Sara
Name of your Device: DAPP-X (Device for analysis of patients with pneumonia - X)

Algorithm Description

1. General Information

Intended use

This algorithm is intented for use on Pneumonia patients (male and female) from ages of 1-100 with suspected lung disease who underwent chest X-rays and may or may not have previous chest X-ray data.

Indication for use

The algoritgm could be used for screening chest X-ray data from male and female patients from ages of 1-100 in the early detection of Pneumonia.

Algorithm limitations

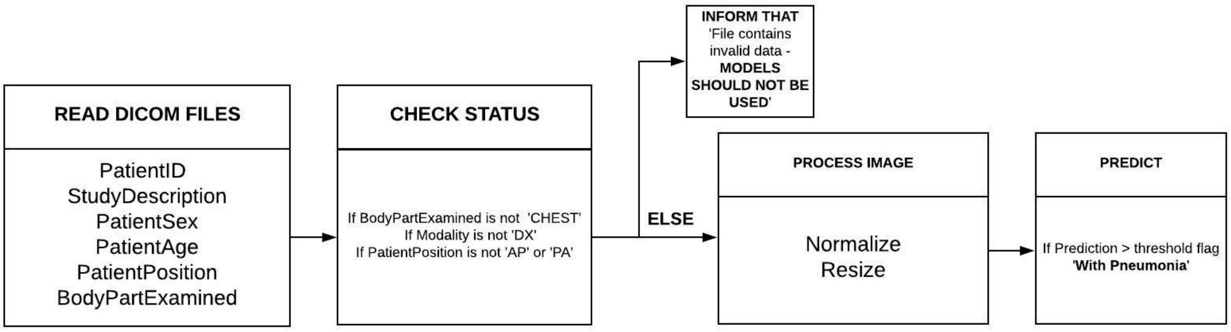
As the trainning dataset have many other diseases and it's very biases (low information on pneumonia patients), it performs somehow poorly on the accurate detection of pneumonia due to the presence of other diseases. In addition, some of the images are rotated or zoomed_out and it is not ideal to process all of then using the same algorithms.

Clinical Impact of Performance

In this specific case, a false positive is not só problematic. If a normal person is said to have pneumonia, other tests should be required (including blood and other chest x-rays). However, a false negative is a major issue as the patient might be sent home and be sick. This is why we need to focus on these results.

2. Algorithm Design and Function

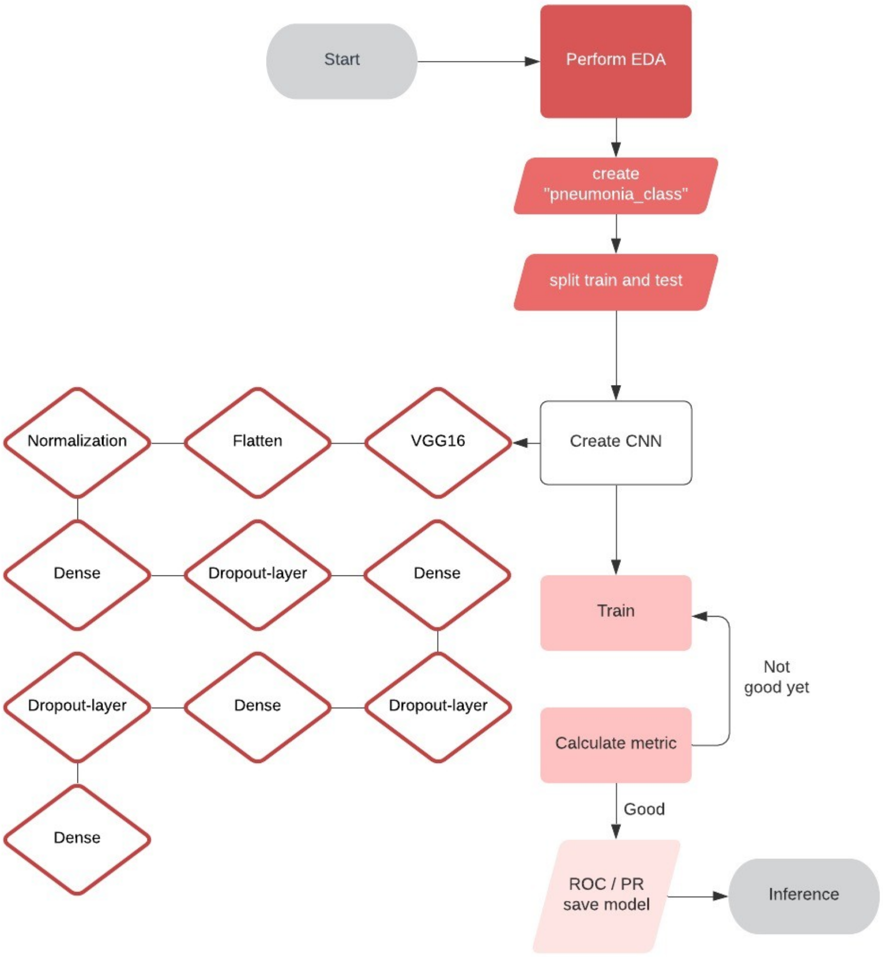
The algorithm works by reading the DICOM images and identifying the image parameters. If the image is not of the chest, or is not taken in PA or PX position or is not DX modality, the algorithm will inform that the model should not be used (although it will still be calculated, for later analysis). The images are then treated based on the average (normalized) and also resized. The weights and model are loaded and forecasts are made. If the prediction is above the threshold, then Pneumonia is suspected.

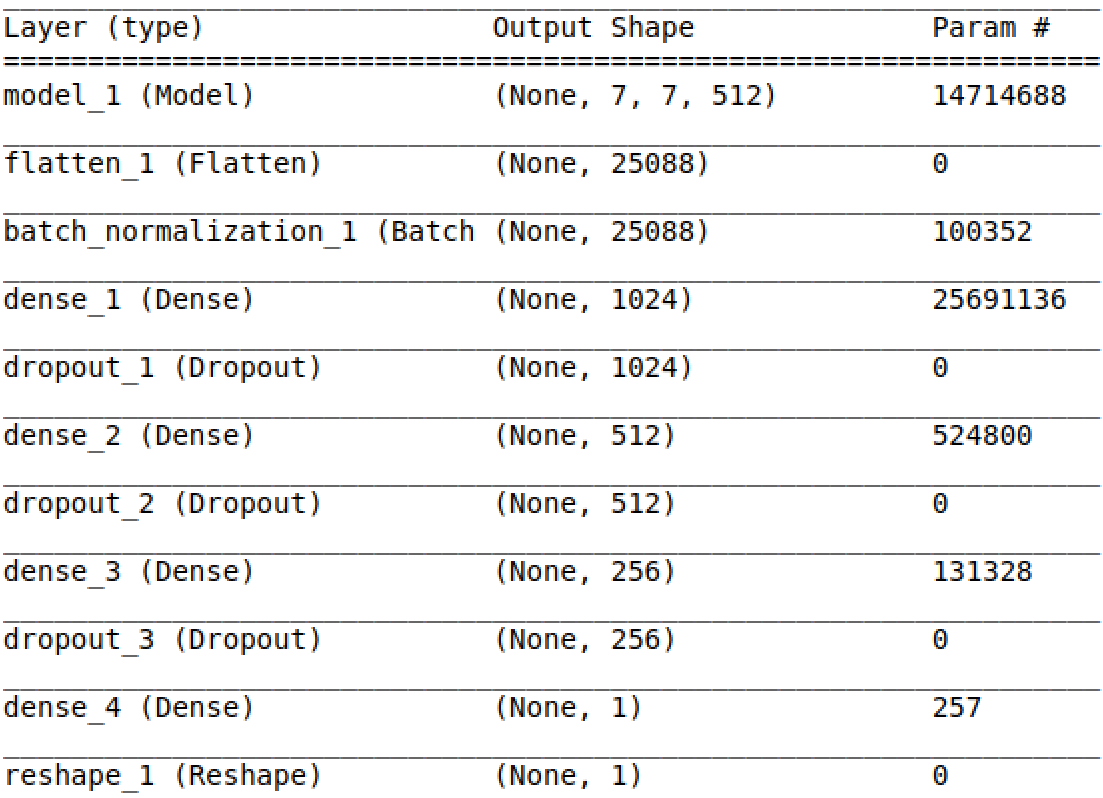


READ DICOM FILES
PatientID
StudyDescription
PatientSex
PatientAge
PatientPosition
BodyPartExamined
CHECK STATUS
If BodyPartExamined is not 'CHEST'
If Modality is not 'DX'
If PatientPosition is not 'AP' or 'PA'
INFORM THAT 'File contains invalid data - MODELS SHOULD NOT BE USED'
ELSE
PROCESS IMAGE
Normalize
Resize
PREDICT
If Prediction > threshold flag 'With Pneumonia'

The algorithm is designed based on the pretrained vgg16 model in a sequential maner. Each layer was added considering these specific needs:

1. Add VGG16 Convolutional layer
2. Flatten the output of the VGG16 model because it is from a convolutional layer
3. Add a dropout-layer which may prevent overfitting and improve generalization ability to unseen data e.g. the test-set.
4. Add a dense (fully-connected) layer. This is for combining features that the VGG16 model has recognized in the image.
5. Add a dropout-layer which may prevent overfitting and improve generalization ability to unseen data e.g. the test-set.
6. Add a dense (fully-connected) layer. This is for combining features that the VGG16 model has recognized in the image.
7. Add a dropout-layer which may prevent overfitting and improve generalization ability to unseen data e.g. the test-set.
8. Add a dense (aka. fully-connected) layer.
9. Add a dropout-layer which may prevent overfitting and improve generalization ability to unseen data e.g. the test-set.
10. Add a dense (aka. fully-connected) layer.





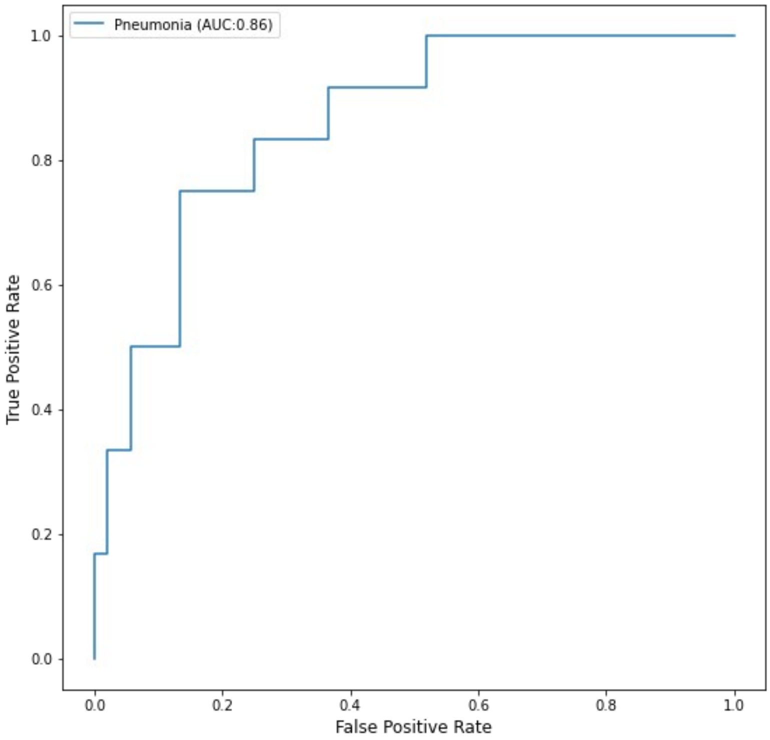
```
_________________________________________________________________
Layer (type)                 Output Shape              Param #   
=================================================================
model_1 (Model)              (None, 7, 7, 512)         14714688  
_________________________________________________________________
flatten_1 (Flatten)          (None, 25088)             0         
_________________________________________________________________
batch_normalization_1 (Batch (None, 25088)             100352    
_________________________________________________________________
dense_1 (Dense)              (None, 1024)              25691136  
_________________________________________________________________
dropout_1 (Dropout)          (None, 1024)              0         
_________________________________________________________________
dense_2 (Dense)              (None, 512)               524800    
_________________________________________________________________
dropout_2 (Dropout)          (None, 512)               0         
_________________________________________________________________
dense_3 (Dense)              (None, 256)               131328    
_________________________________________________________________
dropout_3 (Dropout)          (None, 256)               0         
_________________________________________________________________
dense_4 (Dense)              (None, 1)                 257       
_________________________________________________________________
reshape_1 (Reshape)          (None, 1)                 0         
```

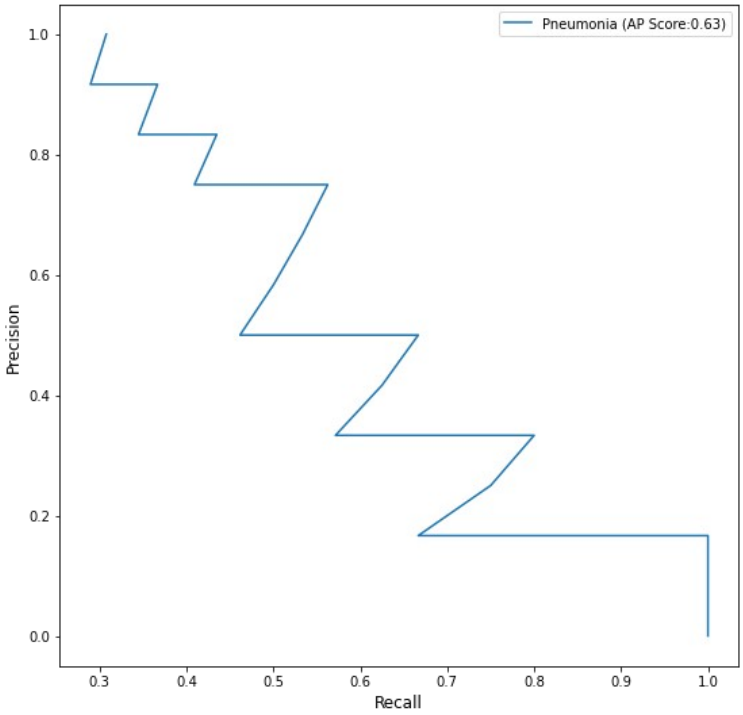
3. Algorithm Training

For training data augmentation the following procedures were conducted:

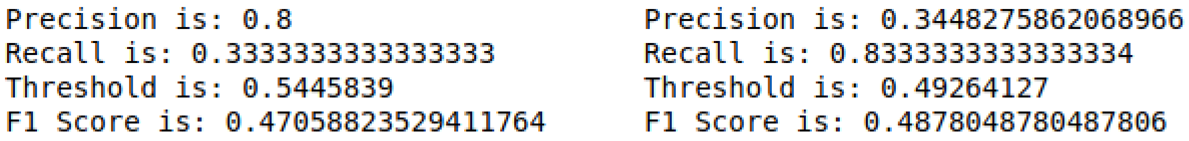
- Image rescaled to 1/255;
- horizontal flip;
- height shift range of 0.05;
- width shift range of 0.1;
- rotation range of 20 (slightly rotation);
- shear range of 0.1;
- zoom range of 0.1 (slightly zoomed);

For the training image generator, a batch size of 32 images is used and for valid images 64. The algorithm is trained based on Adam optimizer and a learning rate of 0.001. The current model has very low training loss,cval_ loss and val_acc.





The algorithm has an area under the curve for True positive rate and false positive rate of 0.86 and the precision-recall curve has an AP score of 0.63. From the above, the calculated threshold is 0.5445839 for the precision of 0.8 and the F1 score of ~0.47. For the recall of 0.8, the threshold is 0.49264127 with F1 score 0.49. We are using the average of both thresholds as final threshold.



```
Precision is: 0.8
Recall is: 0.3333333333333333
Threshold is: 0.5445839
F1 Score is: 0.47058823529411764
```

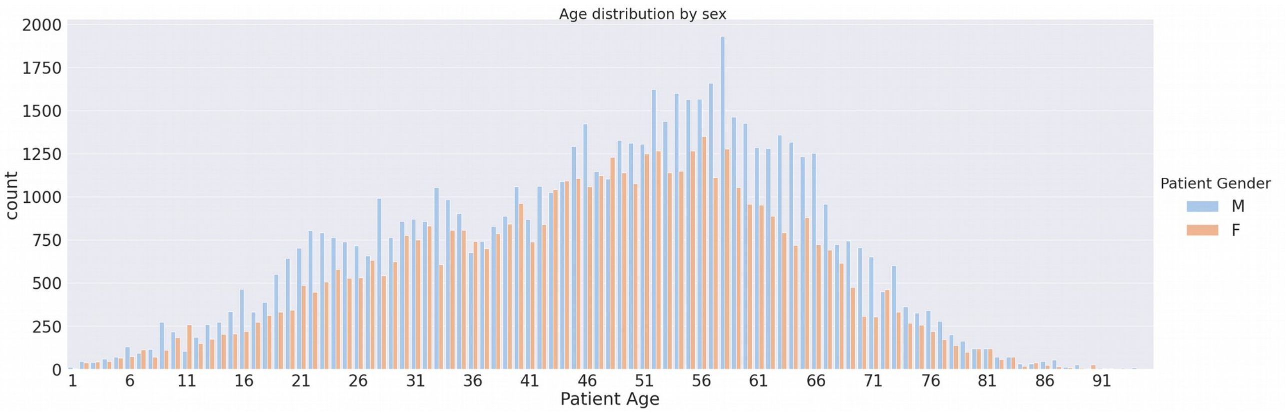
```
Precision is: 0.3448275862068966
Recall is: 0.8333333333333334
Threshold is: 0.49264127
F1 Score is: 0.4878048780487806
```

4. Databases

The dataset was curated by the NIH and have 112,120 frontal-view chest X-ray images from 30,805 unique patiens. The disease labels were created using NLP to mone the associated radiological reports. Based on this dataset, some EDA can be done to define bias. We tried to answered some questions bellow.

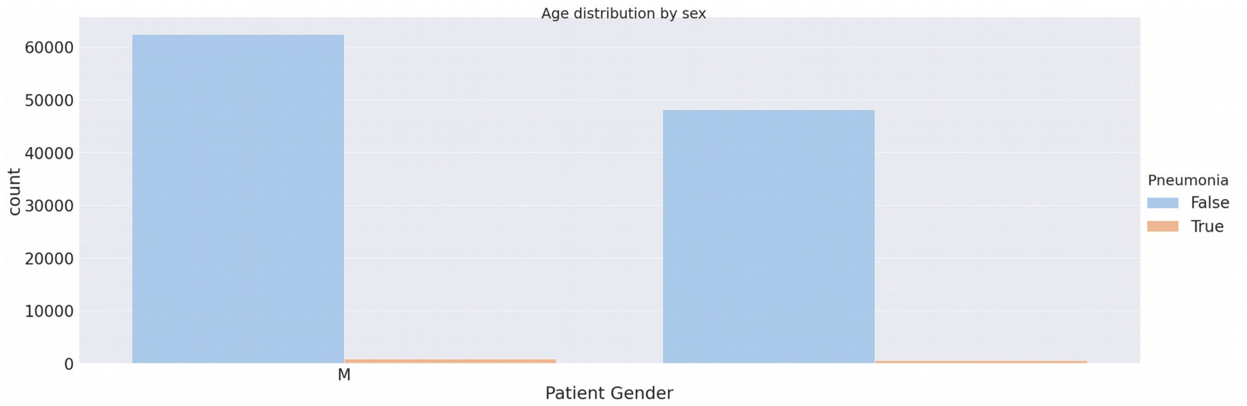
For training the model, I have split the dataset into 80% training data and a 20% dataset. Then the training data is balanced for 50% pneumonia cases and 50% non-pneumonia cases. The training dataset is augmented as described in the previous section then the training data frame is generated with the target size of 224 x 224 and a batch size of 32. Similarly, the validation dataset is generated as the training dataset but with image augmentation ie just rescaling to 1/255.

What is the gender and age distribution?



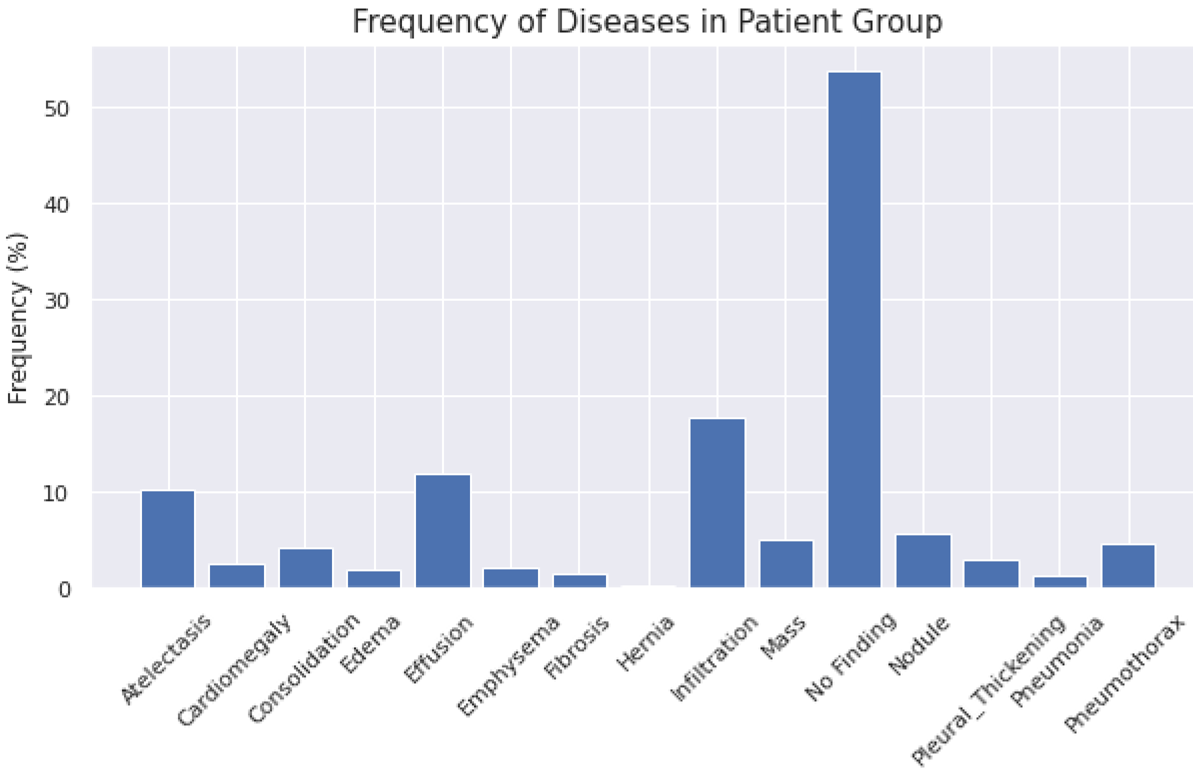
Apparently,there are more men than women in our dataset. In addition, the distribution is more normal for women based on age, while for men it’s skewed, having more data from older men.

What is the gender distribution for Pneumonia patients?



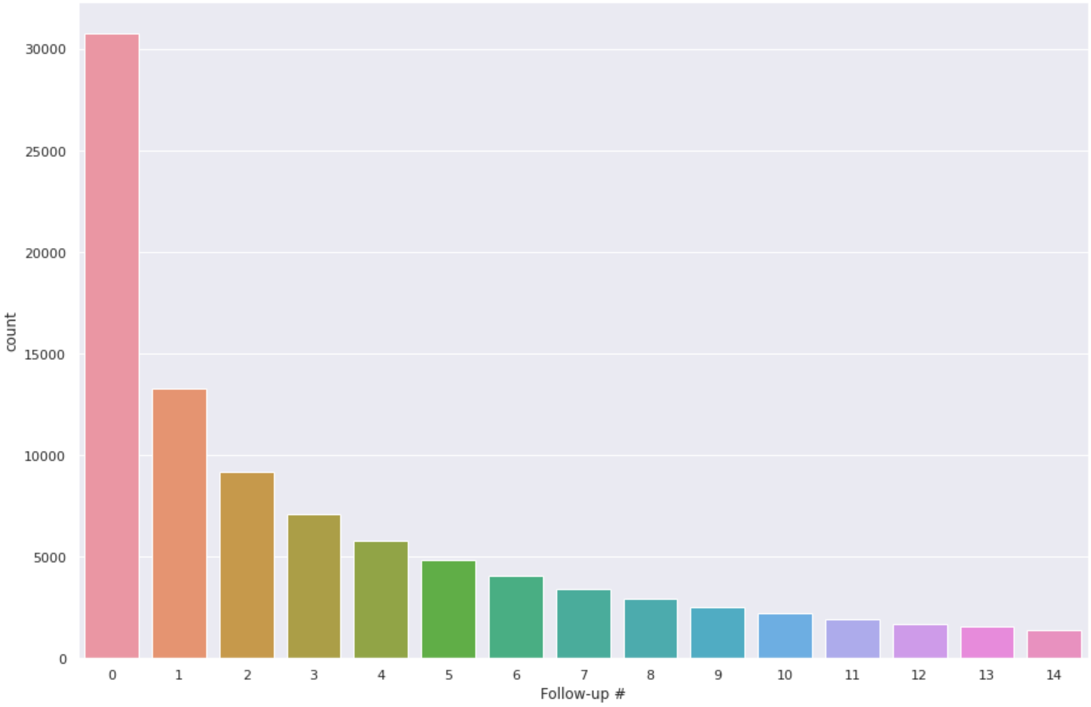
As the amount of men in the dataset is larger, so is the amount of male patients with Pneumonia.

What is the frequency of the diseases?



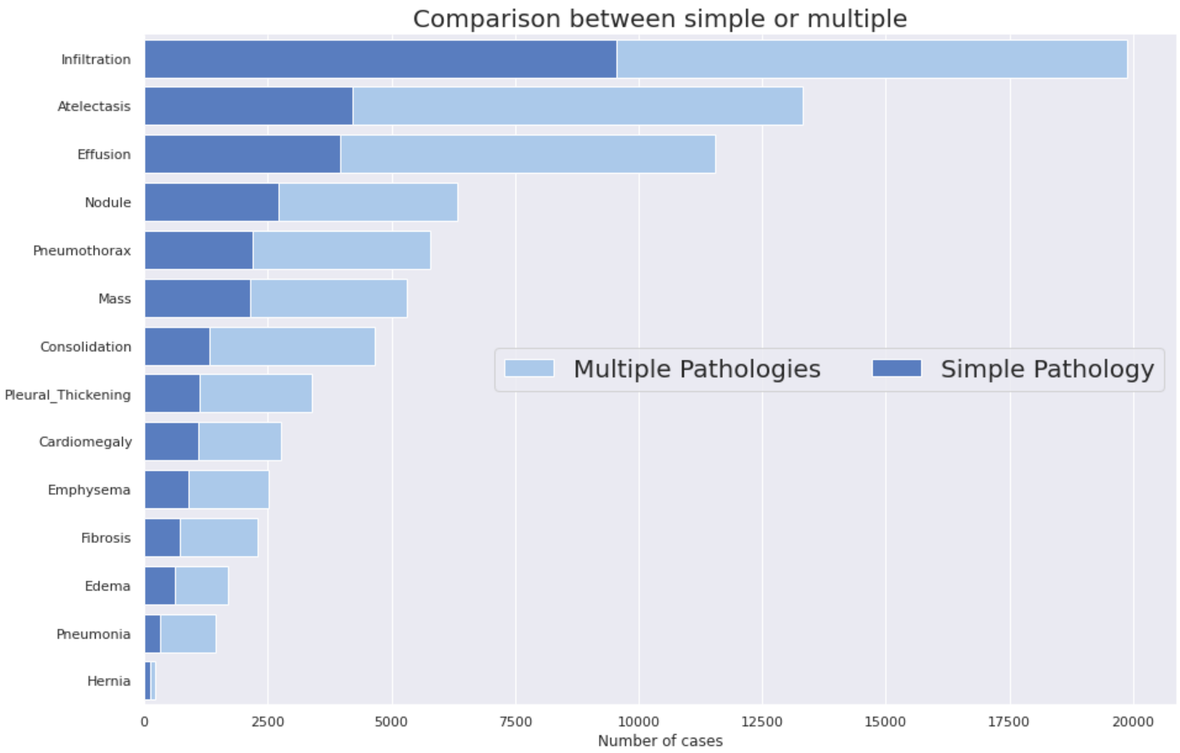
Checking the frequency of each disease (considering patients with more than one disease also), we see that many of then (more than 50%) has no apparent problems. After that, there are many cases of infiltration, atelectasis and effusion. Less than 3% of then have Pneumonia.

How many images are follow-ups?



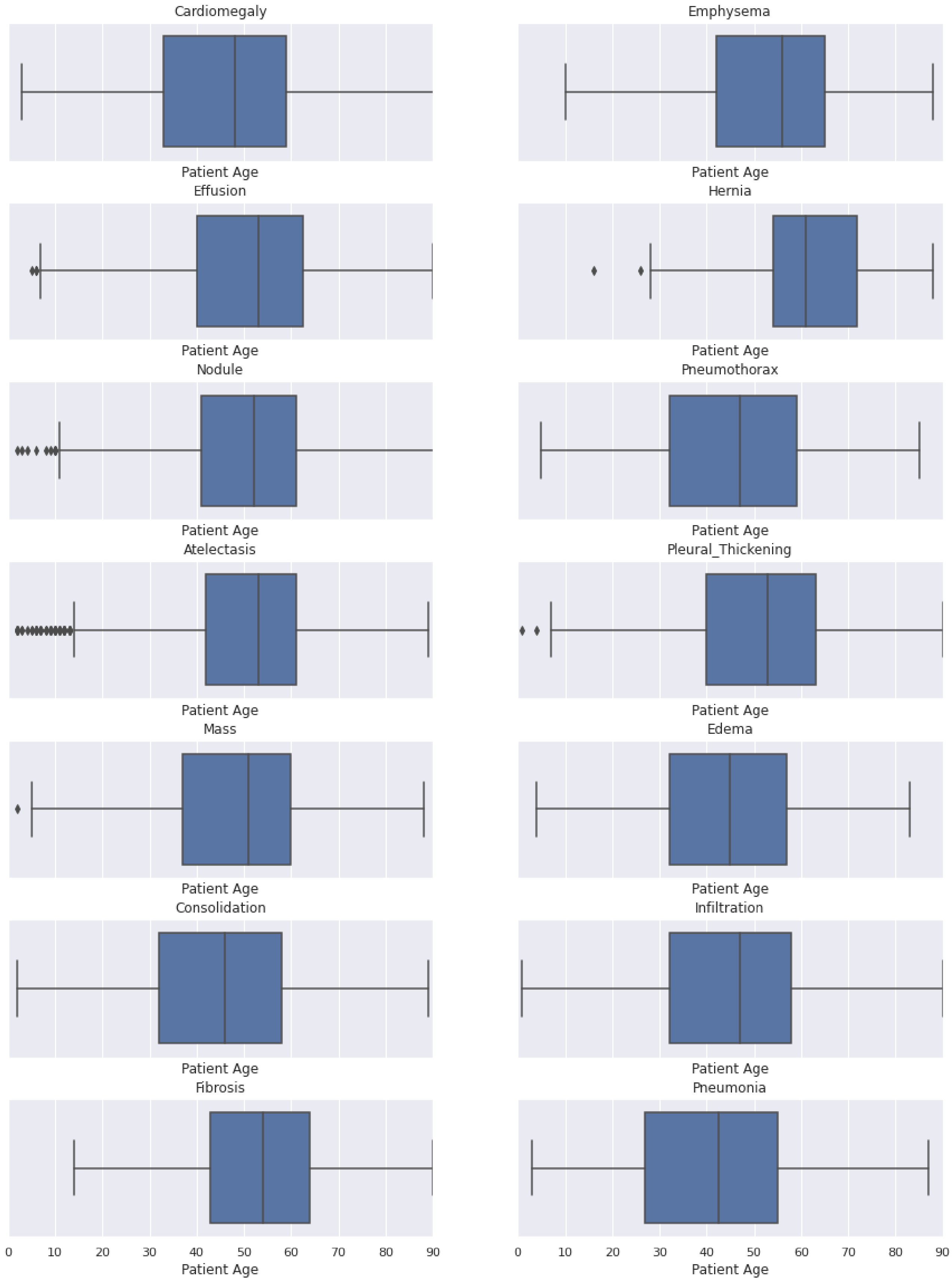
Most images are the first x-ray taken, but some of then are follow-ups, going up to more than 15.

What is the frequency of diseases when occuring alone and on multiple diagnostics?



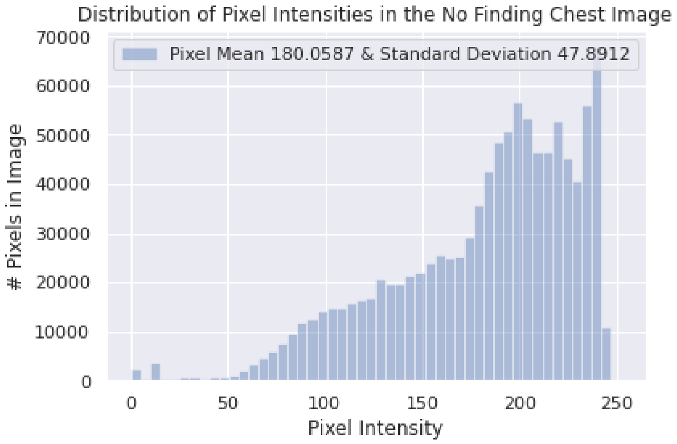
Many of the diseases don’t appear alone. We investigated what is the frequency of diseases when happening alone or not. We can see that infiltration (the disease with higher frequency) wins, followed by atelectasis and effusion. Pneumonia usually appear with other pathologies.

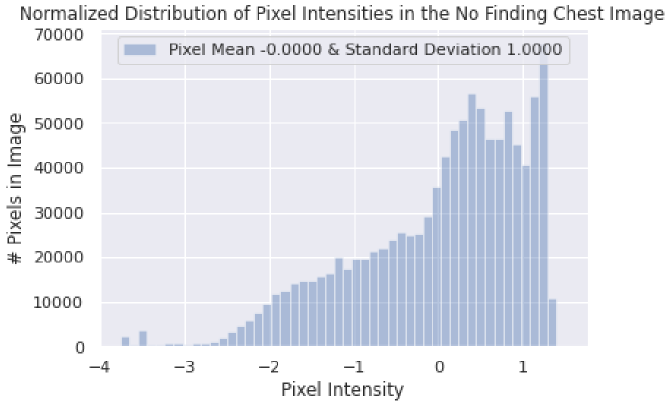
What is the distribution of patients' age based on specific diseases?



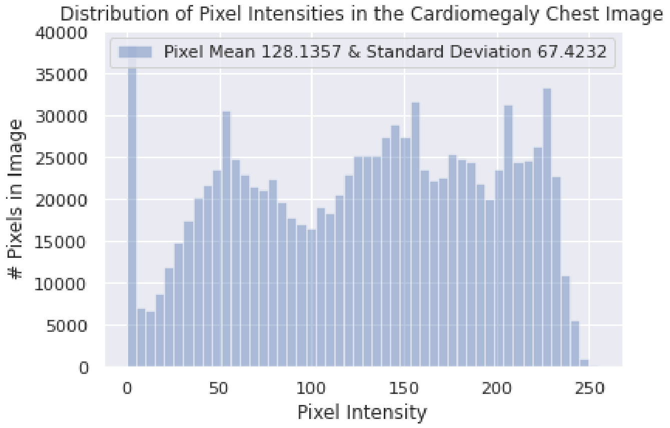
We also checked the boxplots of patients’ age and pathology. Hernia, for example, also happens on older people (with a few outliers), while pneumonia has the younger patients

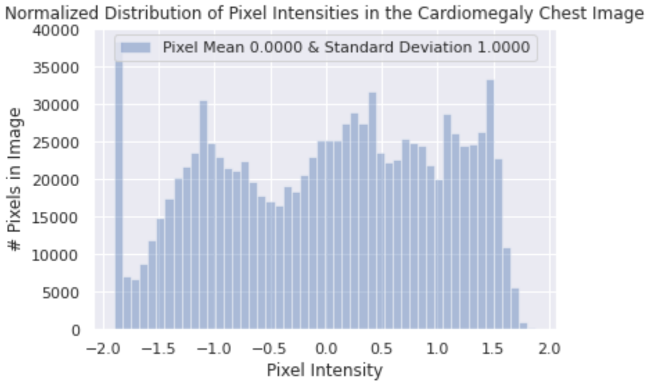
What is the distribution of pixels of No Finding X-ray?



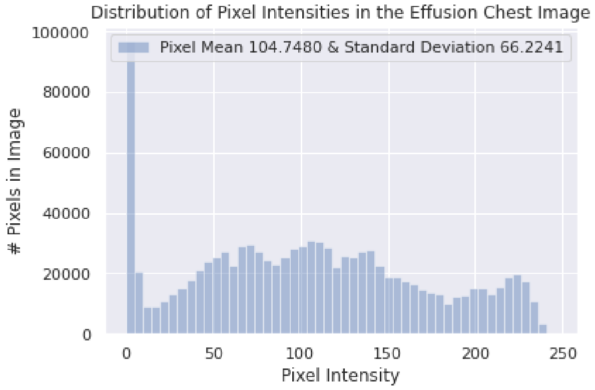


What is the distribution of pixels of Cardiomegaly image?

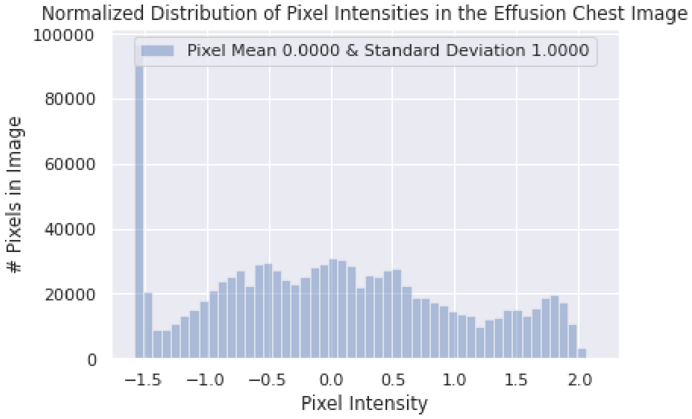




What is the distribution of pixels of Effusion X-ray?



5.



We can see from the pixel distribution that the intensity of the pixel determines if we have a skewed distribution or not, which indicates the type of pathology.

Ground Truth

A considerable effort has been devoted to creating large-scale x-ray datasets with more reliable ground truth. The problem with this and any other dataset is the that the image labels were NLP-extracted so there could be errors (accuracy > 90%).

In addition, only frontal radiographs were presented to the radiologists, but it has been shown that some of the diagnoses require lateral view. In addition, radiologists were not given old results to comparison.

6. FDA Validation Plan

Patient Population Description: For validation of the algorithm, the collected dataset should contain frontal view chest X-rays of patients with Pneumonia between the ages 1-100 for both male and female. These dataset should be very accurate and several radiologists should be used to identifty the pathology. Aditional tests could also be used (Silver standard). Other diseases that could be included (but are not going to be identified) are Atelectasis, Cardiomegaly, Consolidation, Edema, Effusion, Emphysema, Fibrosis, Hernia, Infiltration, Mass, Nodule, Pleural Thickening, or Pneumothorax. Radiologists should confirm the diagnoses.

Ground Truth Acquisition Methodology: This algorithm is good to confirm a diagnostic (due to high precision). In this condition, the ground truth should be a dataset labeled by experience radiologists. They should check for the high ratio of false positives and identify other diseases.

Algorithm Performance Standard: Considering the paper *“CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rayswith Deep Learning”*, we know that radiologists F1-score have a wide range, ranging from 0.282 to 0.492 with an average of 0.387. So several radiologists considering a silver approach would be more optimal as performance standard.