	0.02	0.1	0.2	0.1	0.2	0.09	0.4	-0.004
Finding_Fibrosis	0.1	0.05	0.1	0.06	0.1	0.08	1	0.03	0.1	0.09	0.1	0.2	0.09	0.1	-0.09
Finding_Hernia	0.04	0.009	0.03	0.01	0.03	0.02	0.03	1	0.02	0.04	0.01	0.01	0.02	0.02	-0.05
Finding_Infiltration	0.3	0.1	0.3	0.3	0.4	0.1	0.1	0.02	1	0.2	0.2	0.2	0.2	0.2	-0.2
Finding_Mass	0.2	0.04	0.3	0.2	0.3	0.2	0.09	0.04	0.2	1	0.3	0.2	0.1	0.2	-0.1
Finding_Nodule	0.2	0.06	0.2	0.2	0.2	0.1	0.1	0.01	0.2	0.3	1	0.2	0.2	0.2	-0.1
Finding_Pleural_Thickening	0.2	0.08	0.2	0.1	0.3	0.2	0.2	0.01	0.2	0.2	0.2	1	0.1	0.3	-0.08
Finding_Pneumonia	0.2	0.1	0.3	0.3	0.2	0.09	0.09	0.02	0.2	0.1	0.2	0.1	1	0.1	0.02
Finding_Pneumothorax	0.3	0.05	0.2	0.1	0.3	0.4	0.1	0.02	0.2	0.2	0.2	0.3	0.1	1	0.04
Finding_No_Finding	-0.06	-0.1	0.007	0.05	-0.05	-0.004	-0.09	-0.05	-0.2	-0.1	-0.1	-0.08	0.02	0.04	1

Description of Training Dataset:

For the training Dataset we needed to make sure that the classification was balanced, which means we made sure we had a 50-50% split between images with Pneumonia and without Pneumonia. While doing that, we tried to keep the same proportions of Gender, Age, View-Position, ...

> *Train db: 2437 total elements ; 1294 pneumonia; 1143 ; 53.10% pneumonia*



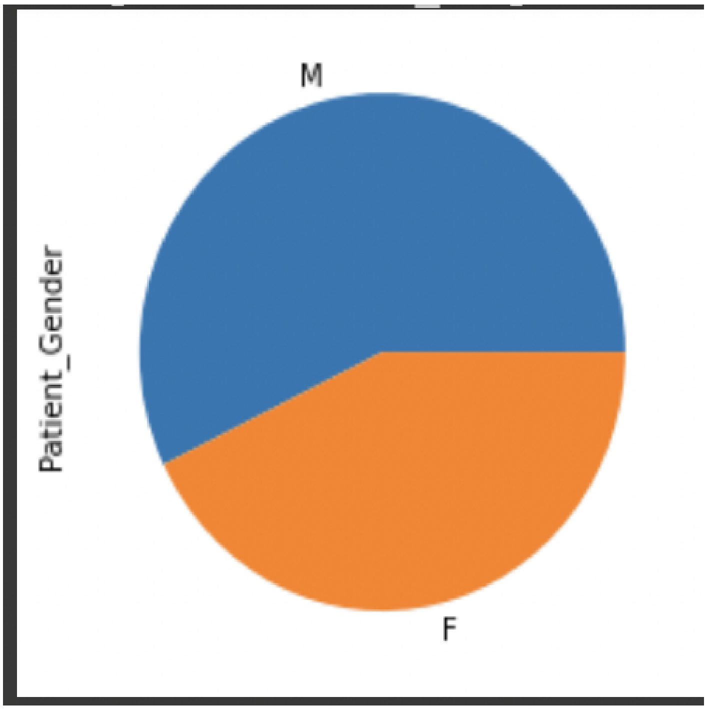


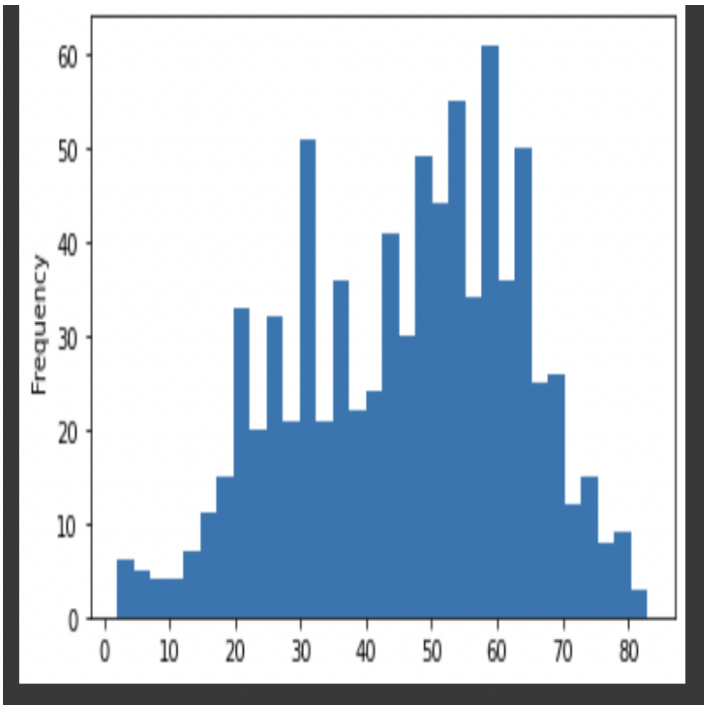
Description of Validation Dataset:

For the validation set, we took 20% of patients with Pneumonia and then multiplied by about 5 times to get No Pneumonia cases.

Another import aspect which I worried **was Data Leakeage**, so made sure that the Patient ID in training data wasn't in the Validation Set.

> *Valid db: 810 total elements ; 135 pneumonia; 675 ; 16.67% pneumonia*





5. Ground Truth

For this experiment we used labels generated by NLP models, so there could be some erroneous labels but the NLP labeling accuracy is estimated to be >90%.

6. FDA Validation Plan

Patient Population Description for FDA Validation Dataset:

Any population between 1 and 95 year old, of any gender.

Ground Truth Acquisition Methodology:

The golden standard for Ground Truth for this case would be for a weighted average of Radiologists (Sputum test) going through every X-ray doing there own rating.

We know that X-Ray analysis is extremely hard so even experience Radiologists can make errors. Therefore to mitigate it, idealy we would have multiple Radiologists doing the rating and we would weight it in according to the radiologists experience.

The for obtaining ground truth would be to perform one of these tests (see this Mayo Clinic Link):

Algorithm Performance Standard:

One of the best papers in the topic is "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning" which states:

> *We assess the performance of both radiologists and CheXNet on the test set for the pneumonia detection task. Recall that for each of the images in the test set, we have 4 labels from four practicing radiologists and 1 label from CheXNet. We compute the F1 score for each individual radiologist and for CheXNet against each of the other 4 labels as ground truth. We report the mean of the 4 resulting F1 scores for each radiologist and for CheXNet, along with the average F1 across the radiologists. We use the bootstrap to construct 95% bootstrap confidence intervals (CIs), calculating the average F1 score for both the radiologists and CheXNet on 10,000 bootstrap samples, sampled with replacement from the test set. We take the 2.5th and 97.5th percentiles of the F1 scores as the 95% bootstrap CI. We find that CheXNet achieves an F1 score of 0.435 (95% CI 0.387, 0.481), higher than the radiologist average of 0.387 (95% CI 0.330, 0.442).*