**Great Learning**

**Capstone Project – Interim Report (Milestone 1)**

**Pneumonia Detection Challenge**

**Authors:**

1. **Ajay Devnani**
2. **Manojkumar Sivaraman**
3. **Naveen Kumar Kaginelli**
4. **Sumeet Kumar**
5. **Umesh Singh Kushwaha**

**Abstract**

Pneumonia is an infection in one or both lungs. Bacteria, viruses, and fungi cause it and the infection causes inflammation in the air sacs in your lungs, which are called alveoli. The alveoli fill with fluid or pus, making it difficult to breathe. Typically, X-ray helps doctors to look for signs of inflammation or opacities in chest which when present can indicate the Pneumonia infections. Since Pneumonia accounts for over 15% of all deaths of children under 5 years old internationally, it is crucial to identify and react swiftly if there are any infections identified.

**Business Perspective**: As the detection and reaction time is vital and the infection is detected using the X-ray Images, Image processing techniques can be leveraged from the emerging AI technology on these images to predict the presence of opacities. Powerful AI techniques can unlock clinically relevant information hidden in the massive amount of data, which in turn can assist clinical decision making. This will also assist physicians to make better clinical decisions or even replace human judgement in certain functional areas of healthcare (eg, radiology).

For this purpose, this project uses instance segmentation and semantic segmentation techniques to create the Pneumonia prediction model which can predict Pneumonia on the patients with an accuracy of XX. XX% and XX.XX% respectively there by helping the doctors to react quickly to save lives.

**Problem Statement**

The problem is about detecting lung inflammations (opacities) corresponding diagnosis of Pneumonia on chest radiographs (images). Tissues with sparse material, such as lungs which are full of air, do not absorb the X-rays and appear black in the image. Dense tissues such as bones absorb X-rays and appear white in the image. In the data, some of these such area labeled as “Not Normal No Lung Opacity”.

All lung opacities may not attribute to Pneumonia as the Pneumonia is one of the several diseases that can occur on a chest (lungs) radiograph. The “Not Normal No Lung Opacity” class indicates that, while pneumonia determined not to be present, there could be nonetheless some type of abnormality on the image. And oftentimes this finding may mimic the appearance of true pneumonia. A radiograph may contain one or more than one locations for any possible Pneumonia case.

**Summary of data and findings:**

The data is spread across different files and folders. The details is as given below,

* **stage\_2\_train\_images:** It contains a set of raw medical images (DICOM files) for training models. The DICOM files contain a combination of header metadata as well as underlying raw image arrays for pixel data.
* **stage\_2\_test\_images:** It contains a set of raw medical images (DICOM files) for testing the model. The file contains a combination of header metadata as well as underlying raw image arrays for pixel data**.**
* **stage\_2\_train\_labels.csv:** This CSV file contains detailed information about the labels (Patient ID, bounding boxes for lung opacity and target 1 or 0 indicate the presence of abnormality i.e. Pneumonia)
* **stage\_2\_detailed\_class\_info.csv:** This CSV file contains information regarding three possible classes in the data - namely normal, lung opacity and no lung opacity (not normal).
* **DICOM files:** The original medical images are stored in a special format called DICOM files (\*.dcm). It contains a combination of header metadata as well as underlying raw image arrays for pixel data.

**Findings:**

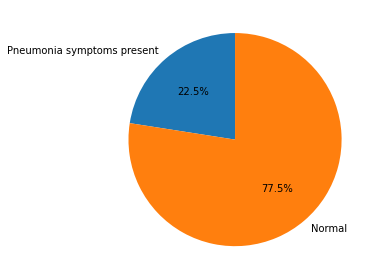
* The most important thing is that all lung opacities may not attribute to pneumonia, as the pneumonia is one of the several diseases that can occur on a chest radiograph.
* A radiograph may contain one or more than one bounding boxes for any possible pneumonia case.

**EDA and Pre-processing**

**Exploratory data analysis (EDA):**

1. **Distribution of Pneumonia Vs Non-Pneumonia:**

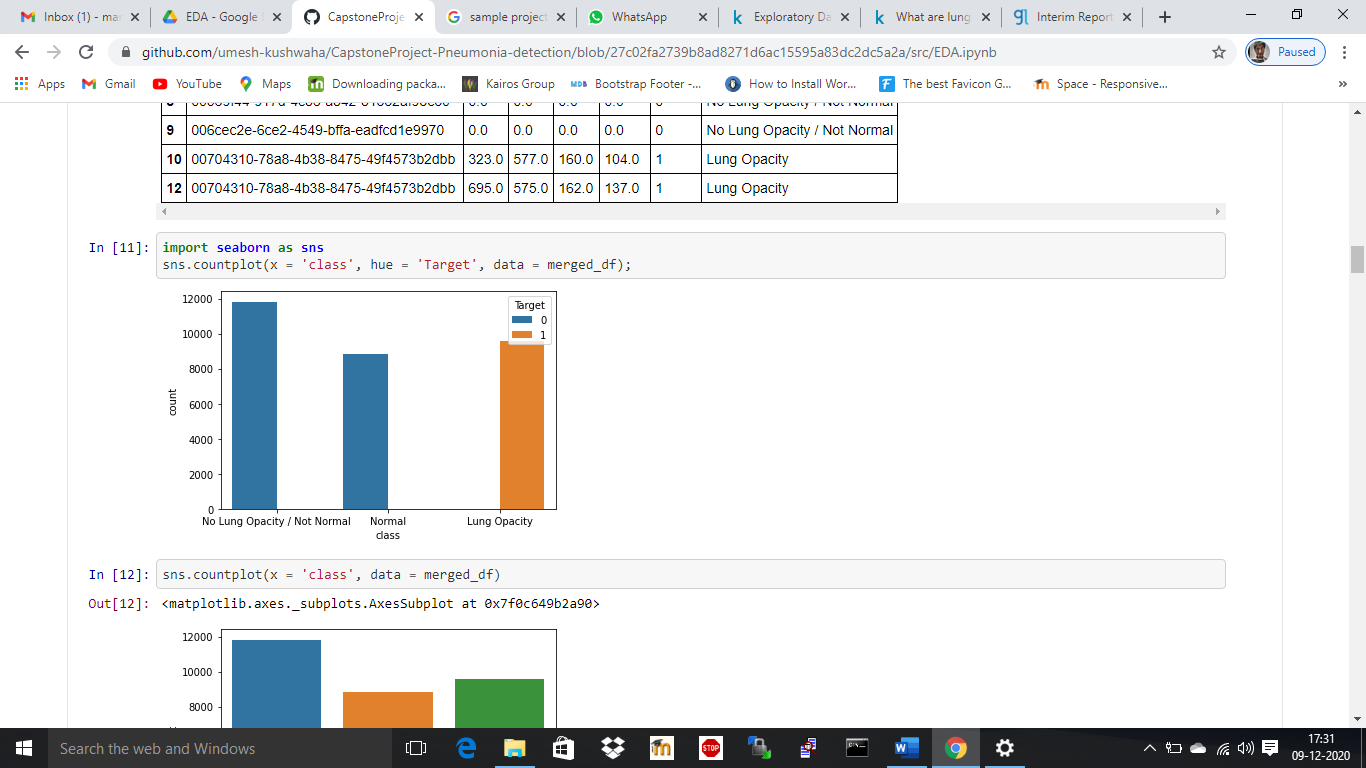
* 33.2% (8,851) patients are normal, do not have any lung related abnormalities
* 22.5% (6012) patients have *lung opacities* which attributes to pneumonia.
* 44.3% (11,821) patients do not have pneumonia but are not normal possibly due to other lung ailments.
* Hence, **22.5% of patients are suffering from pneumonia** and the remaining **77.5% are pneumonia negative**.

****

**Action Taken:** Data is not balanced, to balance the data augmentation technique is used during data generation.

1. **Distribution of 3 different classes data with target:**

* The count of patients with *No Lung opacities/ Not normal* is higher than the pneumonic or normal patients
* the count of normal class is less than other 2 classes indicating that the data has a greater number of Ill health patients

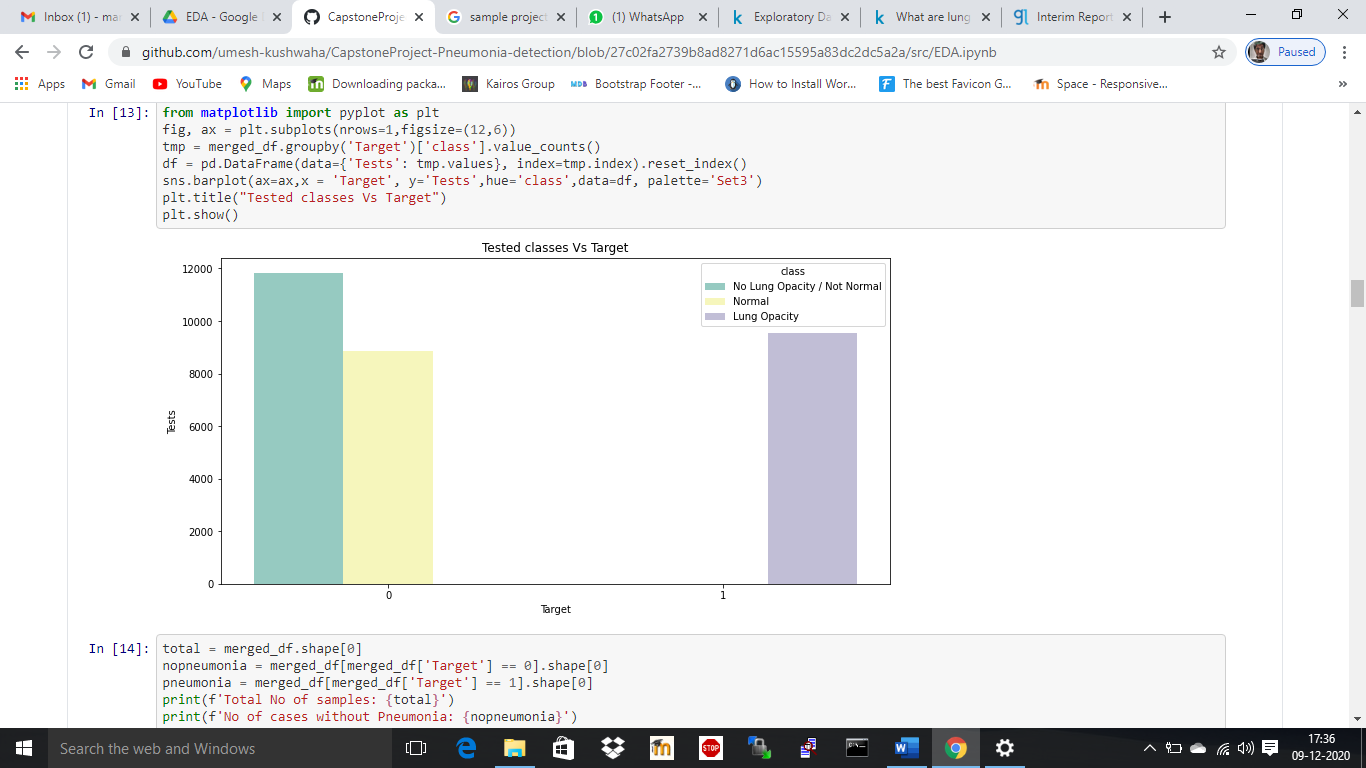


**Action Taken:** No Lung Opacity / Not normal class corresponds to pneumonia but other possible lung ailments. Therefore, this class is considered as no pneumonia i.e. normal patients.

1. **Distribution of target (Pneumonia & Non-Pneumonia) with 3 different classes:**

The same can be observed when plotting the count of Target values segregating the classes.

* Patients with *No Lung Opacity/ Not normal* observations are more than there other 2 classes



**Action Taken:** No Lung Opacity / Not normal class corresponds to pneumonia but other possible lung ailments. Therefore, this class is considered as no pneumonia i.e. normal patients.

1. **Train set:**

* 26,684 images are available in the training set are unique (equal to unique patient IDs).

1. **Bounding box:**

* Out of 26,684 images available in the training set, 23,286 images have only 1 bounding box, 3,266 images have 2 bounding boxes, 119 images have 3 boxes and 13 images have 4 boxes.
* 3,398 patients have more than 1 bounding box.
* If any patient has more than one lung opacity area, then the patient does have pneumonia or high likely to have pneumonia.

|  |  |
| --- | --- |
| **No of Occurrences** | **Count of the Patient ID** |
| 1 | 23286 |
| 2 | 3266 |
| 3 | 119 |
| 4 | 13 |

1. **Data characteristics:**

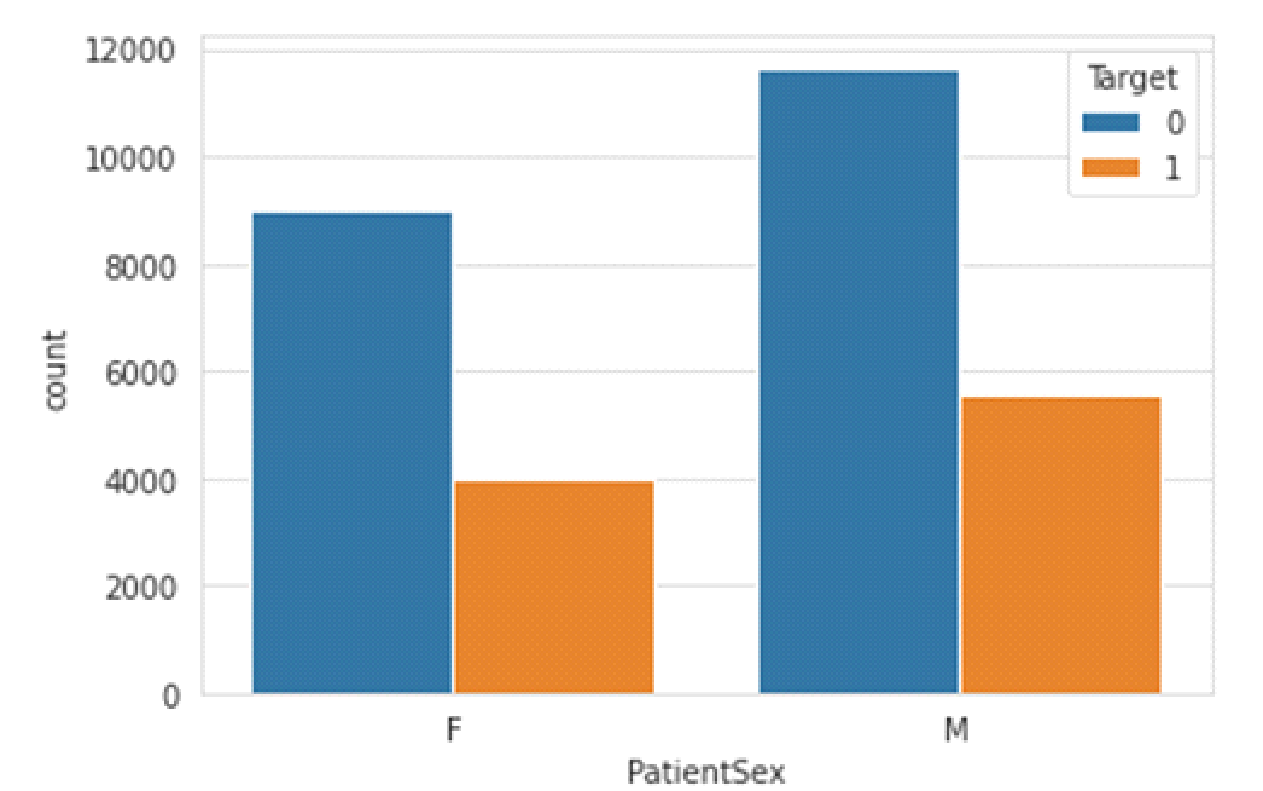
* We have different parameters or characteristics of available information – patient age, sex, body part examined, view position, rows and columns, pixel spacing, etc.

1. **Correlation:**

* We have observed that ‘Target’ and ‘View Position’ have a higher correlation and stand at 0.42.

1. **Gender mix: Distribution Gender Vs Target information**

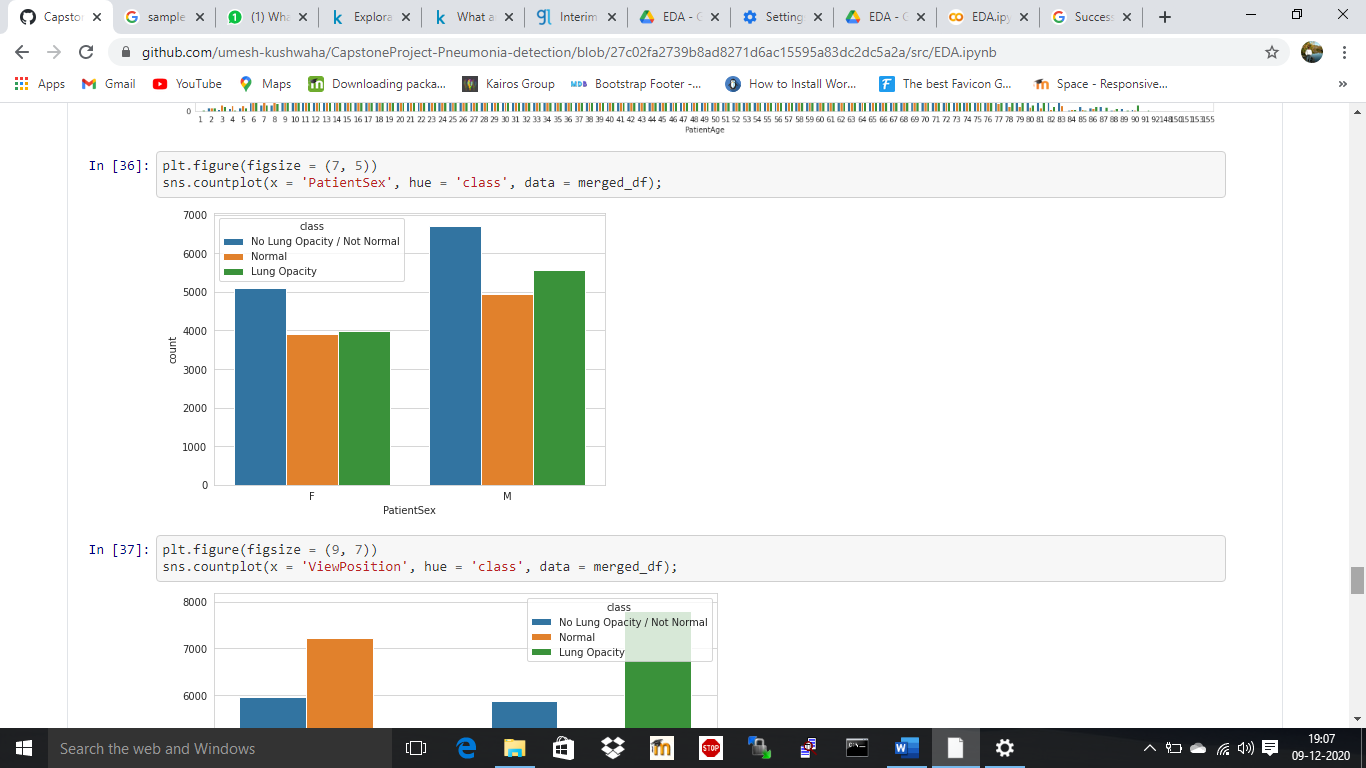
* Looking at the high pillars, there are more test samples (radiographs) for males than females.
* Out of total 9,555 cases of Lung opacity, ~60% is male and rest 40% is female.
* Approximately, one third of the total cases are diagnosed as pneumonia for both the genders.



**Action Taken**: To balance the data augmentation with random shuffling is used.

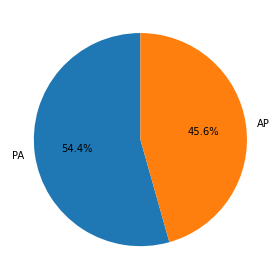
1. **Gender mix: Distribution Gender Vs class information**

* No opacity but Not Normal cases constitute higher number indicating that the patients could be suffering from other lung related illness but pneumonia.
* Males who are diagnosed for lung opacities (Pneumonia) are slightly high in number comparatively.

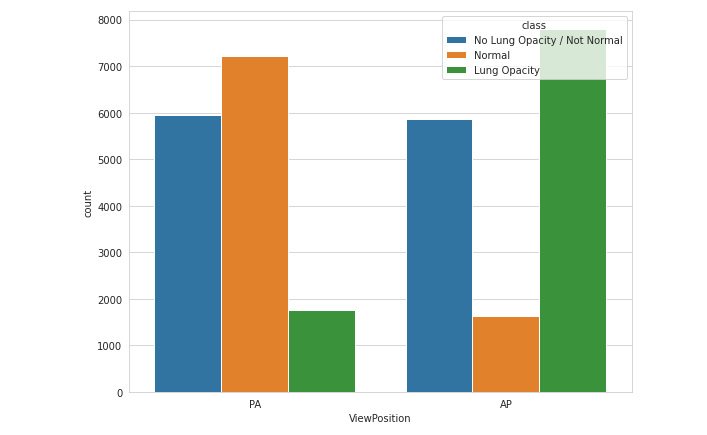


**Action Taken**: To balance the data augmentation with random shuffling is used.

1. **Distribution of 3 classes data over View position:**



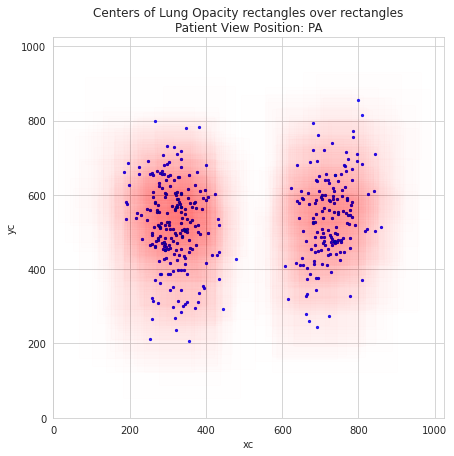
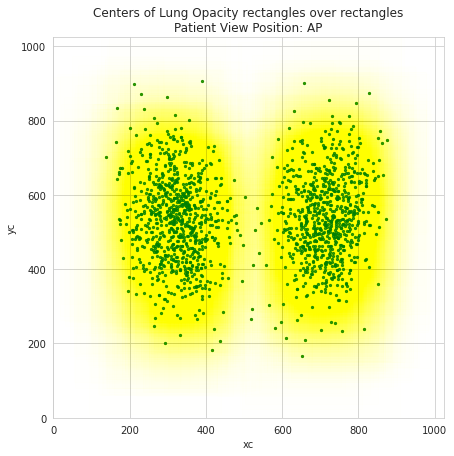
* Radiographs images with PA position are considered to be of good standard in medical profession. The PA amounts to 54.4% of total radiograph images. Therefore, data augmentation with random shuffling to balance the data to spread the impact of AP evenly.



* Evidently **PA position pointing considerably less lung opacities** than the AP position. Whereas the *no opacity but not normal* class seems to be the same in both the position.

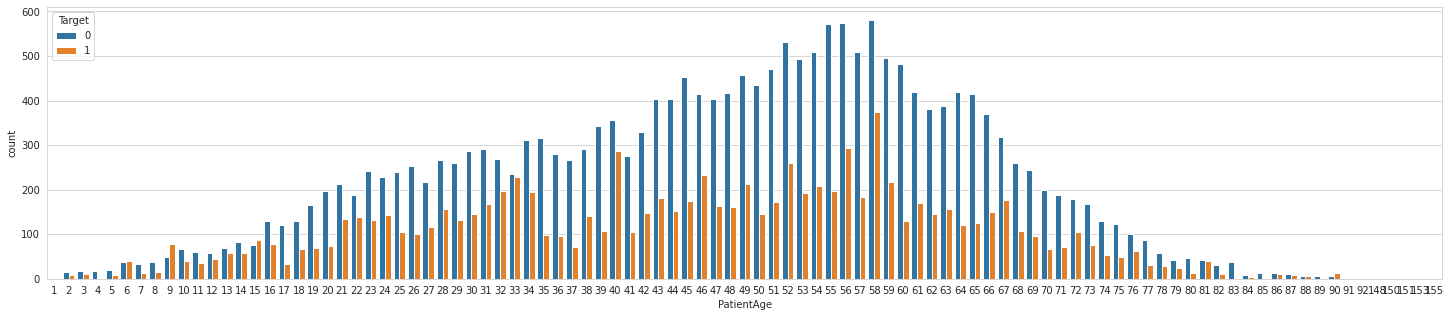
**Action Taken**: Data augmentation with random sampling is used to balance the data. However, this is an important insight to be noted and taken care in the model improvement task.

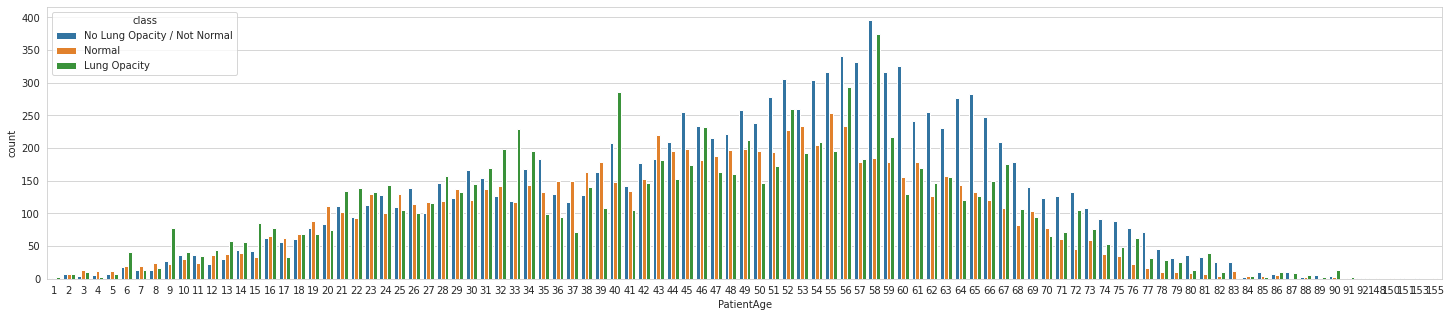
* In the below chart, concentration of Lung opacities for AP is larger, whereas for the PA cases, it’s less.



1. **Distribution data over different age group:**

* Three forth of the total test reports falls within age group of 25 – 70. The peak is between 50-60. It can be observed that lung related ailments are more common in this age group.
* Pneumonia cases are higher between the age group of 30-65 years.
* The lung opacities are spread across the patient age and have peaks within age 50 to 65 years.

****



**Action Taken**: Data augmentation with random sampling is used to balance the data

**Data Preprocessing:**

* **Image scale:** The images have been windowed and leveled already, as have been rescaled to 8-bit encoding and the resolution has been rescaled to (1024, 1024).
* **Data discrepancy:** There is no discrepancy in the data, as the data in the class csv and label csv is the same.
* **Data imbalance:** Data augmentation and random shuffling is used to balance the data.

**Deciding Models and Model Building**

The goal is to predict whether the patient is suffering from Pneumonia or not, therefore there are 2 related task, firstly the radiograph images have to analyzed for potential lung opacities and secondly, whether the predicted lung opacities attribute to the pneumonia or not.

Keeping the goal in mind the following algorithms are used,

1. **Mask RCNN**

Mask RCNN is a deep neural network aimed to solve instance segmentation problem. It can separate different objects in an image or a video. You give it an image, it gives you the object bounding boxes, classes and masks.

There are two stages of Mask RCNN,

* First, it generates proposals about the regions where there might be an object based on the input image.
* Second, it predicts the class of the object, refines the bounding box and generates a mask in pixel level of the object based on the first stage proposal.

**Model Performance**:

Accuracy: 78.36%

loss: 1.2235

mrcnn\_class\_loss: 0.1848

mrcnn\_bbox\_loss: 0.3528

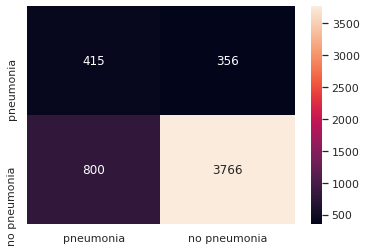
mrcnn\_mask\_loss: 0.3628

val\_loss: 1.2829

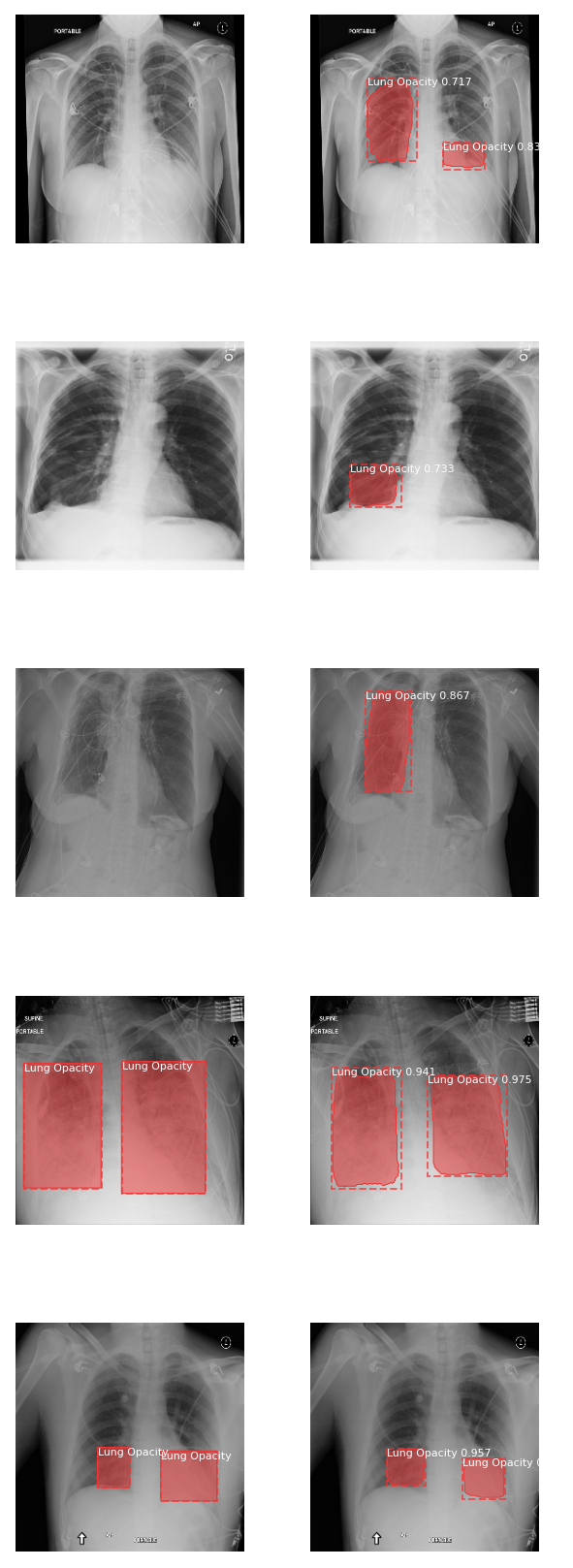
val\_mrcnn\_class\_loss: 0.19

val\_mrcnn\_bbox\_loss: 0.38

val\_mrcnn\_mask\_loss: 0.37



Attached a screenshot which indicates bounding boxes in the lung radiograph images.



**How to improve your model performance?**

* Hyper parameter tuning – tuning the learning rate
* Retraining the model with larger image size (resolution)
* Introducing other techniques in terms of augmentation
* Adding dropped out and normalization layers on top

1. **VGG -** Very Deep Convolutional Networks for Large-Scale Image Recognition.

VGG has achieved 92.7% top-5 test accuracy in ImageNet, for a dataset of over 14 million images belonging to 1000 classes.

* Due to the high accuracy score, this model has been chosen to predict the lung opacity locations in the chest radiographic images.
* Using Transfer learning technique VGG16 used as base network.
* Added a layer to upsample the output tensor to the size of original image in order to fins the mask.
* Trained 4 layers for all the 5K images.

**Model Performance**:

Model Accuracy: 0.8839%

Validation Accuracy: 0.7278

Training Loss: 0.6395

Validation Loss: 0.7274

**How to improve your model performance?**

Even though accuracy and loss metrics show excellent score, the model (trained on 5k) didn’t perform well in locating the opacities (bounding boxes). Therefore, the model has been dropped.

Planning to try with skip connection and making it a proper Unet.

1. **Inception Network**

Inception network known for its speed and accuracy. Due to the large number of training images (~26K, the network is used as backbone network to find the bounding boxes.

* Added a layer to upsample the output tensor to the size of original image in order to fins the mask.
* Trained 4 layers for all the 5K images.

**Model Performance**:

Model Accuracy: 0.9340%

Validation Accuracy: 0.9380

Training Loss: 0.7086

Validation Loss: 0.9380

**How to improve your model performance?**

Even though accuracy and loss metrics show excellent score, the model (trained on 5k) didn’t perform well in locating the opacities (bounding boxes). Therefore, the model has been dropped.

Planning to try with skip connection and making it a proper Unet.

**Future course of action:**

* Analyzing the different algorithms/techniques already tried to achieve the goal.
* Employing hyper parameter tuning technique on the models already trained to improve the accuracy.
* Look for semantic segmentation algorism like YOLO to achieve the goal.
* Choosing the best working and fitting model for the current problem.
* Deploying the model on cloud.

**Appendix**

**Code:**

|  |  |  |
| --- | --- | --- |
| **1** | Mask RCNN Model |  |
| **2** | VGG16 based Model |  |
| **3** | Inception based Model |  |