AIML CAPSTONE PROJECT ON PNEUMONIA DETECTION

Abstract

Pneumonia is one of the common respiratory problems, considered as inflammation in lungs due to viral and other bacterial infection. Pneumonia patients suffer from serious breathing problems and the infection in lungs can also lead to other health problems such as fever and cough. The treatments of this disease depend on the severity. Doctors diagnose this disease based on the symptoms and observing the breathing pattern. The bacterial pneumonia can cause severe unrest compared to viral one. Hence a chest X-Ray followed by optional blood tests are recommended for proper diagnosis and subsequent medication. In this aspect Machine learning and deep learning models play an important role as it would be difficult for the doctors to read and infer the outcome from the X-ray images. In addition, the factors of external influence such as Lung cancer, fluid overload (pulmonary edema), bleeding, volume loss (atelectasis or collapse), or post-radiation or surgical changes appear as increased opacity on CXR make it more difficult for optimal diagnosis accurately. The aspect of sensitivity of CXR with respect to the positioning of the patient and depth of inspiration cannot be ignored. Most importantly, the doctors and medical community are faced with reading high volumes of images every shift on a regular basis and hence a smart system such as machine learning based diagnostics help to improve the efficiency and reach of diagnostic services. Sparse material tissues like lungs appear black in X-ray images as they don't absorb X-rays. Dense tissues like bones absorb X-rays and appear white. Pneumonia manifests as an area of opacity in X-ray.

In the present work the dataset contains X-ray images of patients' lungs. Here we develop models for both segmentation techniques to detect lung opacities on chest X-ray film images, i.e., i.e., to locate the position of inflammation in the DICOM image. The present work aims at developing CNN architectures to detect lung opacities on chest X-ray film images and subsequently building classifiers pertaining to this disease. The subsequent part of the work aims at applying transfer learning, fine tuning the parameters and improving the model accuracy architectures for building classifiers pertaining to this disease.

1. Introduction:

Respiratory infections are found to be prominent cases for hospitalization, for example, in Iraq they represent 60% mean consultations corresponding to 45% of the average population hospitalized [1]. On the other hand, 22-42% of adult pneumonia patients require hospitalization and 5-10% of them require ICU [2].

Considering Among adults suffering from pneumonia, it is estimated that between 22 and 42% require hospitalization and between 5 and 10% need an intensive care unit, and the lethality varies between 5 and 50% depending on the severity of the condition, which is higher in the elderly and immunosuppressed patients [2]. According to UNICEF data [3], pneumonia is increasingly becoming a life threatening disease among children compared to other diseases among them claiming the lives of over 700,000 children under five every year, or around 2,000 every day. South Asia is prominent as 2,500 cases per 100,000 children are noticed by West and Central Africa (1,620 cases per 100,000 children).

Why do we need to solve it?

Pneumonia accounts for over 15% of all deaths of children under 5 years old internationally. In 2015, 920,000 children under the age of 5 died from the disease. Pneumonia accounts for over 500,000 visits to emergency departments in the United States. There were 50,000 deaths in 2015 keeping the ailment on the list of top 10 causes of death in the United States.

AIML CAPSTONE PROJECT ON PNEUMONIA DETECTION

Why can't doctors themselves handle the issue?

Well, reading the X-ray/CXR is a complicated thing because of following reasons:

1. Lung cancer, fluid overload (pulmonary edema), bleeding, volume loss (atelectasis or collapse), or post-radiation or surgical changes appears as increased opacity on CXR.
2. Outside of the lungs, fluid in the pleural space (pleural effusion) also appears as increased opacity on CXR.
3. Positioning of the patient and depth of inspiration can alter the appearance of the CXR
4. Clinicians are faced with reading high volumes of images every shift.

To improve the efficiency and reach of diagnostic services, an automated Pneumonia detection system is very much necessary.

In the present work, our objective is to build a pneumonia detection system i.e., to locate the position of inflammation in the DICOM image. In other words, to build an algorithm that needs to automatically locate lung opacities on chest radiographs. We have developed CNN model to detect lung opacities on chest X-ray film images. The subsequent part of the work aims at applying transfer model learning, fine tuning the parameters and improving the model accuracy architectures for building classifiers pertaining to this disease. The dataset contains X-ray images of patients' lungs.

The data description for the present problem is provided below:

- In the dataset, some of the features are labeled “Not Normal No Lung Opacity”. This extra third class indicates that while pneumonia was determined not to be present, there was nonetheless some type of abnormality on the image and oftentimes this finding may mimic the appearance of true pneumonia. Dicom original images: - Medical images are stored in a special format called DICOM files (*.dcm). They contain a combination of header metadata as well as underlying raw image arrays for pixel data.
- Dataset has been attached along with this project. Please use the same for this capstone project.
- Original link to the dataset: https://www.kaggle.com/c/rsna-pneumonia-detection-challenge/data

The methodology used and the associated tasks performed are described in the figure 1 below:

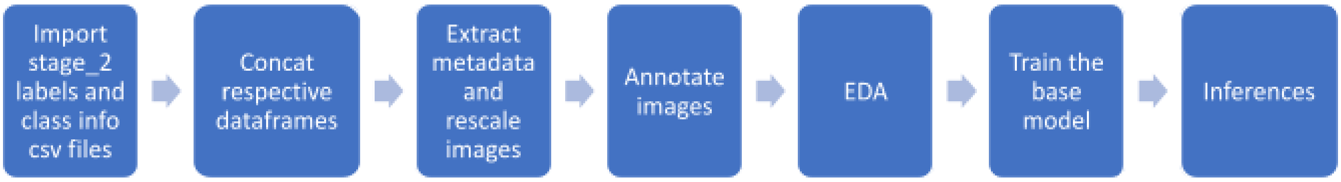

Import stage_2 labels and class info csv files
Concat respective dataframes
Extract metadata and rescale images
Annotate images
EDA
Train the base model
Inferences

Figure 1: Pneumonia segmentation and prediction using deep learning methods

We took csv files and imported them first. We performed EDA and verified for null values. However, there are records with multiple pneumonia presence, so we concatenated files to create a dataframe. Metadata from the images is taken and the information so extracted is used as columns in the dataframe.

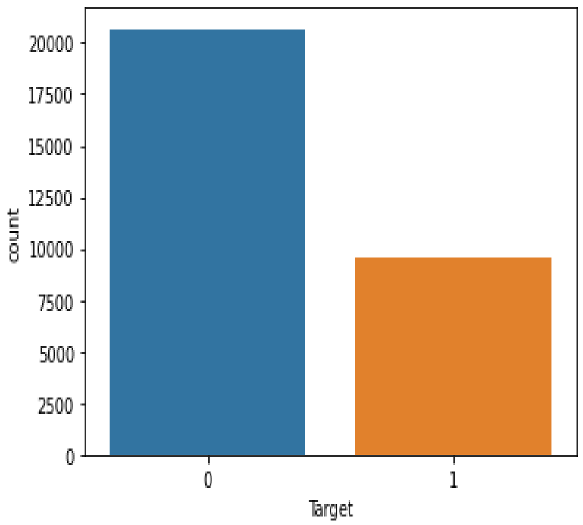
The csv data comprises patient IDs followed by associated bounding box attributes such as x, y, height, width and associated target label (pneumonia :1 its absence:0). The shape of the data frame is `(30227, 6)`

2. Exploratory Data Analysis:

The csv file 'stage_2_train_labels.csv' contains the patient id , the bounding box coordinates (x, y, width and height) along with the Target column consisting of values 0 and 1. There are 30,227 records in this file. The other file is 'stage_2_detailed_class_info.csv' which contains the patient id and the classes: **No Lung Opacity/Not Normal, Normal, Lung Opacity.** The following are the pre- processing / EDA analysis that has been performed on the data :

Out of (30227, 6), the total number of null values in bounding boxes columns is equal to the total number of 0s in the Target column. The class distribution of presence of pneumonia (class label 1);

its absence (class label 0) are: {0: 20672, 1: 9555}. These studies indicate that all the records with class label 0 do not contain bounding boxes as expected because the bounding boxes are indicative of opacity in the images.

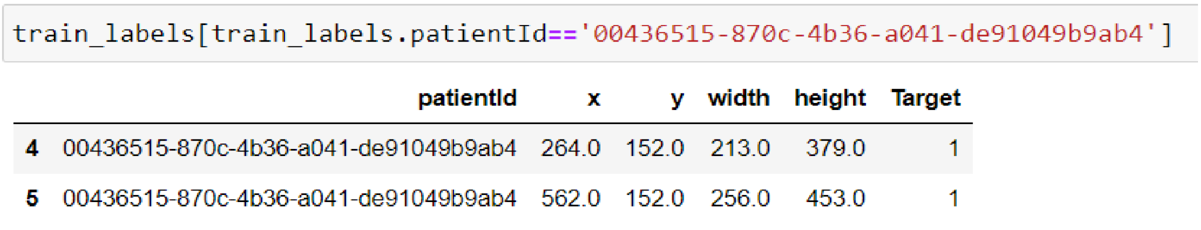


Secondly, one can notice that the instances with class 0 are more than double that of class 1 instances which can be considered as class imbalance problem and to be handled as part of classification task during subsequent phases of this work. One can notice the relative numbers from the adjacent count plot

Figure 2: Distribution of the patients across classes 0, 1

From the above data, we can notice that there are around `30227` rows in `stage_2_train_labels.csv` but we have a total of `26684` unique patientIDs. Hence `3543` patients have more than one bounding box.

For example, the patientId: `00436515-870c-4b36-a041-de91049b9ab4` is the same from 4th and 5th rows of the dataframe:



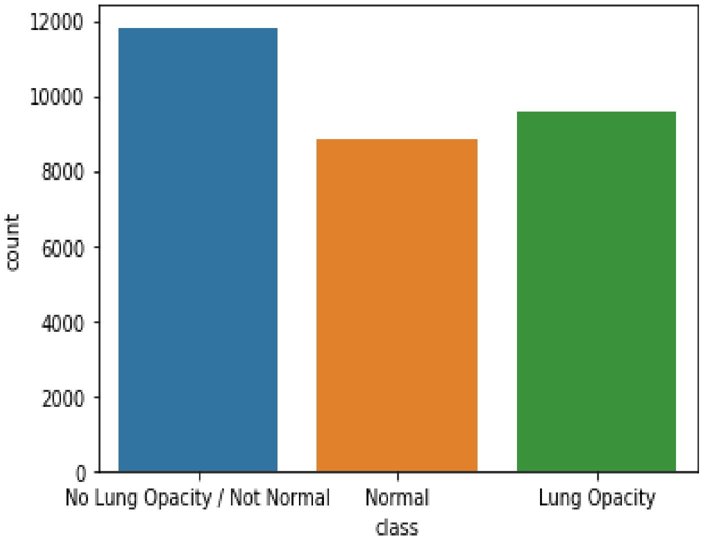
```
train_labels[train_labels.patientId=='00436515-870c-4b36-a041-de91049b9ab4']
```

	patientId	x	y	width	height	Target
4	00436515-870c-4b36-a041-de91049b9ab4	264.0	152.0	213.0	379.0	1
5	00436515-870c-4b36-a041-de91049b9ab4	562.0	152.0	256.0	453.0	1

On the other hand, it is observed that the total number of patients match with the value from training labels is `26684` and the associated classes are: `Normal,Lung Opacity and No Lung Opacity/Not Normal`

The relative population of these classes can be noticed from the count plot below:

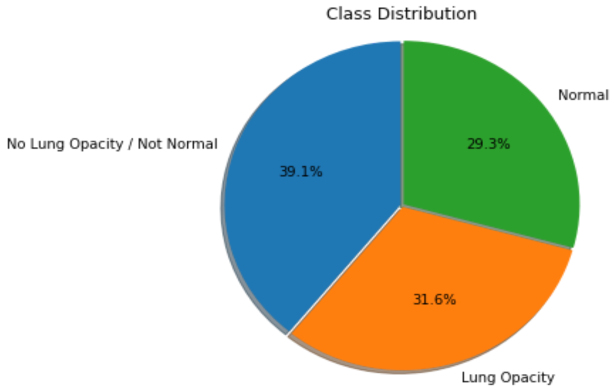
From the countplot, it is very much noticeable that all the classes are almost equally distributed. There are no missing values in the detailed_class_info data



It is observed that total number of null values for "Normal" and "No Lung Opacity / Not Normal" is equal to total number of 0s in Target column i.e., who do not have pnemonia have their bounding box columns as NaN.

Figure 3: Distribution of the patients based on the opacity information from the images.

<u>% of population across opacity groups:</u>



Hence the first two categories are labelled as 0 while the class with Lung Opacity as 1. The resulting distribution can be seen as given in the figure below:

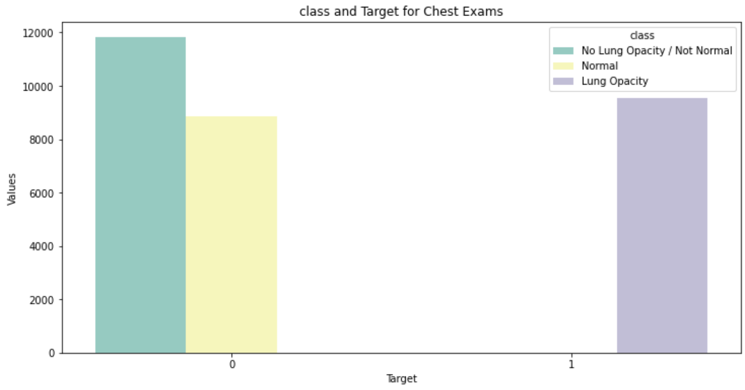


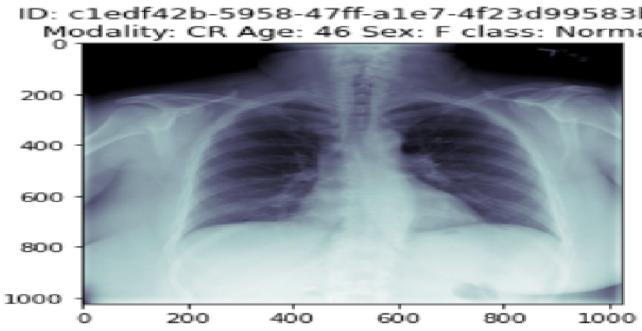
Figure 4: Distribution of the patients from the class labels data

Observations regarding the images and their size:

- Images are stored in DICOM(Digital Imaging and Communications in Medicine) format with .dcm extension. This DICOM image contains more information like patient age, sex, modality, view position and body part and so on.
- `pydicom` library is used to read the images
- Each image size is 1024 x 1024 which requires scaling before employing machine learning models.
- "BodyPartExamined" is CHEST which is expected
- "Modality" is CR (Computer Radiography)

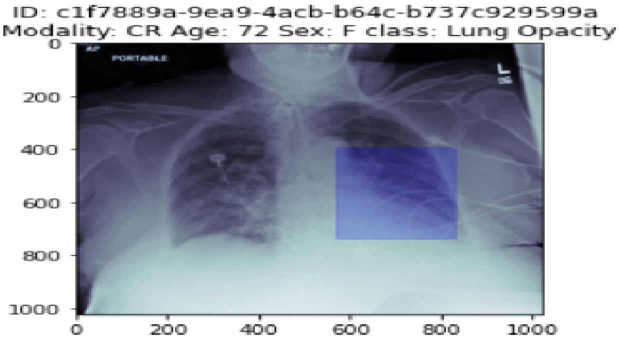
The features '`Age`', '`Sex`', '`BodyPartExamined`' are dropped from the data-frame for further processing as they are not significant for the analysis under consideration. In the next step the metadata information such as Age, gender is mapped for each of the corresponding images and a couple of examples are furnished below considering two random patient ids. Here the function: `MapImagesToAnnotations(patientID)` takes patientID as argument and displays corresponding image and associated patient information:

```
MapImagesToAnnotations('c1edf42b-5958-47ff-a1e7-4f23d99583ba')
```

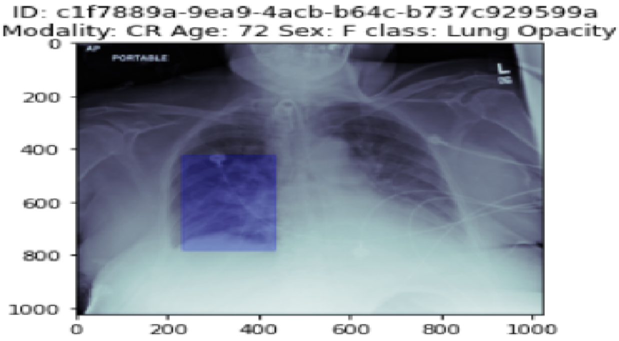


```
ID: c1edf42b-5958-47ff-a1e7-4f23d99583ba
Modality: CR Age: 46 Sex: F class: Normal
```

```
MapImagesToAnnotations('c1f7889a-9ea9-4acb-b64c-b737c929599a')
```



```
ID: c1f7889a-9ea9-4acb-b64c-b737c929599a
Modality: CR Age: 72 Sex: F class: Lung Opacity
```



```
ID: c1f7889a-9ea9-4acb-b64c-b737c929599a
Modality: CR Age: 72 Sex: F class: Lung Opacity
```

Figure 5: Annotated images with Non -opacity and opacity regions (blue) observed.

We have performed detailed analysis on the gender v/s classes. It can also be seen that the number of male patients are more than that of female across the label categories as depicted from the count plot below:

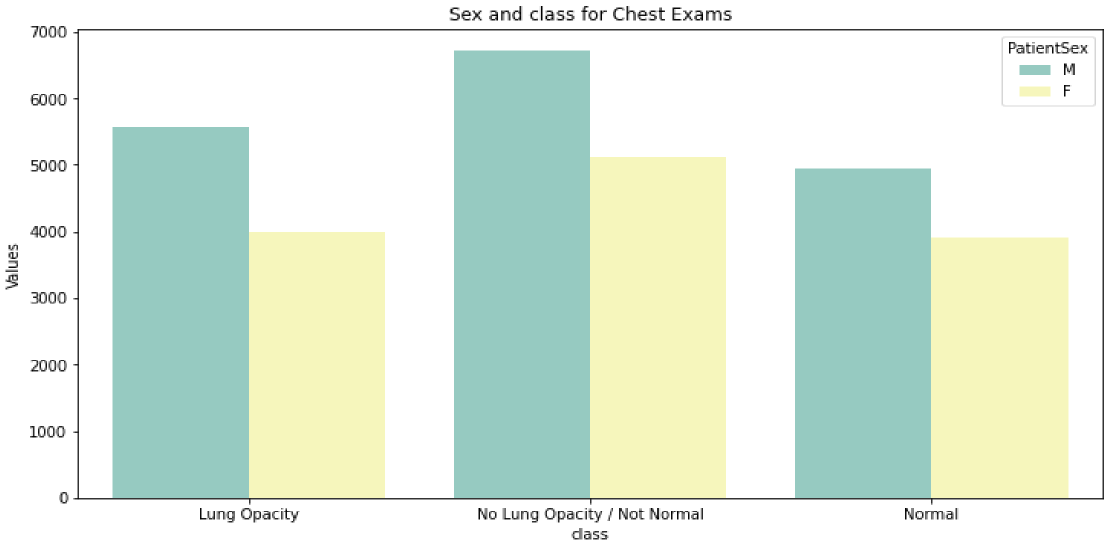


Figure 6: Distribution of the patients from the class labels data

Outlier analysis is performed with respect to the patient age attribute and the opacity. The associated box plots are furnished in the Figure 7 below:

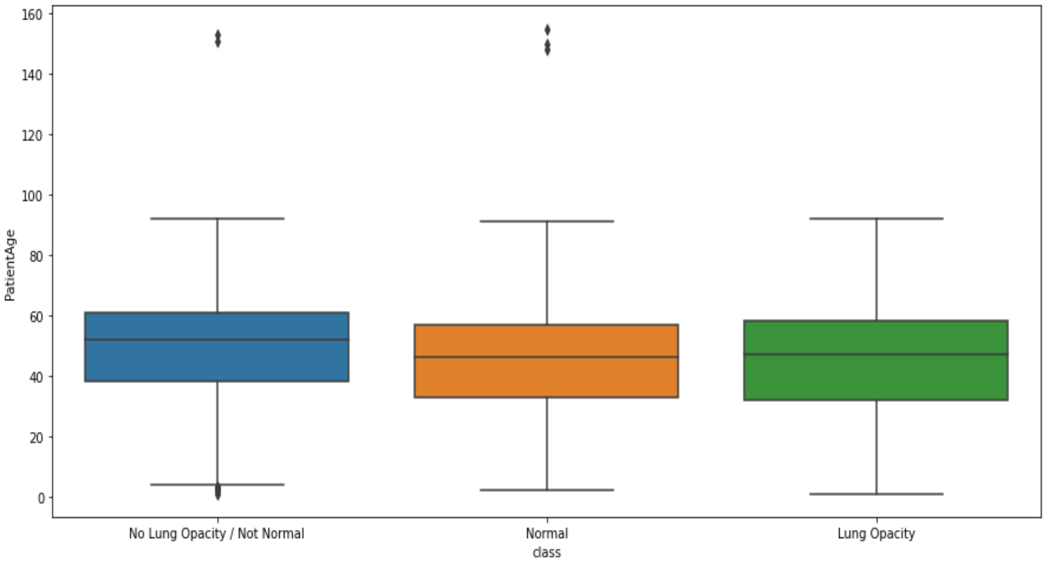
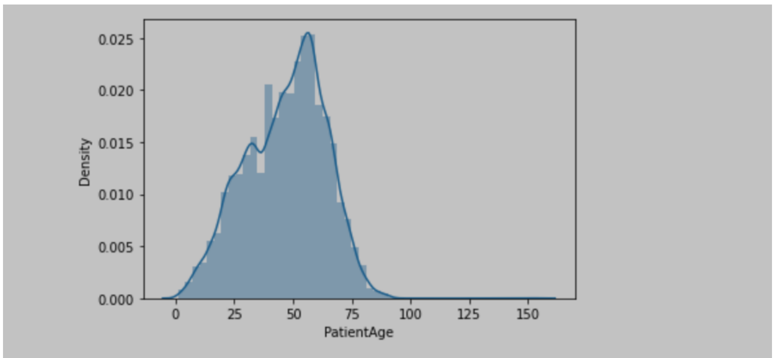
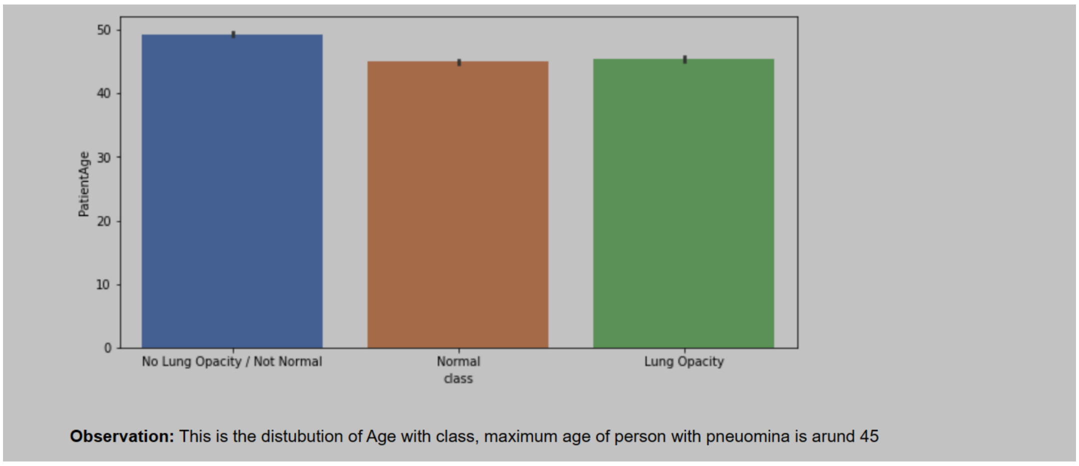


Figure 7: Box plots for opacity distributions



Observation: Looks like normal distubution of age



Observation: This is the distubution of Age with class, maximum age of person with pneuomina is arund 45

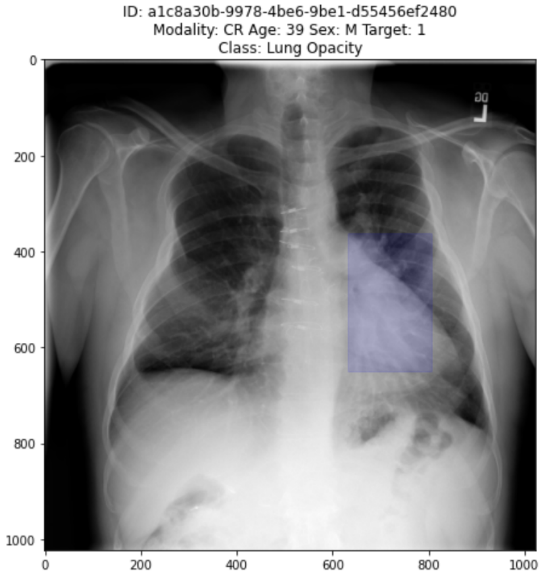
```
Number of patients in age category

<=75      13318
<=50      12157
<=26       3972
<=100       780
```

It is also observed that:

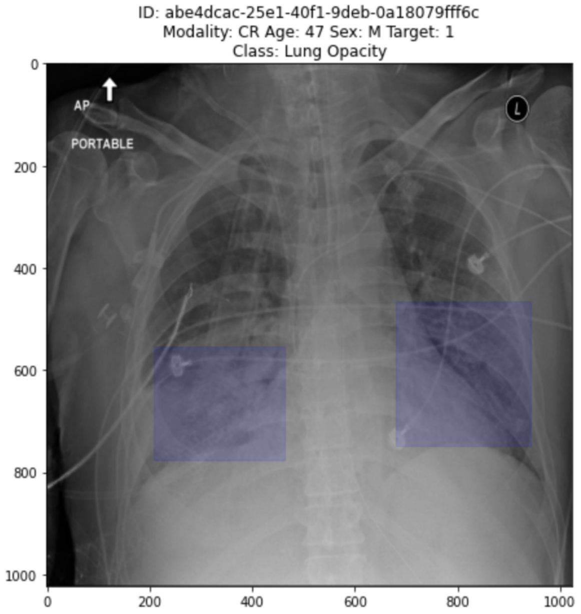
- The mean age is 46 years , whereas the minimum age is 1 year and the max age is 155 which seems to be an outlier.
- 50% of the patients are of around 49 age , the std deviation is 16 which suggest that age is not normally distributed.
- There are 8851 normal cases, people with lung opacity are 9555 and No Lung Opacity / Not Normal are 11821.
- Patients with evidence of Pneumonia are associated with Lung Opacity class and target = 1.
- Patients with no definitive evidence of Pneumonia are either of Normal or No Lung Opacity / Not Normal class and target = 0.

Observations regarding the Bounding boxes

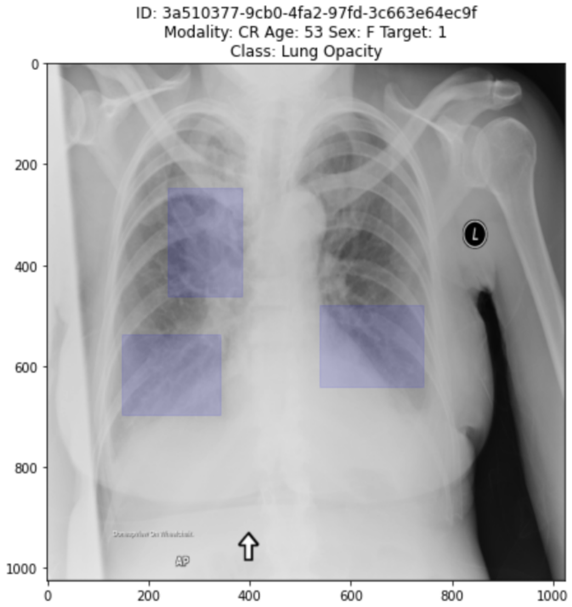
X-ray images with ONE bounding box


ID: a1c8a30b-9978-4be6-9be1-d55456ef2480
Modality: CR Age: 39 Sex: M Target: 1
Class: Lung Opacity
0
200
400
600
800
1000
0
200
400
600
800
1000

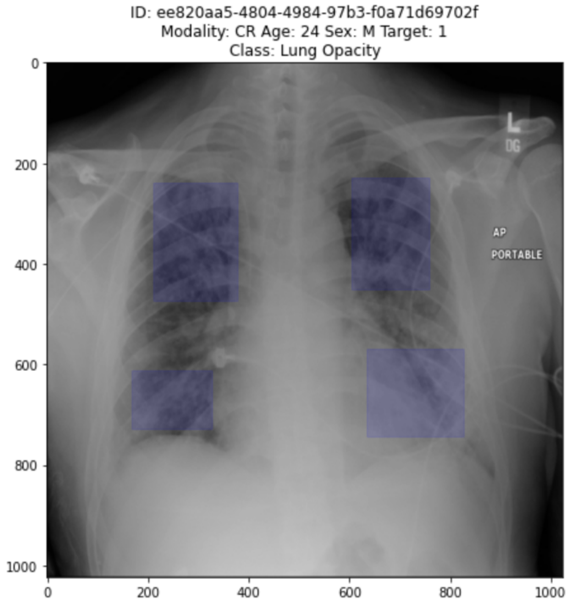
X-ray images with TWO bounding box

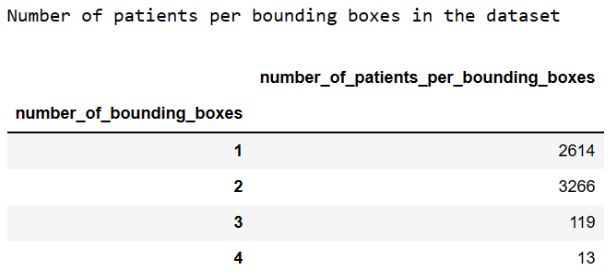

ID: abe4dcac-25e1-40f1-9deb-0a18079fff6c
Modality: CR Age: 47 Sex: M Target: 1
Class: Lung Opacity
AP
PORTABLE
0
200
400
600
800
1000
0
200
400
600
800
1000

X-ray images with THREE bounding box


ID: 3a510377-9cb0-4fa2-97fd-3c663e64ec9f
Modality: CR Age: 53 Sex: F Target: 1
Class: Lung Opacity
AP
0
200
400
600
800
1000
0
200
400
600
800
1000

X-ray images with FOUR bounding box


ID: ee820aa5-4804-4984-97b3-f0a71d69702f
Modality: CR Age: 24 Sex: M Target: 1
Class: Lung Opacity
AP
PORTABLE
0
200
400
600
800
1000
0
200
400
600
800
1000



```
Number of patients per bounding boxes in the dataset
```

number_of_bounding_boxes	number_of_patients_per_bounding_boxes
1	2614
2	3266
3	119
4	13

Observations: Maximum patients have 2 bounding boxes while 13 patients have 4 bounding boxes.

More information from image metadata from the exploratory data analysis:

Metadata in the DCIM image:

A single DCIM image was taken from the dataset and the metadata has been displayed. It includes the patient details such as name, age, modality, gender, view position; details about the image itself, date and time among other parameters. Additional parameters (age, gender, view position, pixel spacing) have been appended to the amalgamated data frame for analysis.

Data frame attribute analysis:

Gender: The dataset consists of a higher percentage of male patients compared to females.

Patient age*:* The histogram plot indicates that there is a greater representation of patients in the 40- 60 age range.

ViewPosition: It indicates if the x-ray is taken from posterior or anterior position.

Duplicate records (records with more than one X-ray):

There are 3543 duplicate entries in our dataset.

Resizing the images:

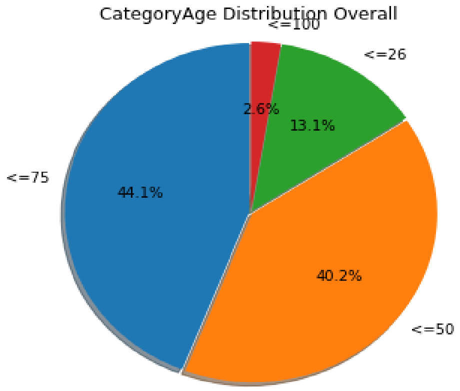
All of the images are of the size 1024 * 1024. this size might slow down the model building process hence we resize the image to a new size of 128*128.

Based on these features following visualizations are generated:

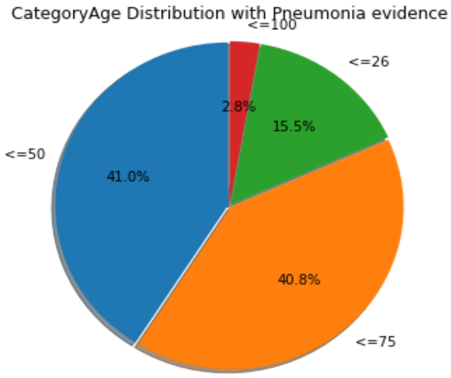
It is observed that

- Numbers of patients are highest for age group 51-75 for overall case
- But the number of patients is highest for the age group 27-50 for Target=1 case.

Age distribution in the dataset:

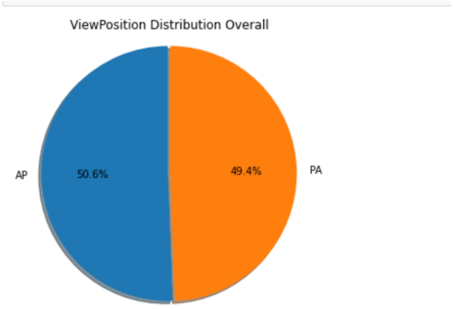


Age distribution of patients with Pneumonia

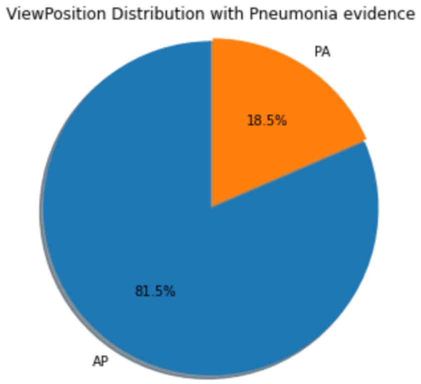


- **Posterior/Anterior (PA)**: X-ray is taken from the back part of the chest. So it hits the posterior part before the anterior part. Patient needs to stand against the X-ray machine.
- **Anterior/Posterior (AP)**: X-ray is taken from the front part of the chest. So it hits the anterior part before the posterior part. This is taken when Patient cannot stand against an X-ray machine but heart size is exaggerated.

View position distribution (overall case)



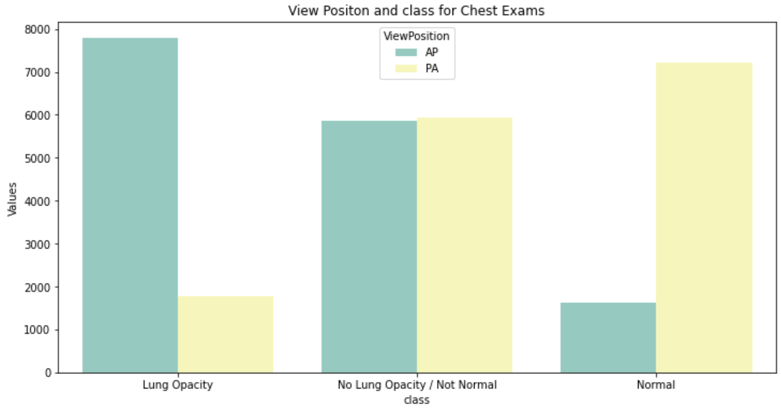
View position distribution with pneumonia evidence



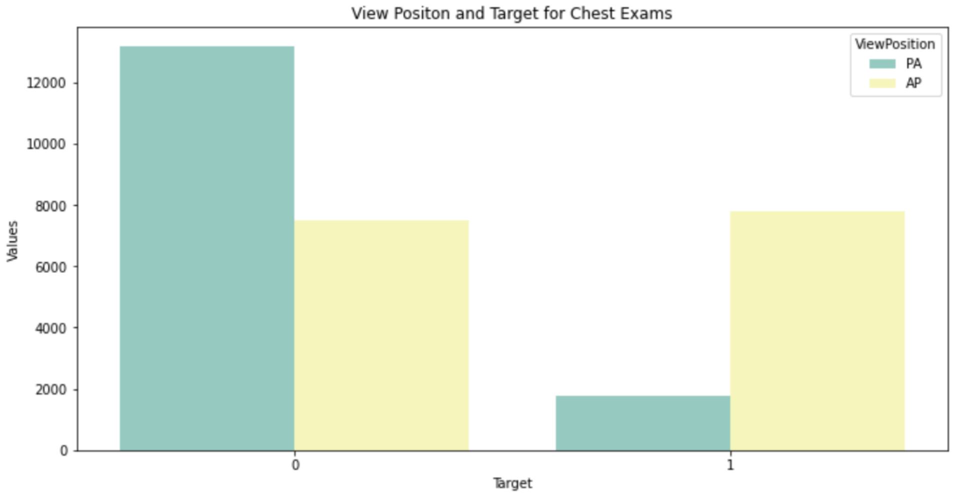
Observation:

1. ViewPosition = PA and ViewPosition = AP are almost equally distributed in the overall train data
2. When Target=1, View position = AP dominates.

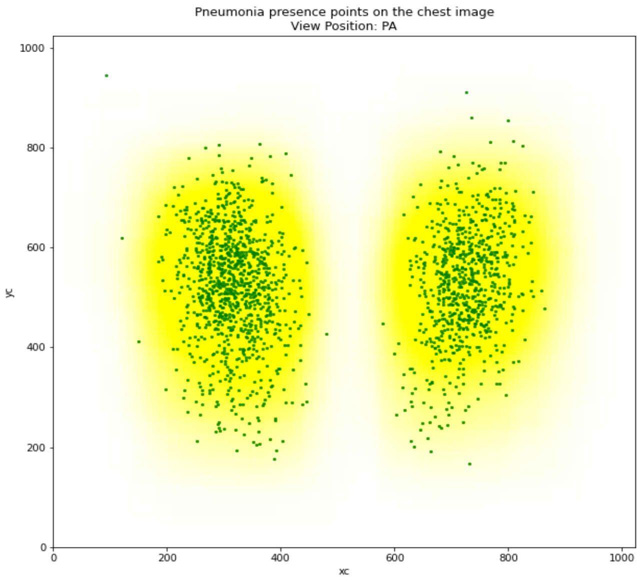
View positions and various states of opacity

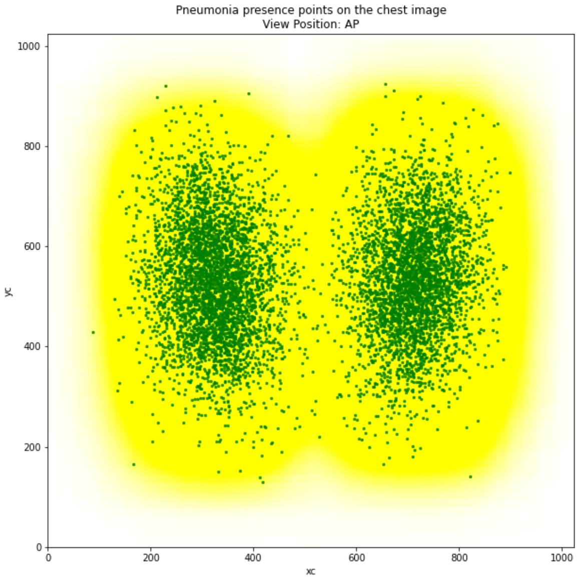


View positions and Target



Pneumonia present points on chest X-Ray





Scatter plots show the concentration of pneumonia affected regions in lungs. It is seen that pneumonia affects the central part of the lungs more than at the boundaries. Since there are more patients with view position = AP, we can see more scatter points in 2nd diagram

Model Building:

We are creating a classification model to classify using images if a patient has pneumonia or not.

We are using a sample of 14,000 images with 7,000 data points having target '0' and 7,000 data points with target '1'.

CNN model:

After preprocessing the data we resize the images in 128*128 format . Add 3 channels to images.

We have 3 convolutional layers with "relu" activation function.

We have 1 DNN layer and 1 output layer with "softmax" activation function and 0.5 dropout.

Optimizer= Adam with a learning rate of 0.001.

Metrics=accuracy

Epochs=10

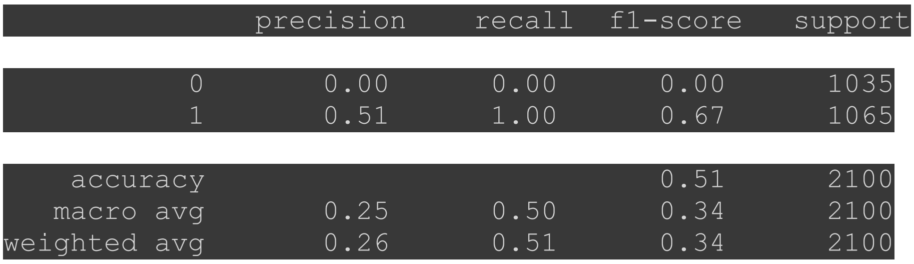
Loss function= binary crossentropy.

Results of CNN model

ACCURACY=51.71%

LOSS=0.62

CLASSIFICATION REPORT:



```
              precision    recall  f1-score   support

           0       0.00      0.00      0.00      1035
           1       0.51      1.00      0.67      1065

    accuracy                           0.51      2100
   macro avg       0.25      0.50      0.34      2100
weighted avg       0.26      0.51      0.34      2100
```

We can see that precision and recall is 0 for class 0 which means the model is not able to predict class 0.

Densenet

As we found out that the sequential model is not able to detect class 0 we are using dense net for better results.

A DenseNet is a type of convolutional neural network that utilises dense connections between layers, through Dense Blocks, where we connect all layers (with matching feature-map sizes) directly with each other. To preserve the feed-forward nature, each layer obtains additional inputs from all preceding layers and passes on its own feature-maps to all subsequent layers.

Optimizer= Adam with a learning rate of 1e-4.

Metrics=accuracy

Epochs=10

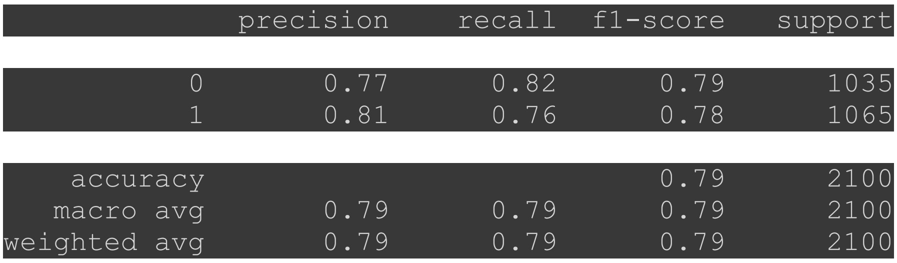
Loss function= binary cross entropy.

Results of CNN model

ACCURACY=79.49%

LOSS=0.45

CLASSIFICATION REPORT:

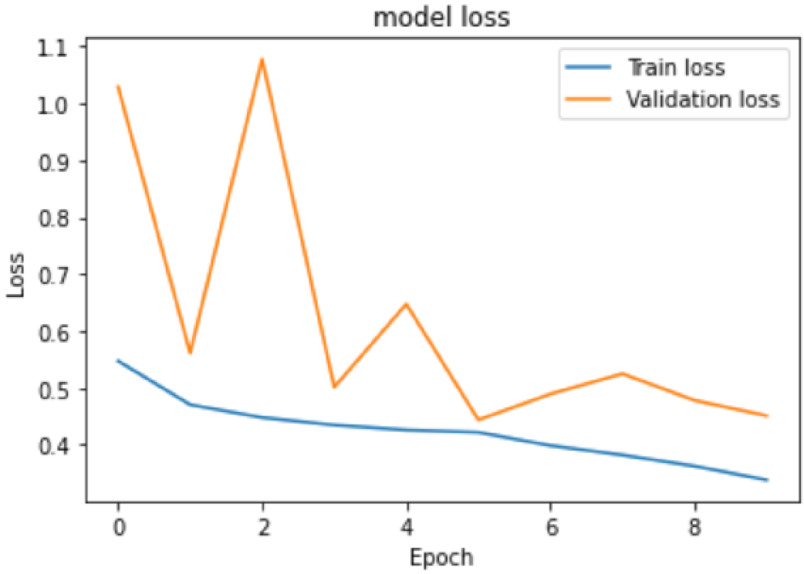


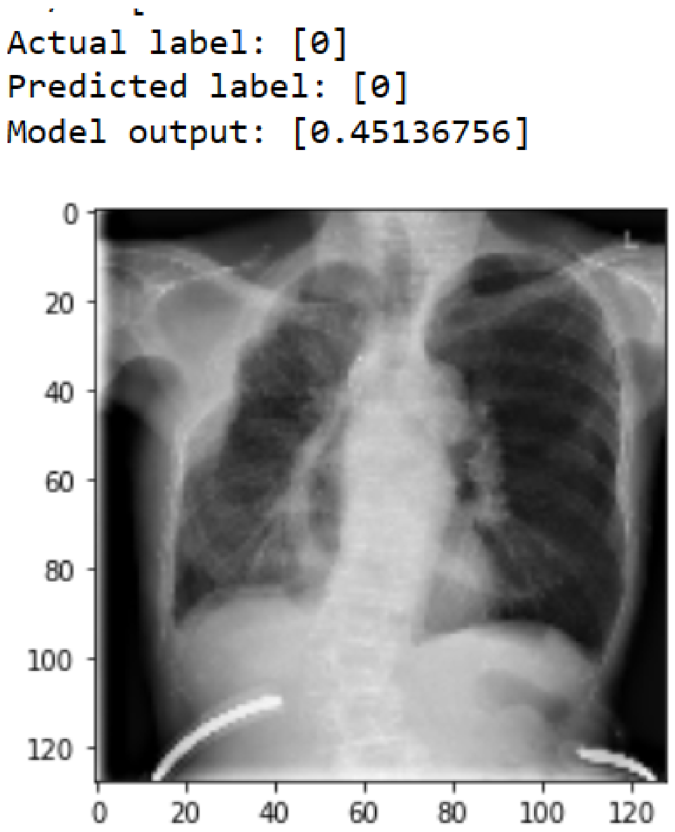
```
              precision    recall  f1-score   support

           0       0.77      0.82      0.79      1035
           1       0.81      0.76      0.78      1065

    accuracy                           0.79      2100
   macro avg       0.79      0.79      0.79      2100
weighted avg       0.79      0.79      0.79      2100
```

here we can see that accuracy has increased to nearly 80% and we are able to detect class 0 better.

Precision for class 0 is 0.77 and class 1 is 0.81

As precision and recall are important measures in the medical field, increasing that means our model is performing better now.





```
Actual label: [0]
Predicted label: [0]
Model output: [0.45136756]
```

So our model has predicted the label correctly.

References:

1. F. Lami, H. Rashak, H. A. Khaleel et al., “Iraq experience in handling the COVID-19 pandemic: implications of public health challenges and lessons learned for future epidemic preparedness planning,” Journal of Public Health, vol. 43, no. Supplement_3, pp. iii19–iii28, 2021

2. V. Sirish Kaushik, A. Nayyar, G. Kataria, and R. Jain, “Pneumonia detection using convolutional neural networks (CNNs),” Lecture Notes in Networks and Systems, pp. 471–483, 2020

3. Pneumonia in Children Statistics - UNICEF DATA (accessed on 5th Jan, 2023)

Acknowledgements:

We have explored the additional information provided at Kaggle web site, available at:

https://www.kaggle.com/c/rsna-pneumonia-detection-challenge/overview/acknowledgements during the present work.