

PNEUMONIA DETECTION Assignment

Interim Report



Team:

|  |  |
| --- | --- |
| **Name** | **E-mail** |
| Arup Kumar Mohanty | Arup.kumar1988@gmail.com |
| Mohit Gupta | Mohitgupta.mait@gmail.com |
| Ved Verma | Ved.verma.273306@gmail.com |
| Venkatesh Doddihal | Ds.venky1975@gmail.com |

Table of Contents

[**1.** **Summary of the problem statement, Data and findings** 4](#_Toc103275136)

[1.1. Problem Statement 4](#_Toc103275137)

[1.2. Project Objectives: 4](#_Toc103275138)

[**2. EDA:** 5](#_Toc103275139)

[2.1. Approach: 5](#_Toc103275140)

[2.2. Analysis: 5](#_Toc103275141)

[2.3 Visualization: 8](#_Toc103275142)

[2.4 Summary: 14](#_Toc103275143)

[**3.** **Pre-Processing** 14](#_Toc103275144)

[3.1 Pre-processing Methods 14](#_Toc103275145)

[3.2 Pre-processing Applied 15](#_Toc103275146)

[3.3 Transformations 16](#_Toc103275147)

[3.4 Summary 16](#_Toc103275148)

[**4.** **Model and Model building** 16](#_Toc103275149)

[4.1 Model Approach 16](#_Toc103275150)

[4.2 Model Evaluation Metrics 16](#_Toc103275151)

[4.3 Model creation 16](#_Toc103275152)

[4.4 Model Summary 16](#_Toc103275153)

[**5** **Improve Model Performance** 18](#_Toc103275154)

[5.1 Approaches 18](#_Toc103275155)

[5.2 Summary 18](#_Toc103275156)

# **Summary of the problem statement, Data and findings**

## Problem Statement

Pneumonia is an infection in the lung, which requires review of a chest radiograph by highly trained specialists. Pneumonia shows up in a chest radiograph as an area of opacity. However, diagnosis of it can be complicated and much time and effort is spent by specialists in reviewing them. Chest radiograph is the most common performed diagnostic imaging study. Due to the high volume of chest radiography, it is very time consuming and intensive for the radiologists to review each image manually. As such, an automated solution is ideal to locate the position of inflammation in an image. By having such an automated pneumonia screening system, this can assist physicians to make better clinical decisions or even replace human judgement in this area.

## Project Objectives:

* To build a deep learning a pneumonia detection system, to locate the position of inflammation in an image.
* Use TensorFlow/Keras as the framework for building the model
* Read Medical images are stored in a special format called DICOM files (\*.dcm).
  1. **Data & Findings :**
* Details about the data and dataset files are given in below link,  
  <https://www.kaggle.com/c/rsna-pneumonia-detection-challenge/data>
* The first step would be to examine the data available for this. The data is given in a zip file “rsna-pneumonia-detection-challenge.zip”, which contains the following items:
* A folder “stage\_2\_train\_images”: This folder contains all the training dataset chest radiograph DICOM images.
* A csv file “stage\_2\_train\_labels.csv”: This file contains the corresponding patientID images to the folder “stage\_2\_train\_images” and contains the bounding box of areas of pneumonia detected in each image along with a target label of 0 or 1 for pneumonia detected.
* A csv file “stage\_2\_detailed\_class\_info.csv”: This file contains the corresponding patientID images to the folder “stage\_2\_train\_images” and contains the target class labels of the images.
* A folder “stage\_2\_test\_images”: This folder contains all the test dataset chest radiograph DICOM images. We will not be using this set of images as they do not contain labels.
* A csv file “stage\_2\_sample\_submission.csv”: This file contains the corresponding patientID images to the folder “stage\_2\_test\_images”. We will not be using this set of file.

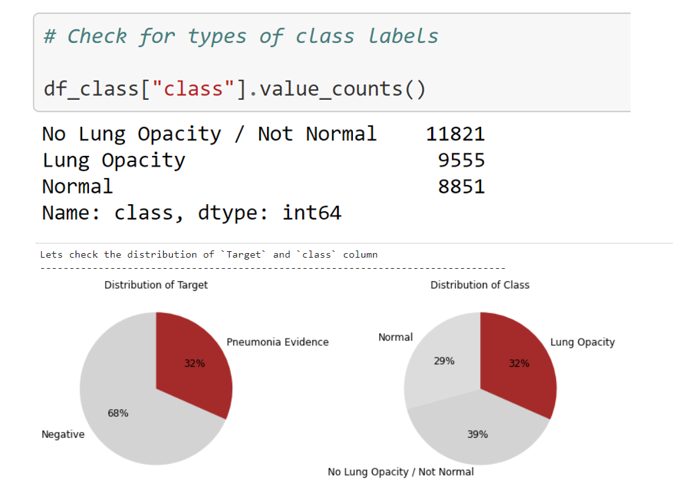
# **2. EDA:**

## 2.1. Approach:

* To support the building of a neural network, the project will be done on **google Colab**.
* The first step is to unzip the zip file to open the above files to the google drive directory.
* Second is to verify the file format of the images provided, and they are all DICOM images in the “.dcm” file format. In order to open and read the DICOM images, we are using the **pydicom** library for that purpose. After that, the next step would be to inspect the csv files.

## 2.2. Analysis:

* Loading the “**stage\_2\_detailed\_class\_info.csv**” file into pandas dataframe, a quick glance reveals that it has only 2 columns:
* **patientId** – which refers to the patientId’s corresponding image name
* **class** – Target label of the patientId’s image



* As illustrate in above figure , there are 3 types of classes: **Normal, Lung Opacity and No Lung Opacity / Not Normal**. 11821(~39%)records belongs to No Lung Opacity / Not Normal, 32% accounts for Lung Opacity and roughly 29% marked as Normal. The primary concern of the project would be to detect images with Lung Opacity, and the others would be in the same group labelling.
* The Target distribution seems to be imbalance as 32% of the patients are having pneumonia evidences where as 68% are normal.
* Loading the “**stage\_2\_train\_labels.csv**” file into pandas dataframe, we can see that it has below fields:

**patientId** – which refers to the patientId’s corresponding image name

**x** - upper-left x coordinate of the bounding box

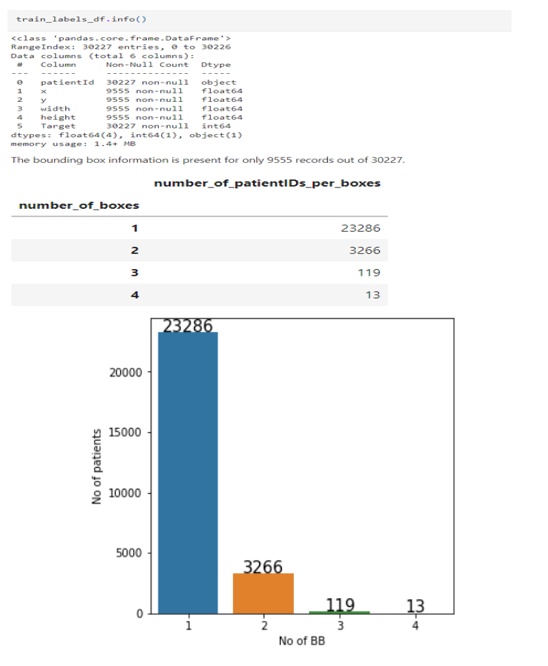
**y** - upper-left y coordinate of the bounding box

**width** – the width of the bounding box

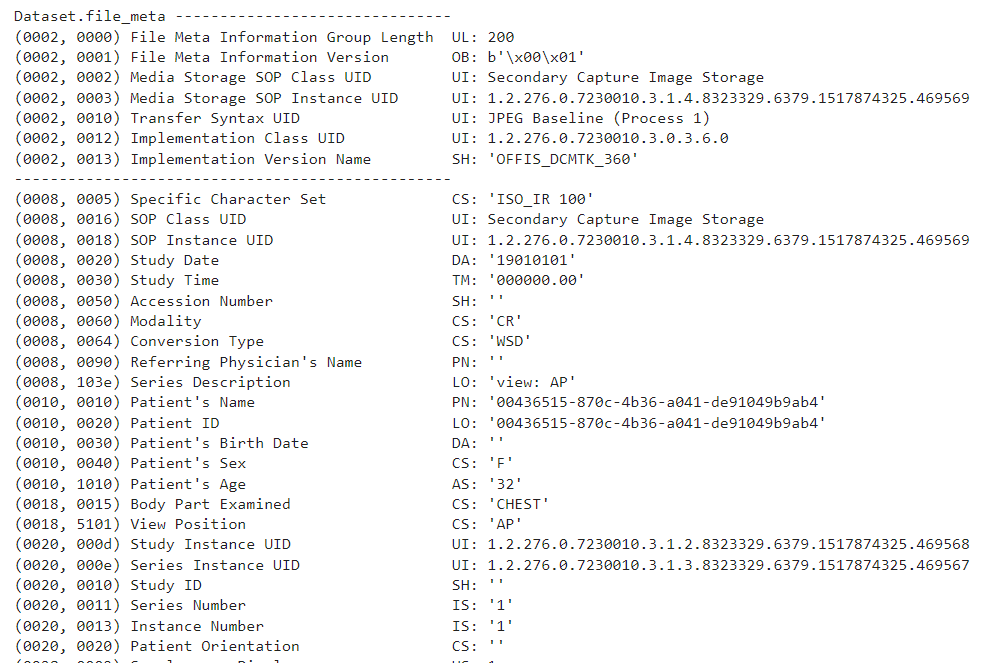
**height** – the height of the bounding box

**Target** – binary target indicating if this image has evidence of pneumonia

* There is a total of 30,227 entries, no missing values with 9,555 images have bounding boxes. This corresponds to the data available “stage\_2\_detailed\_class\_info.csv” file.



* There are multiple records for patients. Number of duplicates in patientID is 3,543.
* After merging both the csv files below are the observations.
* About 23,286 patientI
* ds (~87% of them) provided have 1 bounding boxes while 13 patients have 4 bounding boxes. The reason is because each row records a single bounding box area of pneumonia detected. However, in a patient image, it might be the case of several bounding boxes area of pneumonia detected.
* Chest examinations with Target = 1 i.e. ones with evidence of Pneumonia are associated with Lung Opacity class.
* Chest examinations with Target = 0 i.e. those with no definitive evidence of Pneumonia are either of Normal or No Lung Opacity / Not Normal class.
* The next step is to read the images in the file “**stage\_2\_train\_images**”. Images provided are stored in DICOM (.dcm) format which is an international standard to transmit, store, retrieve, print, process, and display medical imaging information. We will make use of pydicom package here to read the images.

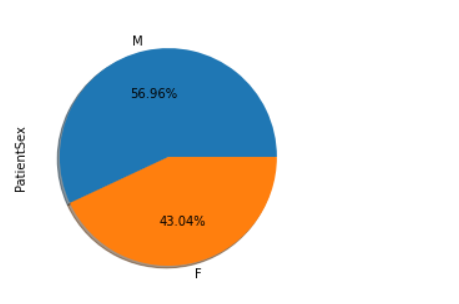


* From the above sample we can see that dicom file contains some of the information that can be used for further analysis such as **sex, age, body part examined , view position and modality**. Size of this image is 1024 x 1024 (rows x columns).
* To examine further we will merge the image features with the existing class data. This will help us understand distribution of age for those with evidence of lung opacity and those with no definite evidence of lung opacity.
* To understand distribution of male and female for those with evidence of lung opacity and those with no definite evidence of lung opacity
* Explore different view positions in the dataset
* Explore modality.
* We will pickle the file and do our analysis on the saved file.

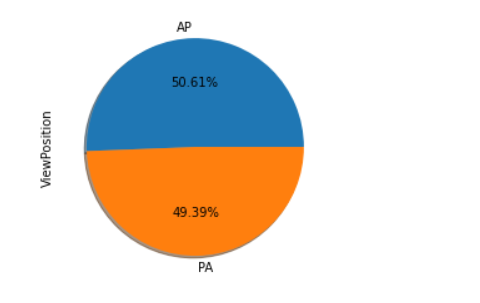
## 2.3 Visualization:

* As we proceed further we will use different visualization techniques like univariate, multivariate analysis to discover patterns and anomalies in the data.

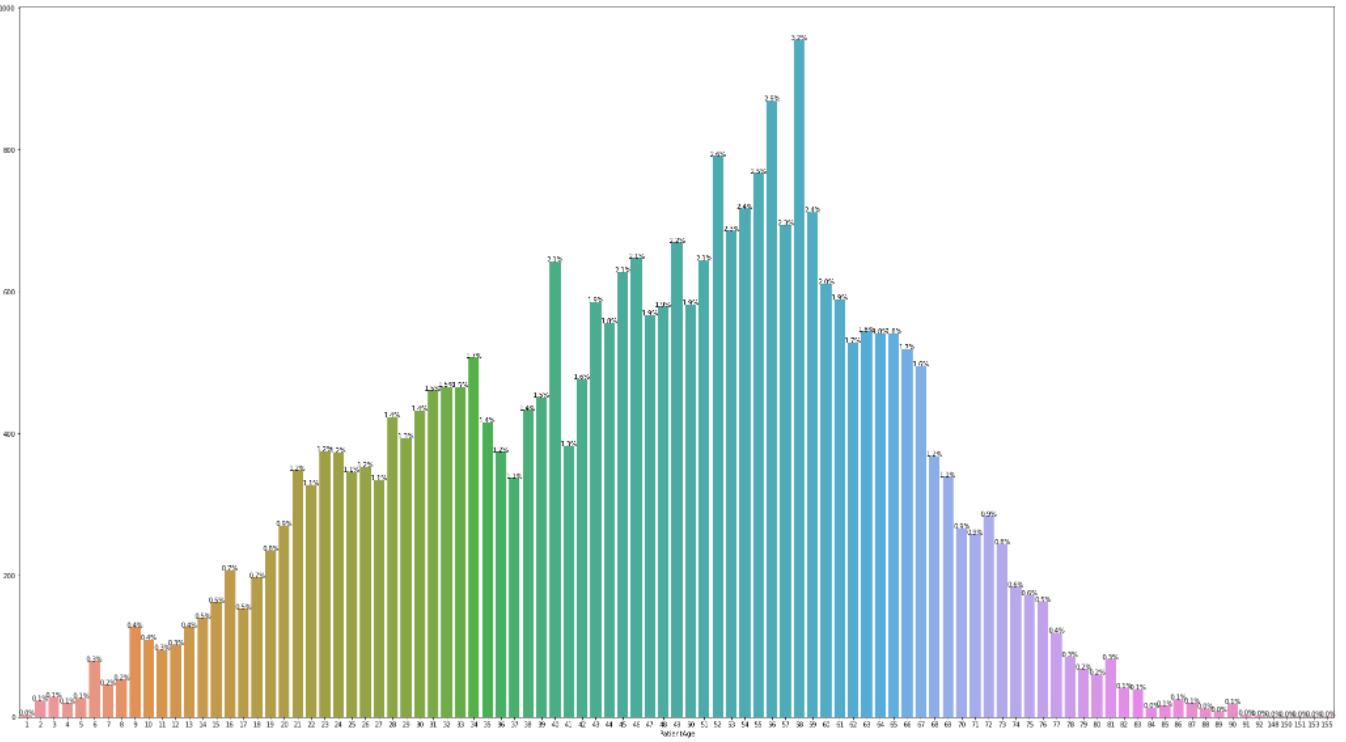
**Univariate Analysis:**



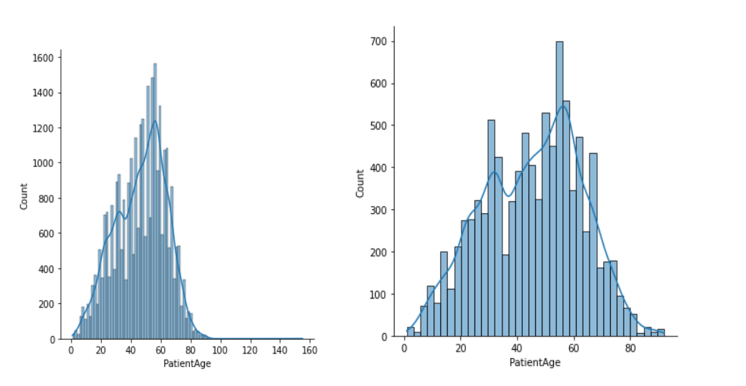
* As shown above, There are 56.96% male patients and roughly 43% female patients. There are also more no of male patients having pneumonia compared to females.



* We have two different view positions AP (Anterior/Posterior) and PA (Posterior/Anterior) in the training dataset.
* **Posterior/Anterior (PA)**: In PA, X-Ray beam hits the posterior (back) part of the chest before the anterior (front) part. While obtaining the image patient is asked to stand with their chest against the film.
* **Anterior/Posterior (AP)**: At times it's not possible for radiographers to acquire a PA chest X-ray. This is usually because the patient is too unwell to stand. AP projection images are of lower quality than PA images. Heart size is exaggerated (cardiothoracic ratio approximately 50%)
* As can be seen above in the chart , the view position attributes is almost equally distributed.



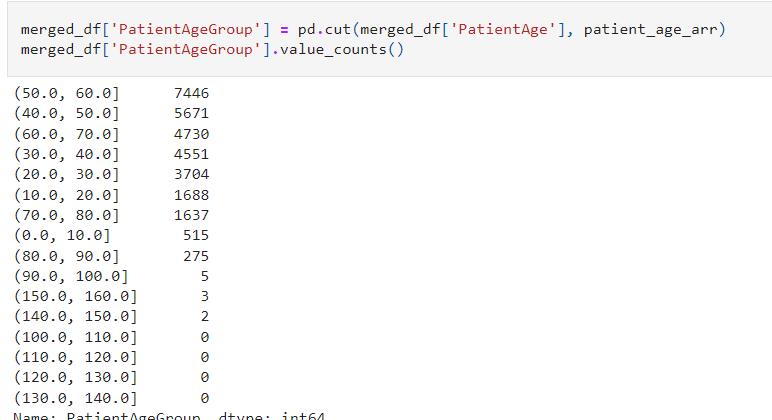
* The above graphs shows the distribution of age across all the patients. As per the data the maximum no of patients are falling within age group 40-60.
* There are also few patients where age > 100. These records seems to be erronous and we can safely remove them from our dataset and will keep the age

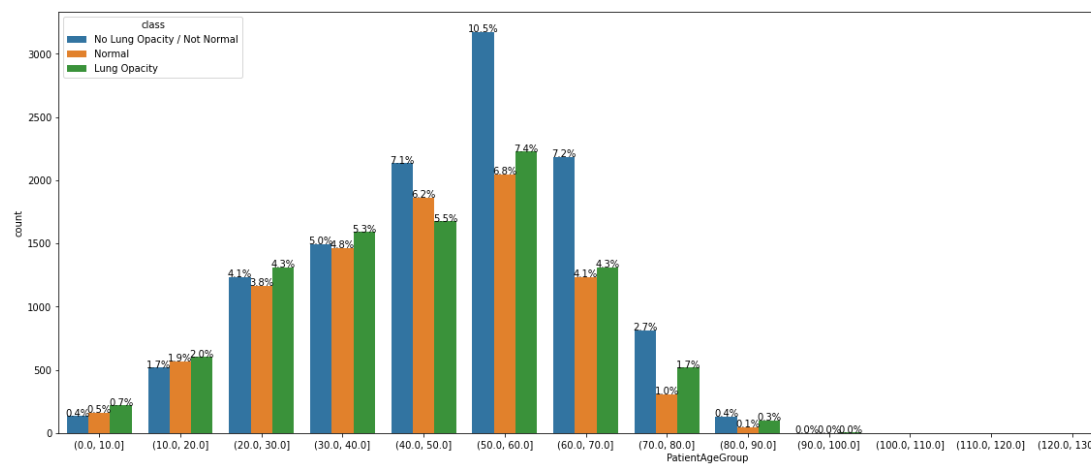


* As per the above histogram, the patient age is normally distributed with more volume of data lying between age group 40-60.
* The Age distribution for pnenumonia patient is slightly left skewed with more no of patient between age group 40-60 are having pneumonia.

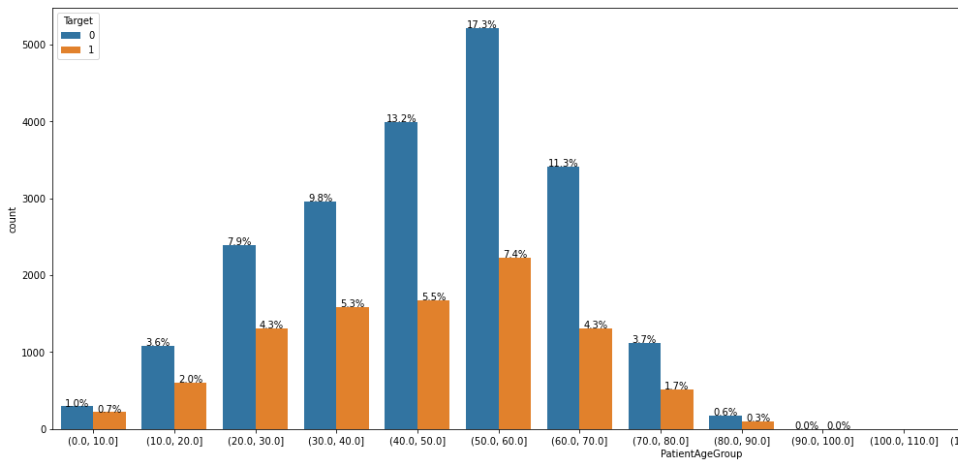
**Bivariate Analysis:**

* To have a better interpretation we will make use of binning concept as shown below to group the patient age and will further visualize on the transformed data.

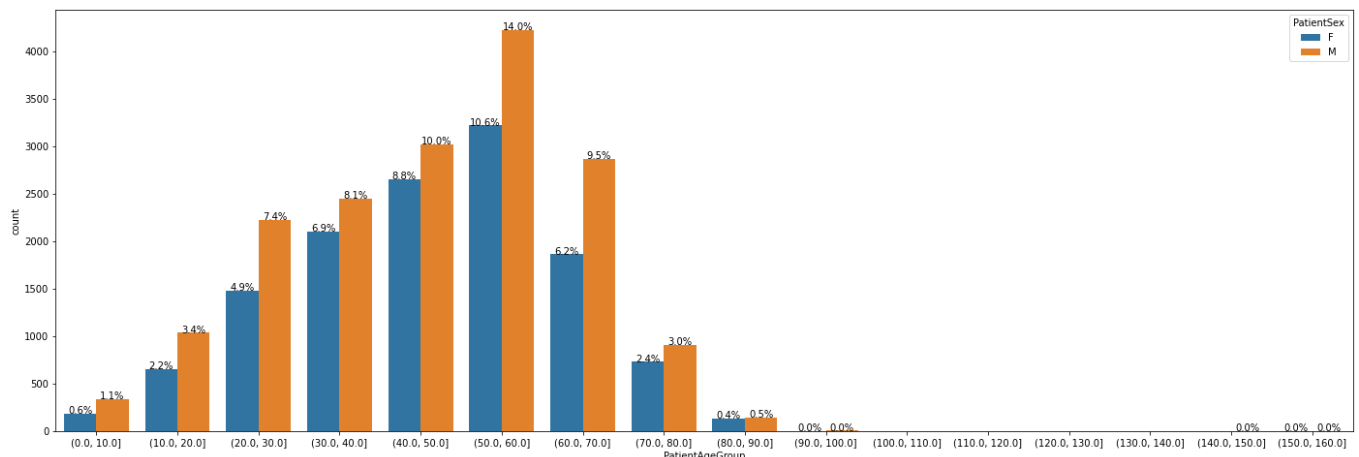




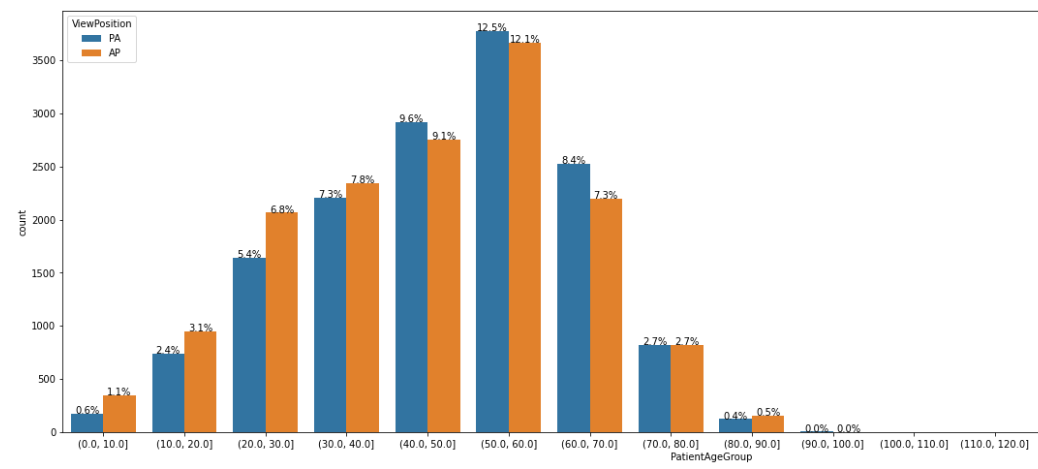
* The above picture depicts a distribution of different classes across the ‘PatientAgeGroup’.
* As can be seen patients with age group between 50-60 are having the highest probability of getting pnenumonia compared to other age groups.
* There are very less no of data points for age group>70.



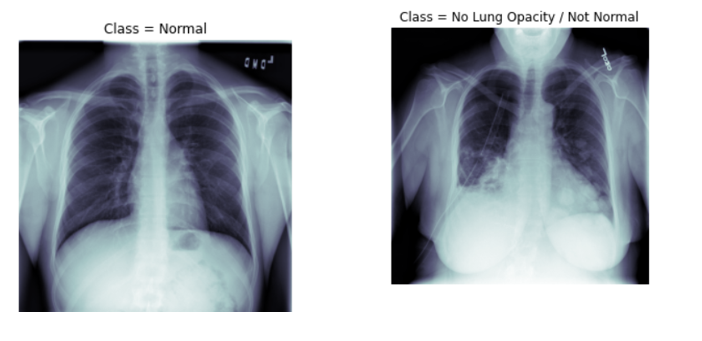
* The distribution of Age group with Target variable also reflects maximum no of postive cases between 40-60. For age group 50-60 there are 17.3% of patients who doesn’t have pneumonia compared to 7.4% who have.



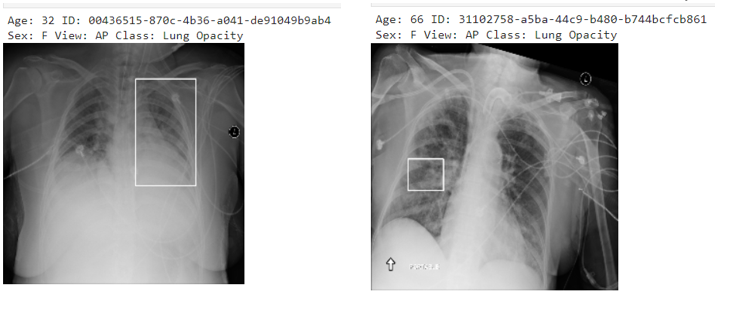
* The distribution of patient sex vs Age group shows there are maximum no of male and female present in age group 50-60 which is 14% and 10.6% respectively.
* Next majority falls between age group 40-50.
* There are very less no of records exists for age group 0-10 and 80-100.



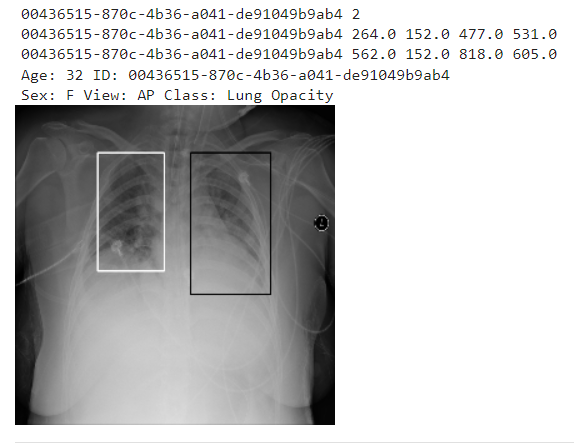
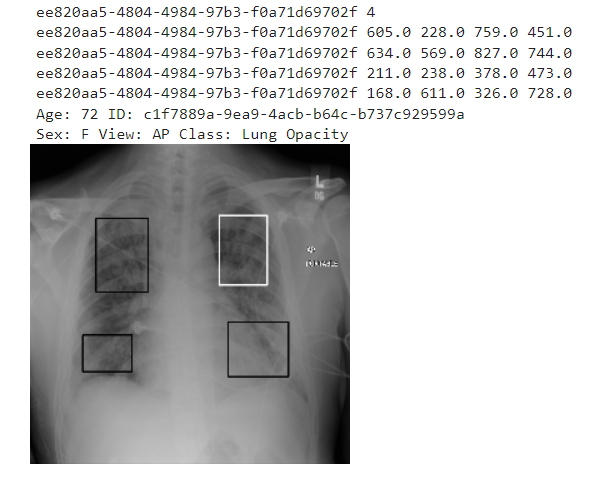
* Above picture illustrate the PA and AP view position for different Age group.
* Data shows for age group 50-60 there are 12.5% patients with AP and 12.1 % with PA position.
* Amongs all the patients between age group 30-40, 9.6% are having AP and 9.1% are having PA view positions.
* For agegroup 0-10 and 80-100 we have very less data available.
* Before proceeding, it would be good to view how the images are displayed. So loading a few DICOM images in google Colab. We will read images from training samples for both normal as well as for pneumonia patients as shown below.



* The above two images represents classes **‘Normal’** and ‘**No Lung Opacity / Not Normal’.** In the normal class the image is quite prominent with no sign of any lungs opacity.
* The second image signifies a patient with **No Lung Opacity / Not Normal’.** Even though it doesn’t have any Lung Opacity, still some portion of the image is blurred giving a notion of Not normal lungs. These sort of cases will further be validated by medical practitioners.



* Above two screenshot shows patients having lungs opacity. As can be seen there are bounding box on each of the images pointing the infections. There will be region of infection based on which the bounding box coordinates have been defined.

* There are also patient id’s which are duplicated or having multiple bounding boxes. This is valid as the patients might have multiple area with infections.
* As shown above the first image shows two bounding boxes/lungs opacity for the same patient id.
* Similarly for the second image there are 4 regions of infection for the same patient.

## 2.4 Summary:

* The training dataset (both of the csv files and the training image folder) contains information of 26684 patients (unique)
* Out of these 26684 unique patients some of these have multiple entries in the both of the csv files
* Most of the recorded patient belong to Target = 0 (i.e., they don't have Pneumonia)
* Some of the patients have more than one bounding box. The maximum being 4
* The classes "No Lung Opacity / Not Normal" and "Normal" is associated with Target = 0 whereas "Lung Opacity" belong to Target = 1
* The images are present in dicom format, from which information like PatientAge, PatientSex, ViewPosition etc are obtained
* There are two ways from which images were obtained: AP and PA. The age ranges from 1-155 (which were further clipped to 100)
* The centers of the bounding box are spread out over the entire region of the lungs. But there are some centres which are outliers.

# **Pre-Processing**

This section describes the pre-processing steps applied to data before modelling. The images are in dicom format which contains lot of metadata along with pixel data. The pixel data needs to be extracted and converted to either jpg or png format.

## 3.1 Pre-processing Methods

The following pre-processing methods can be applied to images.

* Conversion of image to jpg or png format.
* Find the number of channels in images and align to 1 or 3 channels.
  + Convert images to gray scale.
* Image resizing required as per base model requirements like 224\*224 for VGG16.
* Drop duplicate data.
* Set null values to 0 or drop the rows.
* Pixel normalization i.e scale the pixel values.
* Convert the pixel values to float.
* Image transformations like,
  + Rotate.
  + Flip vertically or horizontally.
  + Generate masks for the image.
  + Thresholding.
  + Erosion, Dilation.
  + Cropping.
  + Translation.
  + Noise addition. Etc.

## 3.2 Pre-processing Applied

The data generators are used to pre-process the image. The images are resized to 224\*224 and processed in batches of 32.

Duplicate rows are dropped from merged data frame “train\_feature\_engineered”.

The total number records after dropping duplicates are 26684.

The distribution of target variable and classes are given below.

Distribution of target and classes

0 20672

1 6012

Name: Target, dtype: int64

No Lung Opacity / Not Normal 11821

Normal 8851

Lung Opacity 6012

Name: class, dtype: int64

The shape of data after split into train, test and validation are as below.

Shape of the dataframes:

TRAIN:(21348, 3)

VALID:(2668, 3)

TEST:(2668, 3)

The data distribution is proper across train, test and validation data set.

Distribution of target in the training set:

0 0.78

1 0.22

Name: Target, dtype: float64

Distribution of target in the validation set:

0 0.78

1 0.22

Name: Target, dtype: float64

Distribution of target in the test set:

0 0.77

1 0.23

Name: Target, dtype: float64

|  |  |
| --- | --- |
|  |  |
| TransformationsSummary**Model and Model building**4.1 Model Approach4.2 Model Evaluation Metrics4.3 Model creation4.4 Model Summary |  |

# **Improve Model Performance**

## Approaches

## Summary