FDA Submission

Name: Susan Kolim

Name of your Device: Software that can predict the presence of pneumonia with human radiologist-level accuracy.

Algorithm Description

1. General Information

Intended Use Statement: Assisting a radiologist with identifying the presence of pneumonia.

Indications for Use: Indicated for use in screening pneumonia studies in male and female between ages 1 to 95 who have been administered a screening pneumonia study through digital Chest X-Ray (DX) with view position either AP or PA.

Device Limitations: From the Exploratory Data Analysis (EDA), it is found the most common comorbidities are infiltration (13.92%) and edema & infiltration (9.58%). Also, it is found that infiltration, edema, and pneumonia Chest-X-Ray almost have a similar intensity values distribution. Therefore, there is a high chance that this algorithm performs very poorly on the accurate detection of pneumonia in the presence of infiltration and edema.

Clinical Impact of Performance: It is found that the best threshold is 0.5226 with precision 0.7143, recall 0.8333, and F1-score 0.7692 (see Build and Train Model 3.ipynb).

This means that of *all Positive test results*, only 71% is truly Positive. About 29% (100% – 71%) is actually False Positive (FP). This means that 29% of patients with no pneumonia have to wait to be seen by doctor although they don't have pneumonia.

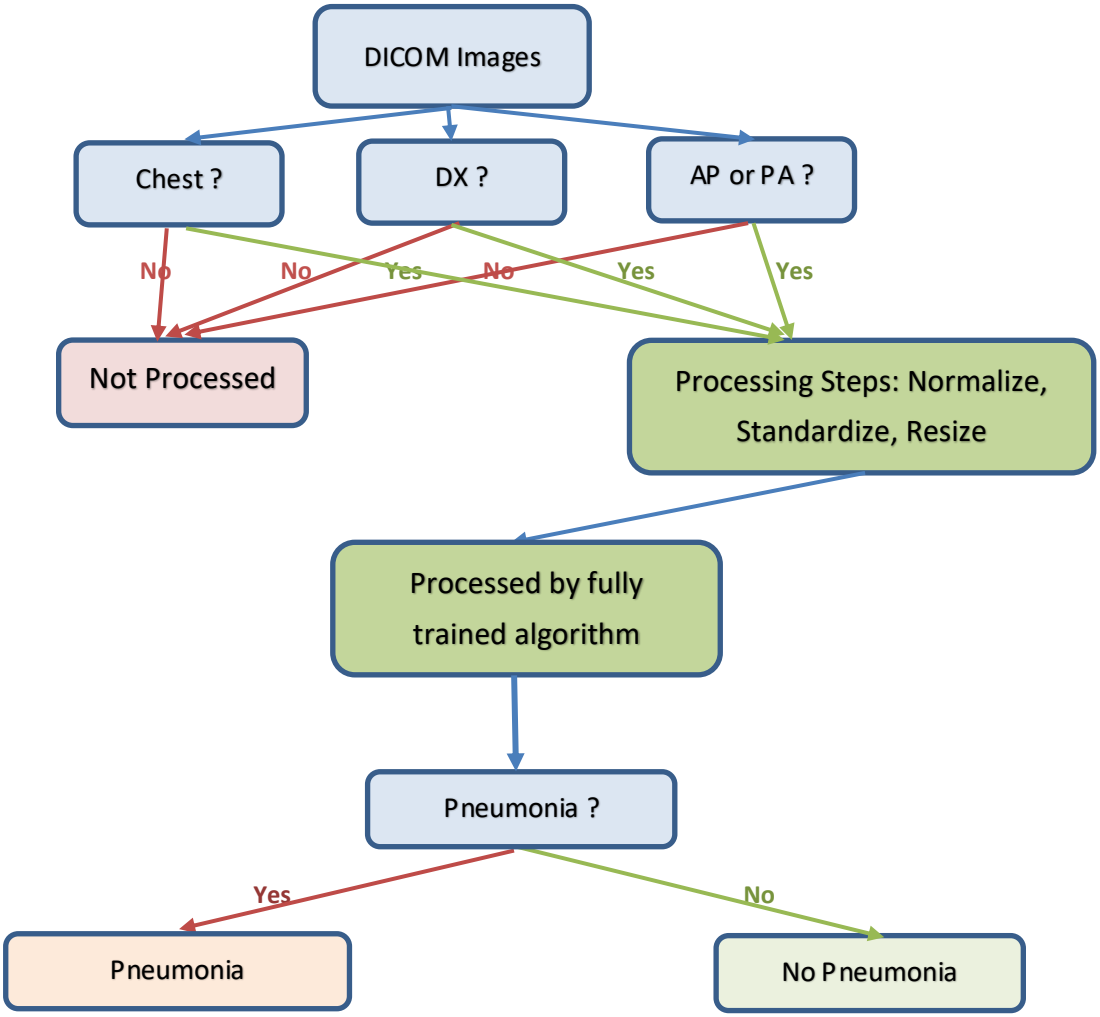
Moreover, from *all actual Positive cases*, only 83% will be tested positive with pneumonia, while 17% (100% - 83%) will be tested as negative. This means that 17% of pneumonia patients will be discharged when they actually need to see doctors. This can have serious impact to patient's health. We call this 17% as False Negative (FN).

Since this model have lower FN than FP, it is more appropriate for screening of pneumonia where we really want to make sure if someone doesn't have a pneumonia, or worklist prioritization where we want to make sure that people without pneumonia are being deprioritized. But we still need to keep in mind of the FN since it can have serious impact to patient's health.

2. Algorithm Design and Function

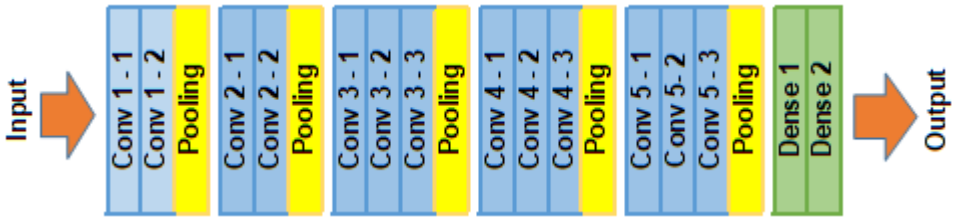
DICOM Checking Steps: Before DICOM images were processed by the algorithm, we performed a checking by accessing DICOM data which showed Body Part Examined, Modality, and Patient Position. If the DICOM data didn't indicate that Body Part Examined is 'Chest', Modality is 'DX', and Patient Position is 'AP' of "PA', we would not process the image in that DICOM.

The flowchart below show how DICOM images enter our fully trained algorithm.



Preprocessing Steps: At the beginning, the images are rescaled or normalized and standardized. For the normalization, we divide the image intensities with 255. As for standardization, we subtract the image by the image mean, then divide it by the image standard deviation. Afterward, we resize the image so it become 224 x 224.

CNN Architecture: Below is the algorithm flowchart. The images enter a pre-trained VGG16 model (Convolution 1 – 1 to Convolution 5 – 3), i.e., we freeze all layers except the final convolutional layer of VGG16, then we fine-tune the model by adding a dense layer of 256 units with RELU activation (Dense 1) before the final dense layer of 1 unit with Sigmoid activation (Dense 2) which outputs the result.

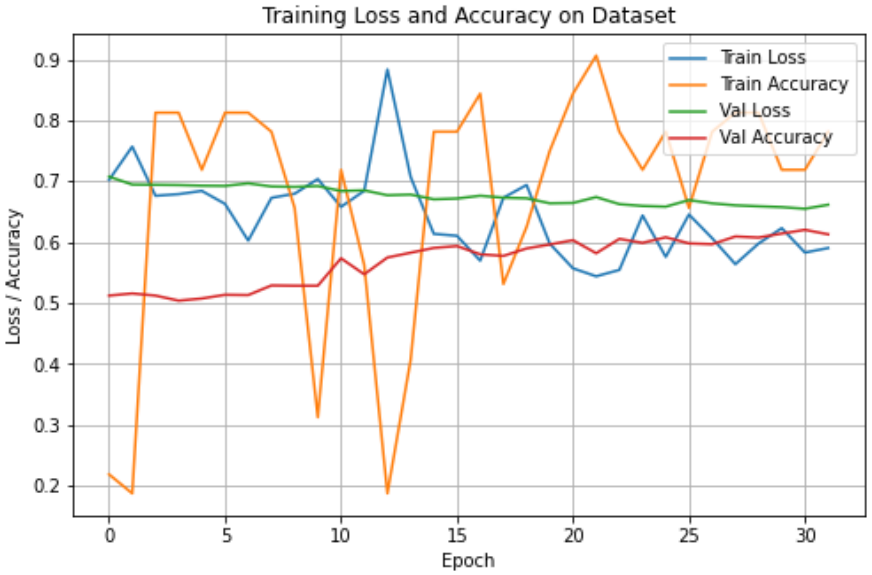


3. Algorithm Training

Parameters:

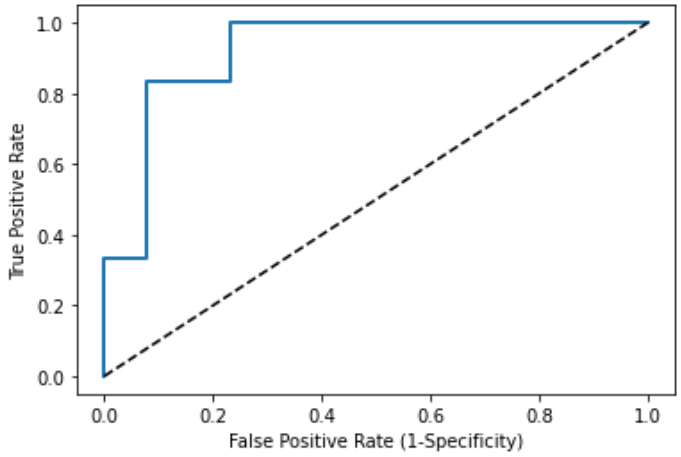
- *Types of augmentation used during training:* Only images in the training set were augmented. The augmentation which we did were horizontal_flip, height_shift_range = 0.1, width_shift_range=0.1, rotation_range=10, shear_range = 0.1, zoom_range=0.1, brightness_range=[0.8,1.2]. These augmentation were chosen since they are often seen in clinical setting.
- *Batch size:* The batch size for the training set is 64 and the one for the validation set is 32.
- *Optimizer learning rate:* The optimizer used is Adam with learning rate 0.0001.
- *Layers of pre-existing architecture that were frozen:* All layers in the VGG16 were frozen except the final convolutional layer. To be precise, the layers of VGG16 which was frozen was **Block 1**: Convolutional 1 and Convolutional 2, **Block 2**: Convolutional 1 and Convolutional 2, **Block 3**: Convolutional 1, Convolutional 2, and Convolutional 3, **Block 4**: Convolutional 1, Convolutional 2, and Convolutional 3, **Block 5**: Convolutional 1, Convolutional 2, and Convolutional 3
- *Layers of pre-existing architecture that were fine-tuned:* **Block 5**: Polling
- *Layers added to pre-existing architecture:* **Dense 1**, layer of 256 units with RELU activation and **Dense 2**, a layer of 1 unit with Sigmoid activation

Below is the plot of algorithm training performance. It shows that the validation loss (green) decreases from epoch 0 to 22, then it does not improve or decrease anymore. This means that the model stops learning after epoch 22. Although the training loss (blue) fluctuates, we see that as the epoch increases, the range of fluctuation becomes smaller which indicates that it starts to stabilize. Also we see a decreasing trend in the training loss as the epoch increases. And, at the last few epochs, the training loss is a little bit below validation loss which indicates that our model is somewhat overfit. However, the gap between training loss and validation loss at the last few epochs is small which means that the model is still acceptable.

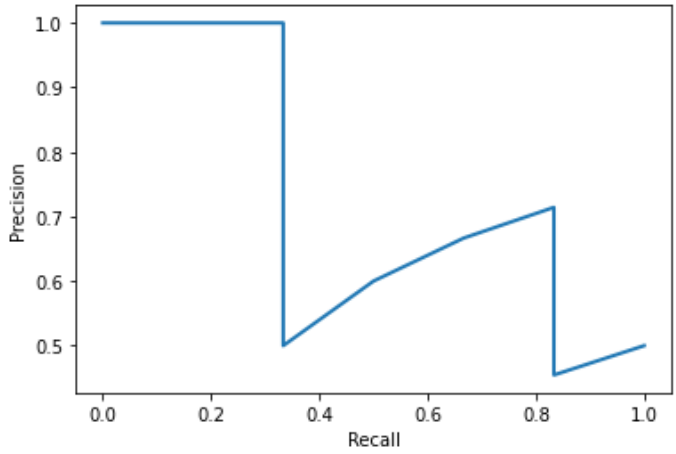


From the ROC curve below (the plot on the left), we found that the AUC value is 0.92, this means that there is a high chance that the model will be able to distinguish the positive class values from the negative class values. This is because the model is able to detect more True Positive and True Negative than False Negative and False Positive.

However, note that ROC curves can present an overly optimistic or misleading picture of the model on datasets with a class imbalance. Our model have 20% of Pneumonia and 80% of no Pneumonia cases in the validation set. Therefore, there is a little class imbalance.



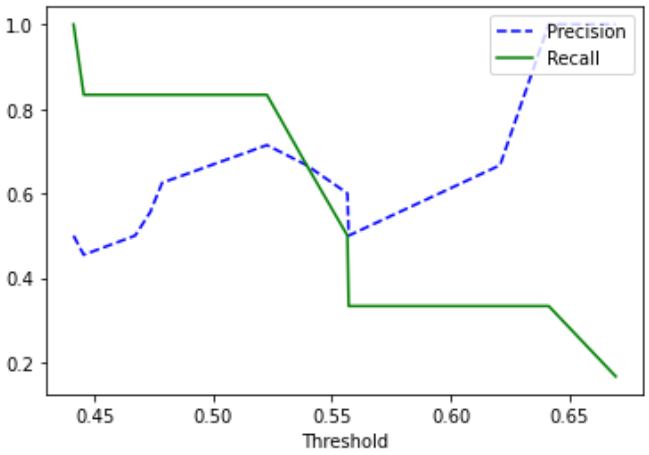
AUC: Pneumonia (AUC:0.92)

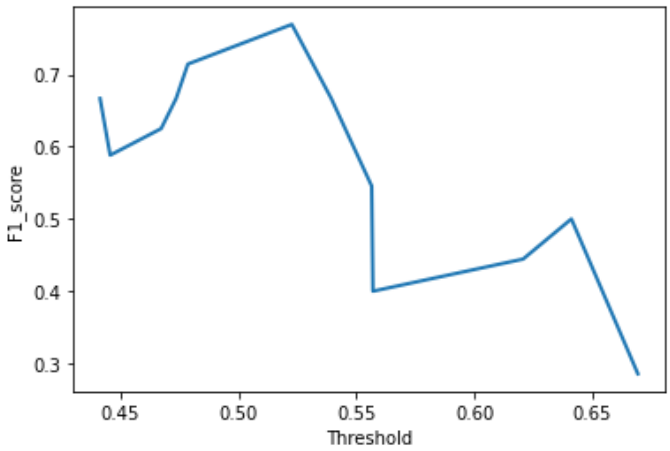


AP: Pneumonia (AP Score:0.75)

From the Precision and Recall curve above (the plot on the right), we see that the Average Precision (AP) score is 0.75 which gives us more confidence that a positive test result is actually positive since it has low False Positive.

Using the Precision and Recall curve, we have to decide what value of recall and precision which we want for our model. We have to balance the precision and recall score since if a model has high recall but low precision, then the model classifies most of the positive samples correctly but it has many False Positives (i.e. classifies many negative samples as positive). When a model has high precision but low recall, then the model is accurate when it classifies a sample as positive but it may classify only some of the positive samples.





The plots above shows the Precision & Recall vs Threshold (left) and F1-score vs Threshold (right). We see that as recall decreases, precision increases. This happens because of the trade-off between precision and recall values. The F1-score curve is calculated by using the formula 2*[(Precision * Recall)/(Precision + Recall)].

Final Threshold and Explanation: From our previous explanation, we want to have a balance precision and recall value. Therefore, we use F1-score which is the harmonic average of the precision and recall to choose our threshold. Using the above plot, we see the F1-score is the highest at about 7.6, and the threshold is about 0.52. If we trace back this threshold to the Precision & Recall vs Threshold plot, we see that the recall value is about 0.83 and the precision value is about 0.71.

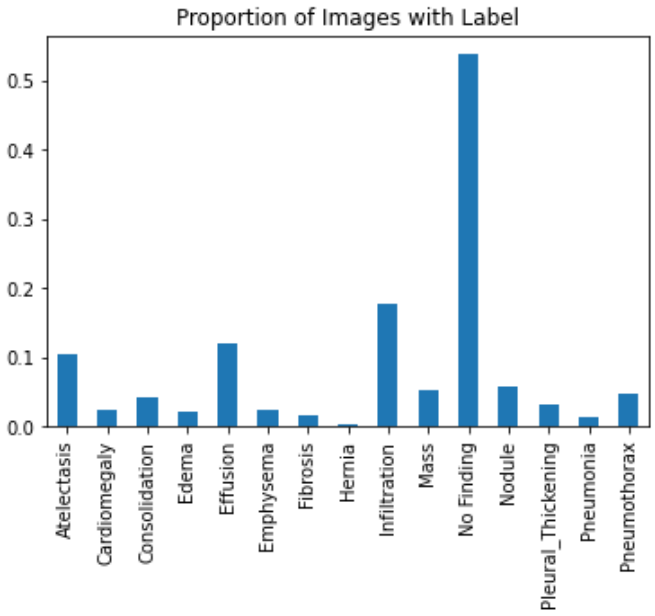
After doing a little calculation, we found that the precise threshold is 0.5226 with precision 0.7143, recall 0.8333, and F1-score 0.7692. With these values of Precision and Recall, we can have more confidence that a

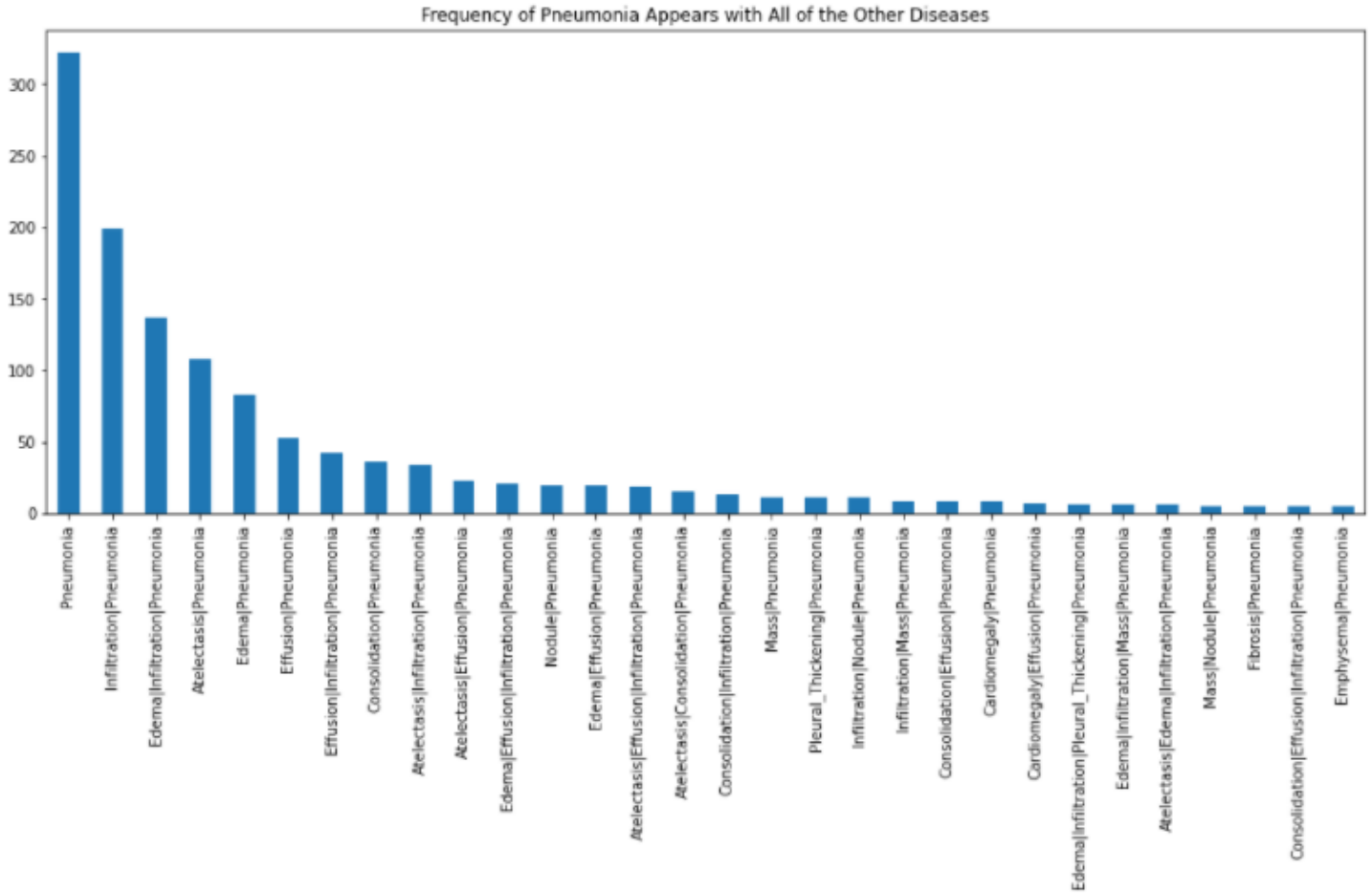
positive test result is actually positive since it has low false positive (precision) and be confident that a negative test result is truly negative since it has low false negative (recall).

4. Databases

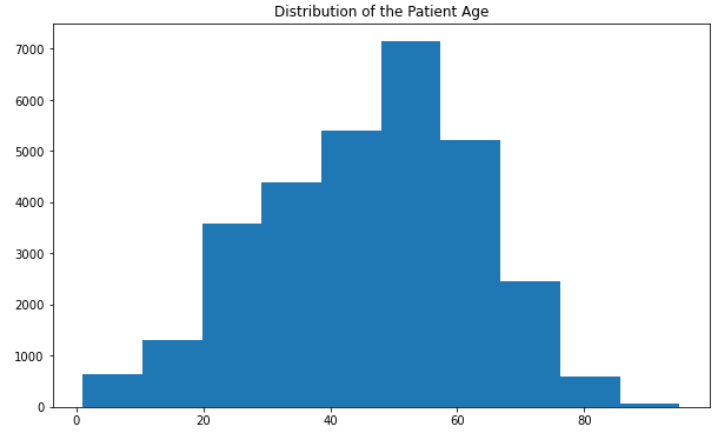
The dataset which was used for training and validation was extracted from Data_Entry_2017.csv which is a NIH Chest X-ray Dataset which comprised of 112,120 X-ray images with disease labels from 30,805 unique patients. To create these labels, the authors used Natural Language Processing to text-mine disease classifications from the associated radiological reports. The labels are expected to be >90% accurate and suitable for weakly-supervised learning.

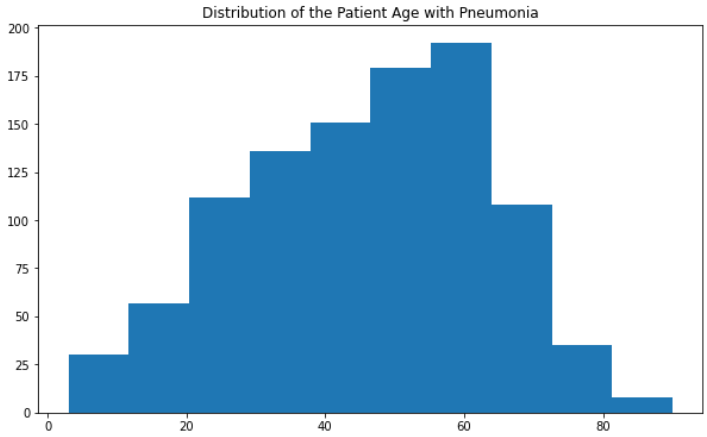
From the Exploratory Data Analysis, we found that 'No Finding' is the most common occurrence. 'No Finding' appears in 53.8% of this dataset. As for pneumonia, it's only appears in 1.3% of this dataset. Moreover, pneumonia actually occurs alone for the most part, and its most-common comorbidities are Infiltration, Edema and Infiltration, and Atelectasis (see the plots below).



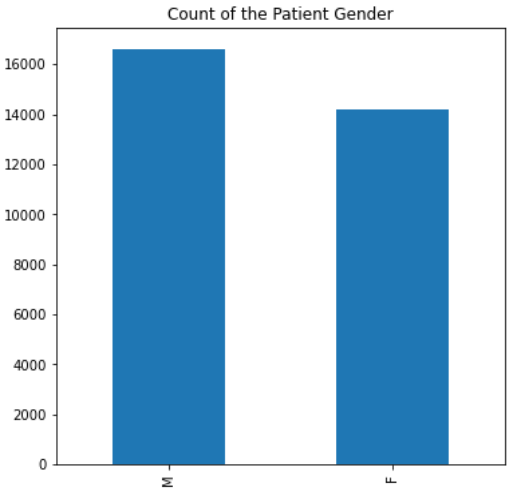


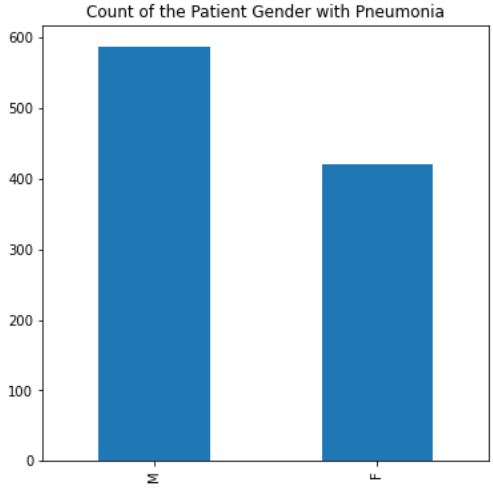
The distribution of age across the whole population is slightly different than it is with Pneumonia (see the plots below). The distribution of the whole population (left) spans the age range but has a large peak around 50. On the other hand, the distribution of the patients with Pneumonia seems to be heavier from age 3 to 65.



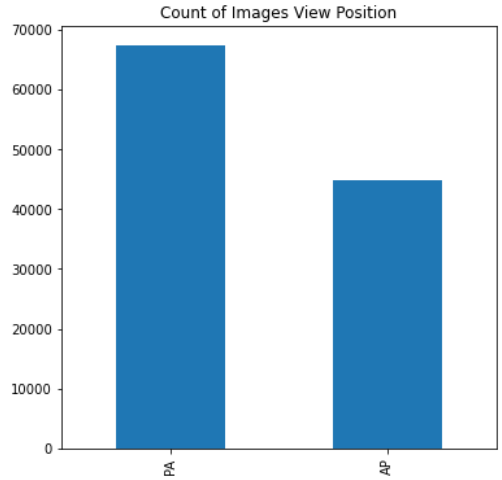


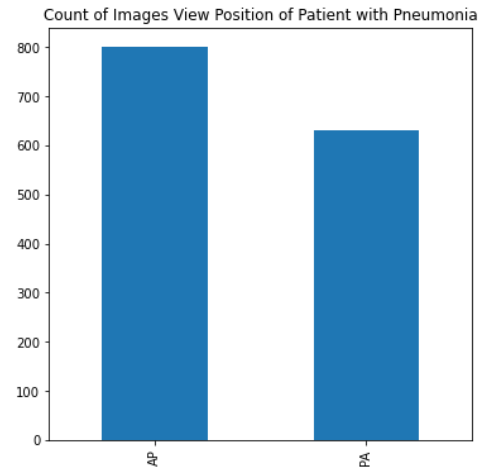
However, gender distribution (see the plots below) seems to be pretty equal in the whole population (left) as well as with Pneumonia (right), with a slight preference towards males in the Pneumonia distribution.



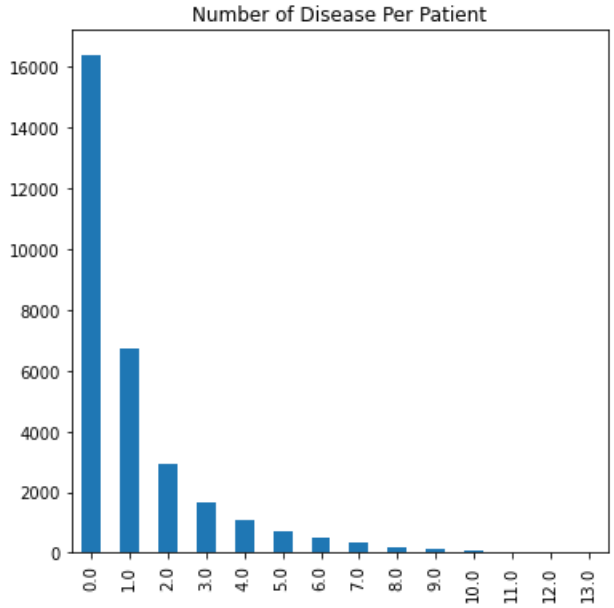


We found that the distribution of the View Position of the whole population (left plot below) in the dataset is different from the distribution of the patient with Pneumonia (right plot below) with the slight preference toward AP for the patient with Pneumonia.





Finally, we found that most of the patients (53.25%) in the dataset has no disease, follows by 21.89% has one disease, and 9.44% has two diseases (see the plot below).



There are 112120 rows of data in Data_Entry_2017.csv provided by NIH (National Institutes of Health). First, the patients whose age are above or equal to 148 were removed from the dataset because it doesn't make sense. Then we split up the 'Finding Labels' column into one additional column per disease and put a binary flag in that column to indicate the presence of the disease. Afterwards, a new column called 'Pneumonia_class' was created. This would allow us to look at images with or without pneumonia.

Later, the dataset was split into training set (train_df) and test set (val_df) with the test size equal to 20% of the whole dataset. The Pneumonia cases was stratified. After the split, we get 89,683 rows x 28 columns of the training set and 22,421 rows x 28 columns of the validation set.

Description of Training Dataset: Some of images that didn't contain Pneumonia in the training set was discarded through random sampling so we have the same amount of positive and negative cases of Pneumonia in the set, i.e., 50% of the data has Pneumonia cases and 50% of the data has no Pneumonia cases. The final training set consist of 2,288 rows x 28 columns.

Description of Validation Dataset: For the validation set, we also discarded some images with no Pneumonia through random sampling so we have 20% of positive pneumonia cases and 80% of negative pneumonia cases in the set. The final validation set consist of 1,430 rows x 28 columns.

5. Ground Truth

To create these labels, NIH used Natural Language Processing (NLP) 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 above method to create the dataset has some advantages and disadvantage. Some of the advantages of using NLP to extract the ground truth are it is less time consuming and more cost effective than using a human. On the other hand, NLP is not as accurate as a human in extracting the ground truth. This can have a negative effect on our model if our dataset contains a lot of false ground truth.

6. FDA Validation Plan

Patient Population Description for FDA Validation Dataset: Male and female distributed between the ages of 1 – 95 whose images were taken by using digital Chest X-Ray (DX) with view position either AP or PA. About 53.8% of the dataset should have 'No Finding', and as for pneumonia, it should appear in about 1.3% of the dataset. Additionally, comorbidities among pneumonia should be atelectasis, cardiomegaly, consolidation, effusion, emphysema, fibrosis, hernia, mass, nodule, pleural thickening, and pneumothorax. Infiltration and Edema should not be included in the dataset since from our EDA, we found that they have the same distribution as that of pneumonia. If edema and infiltration are part of the comorbidities, there is a high chance that our algorithm will label these disease as pneumonia too and this may lower the accuracy and reliability of our model.

Ground Truth Acquisition Methodology: To obtain an optimal ground truth, four radiologists with more than 3 years of experiences, independent from our department, should extract the label from the Chest-X-Rays. Radiologists should not have access to any patient information or knowledge of disease prevalence in the data.

Algorithm Performance Standard: From CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning (arXiv:1711.05225**,** we found that the average F1 score obtained from 4 radiologists was 0.387. Also, the best AUROC published by Yao et al.(2017) was 0.713. Our algorithm has a F1 score of 0.7692, which is about 0.3822 point better than the average radiologist. Similarly, our algorithm has AUCROC value of 0.92, which means that our algorithm performs 0.207 point better than the best published result.