Sailesh , Karthik, Krishna Kumar

AIML Online CAPSTONE

Pneumonia Detection

Contents

[Project Description: 2](#_Toc45481199)

[Business Value: 2](#_Toc45481200)

[Dataset 2](#_Toc45481201)

[Data Description: 4](#_Toc45481202)

[stage\_2\_detailed\_class\_info: 4](#_Toc45481203)

[Data Fields: 4](#_Toc45481204)

[stage\_2\_train\_labels: 4](#_Toc45481205)

[Data Fields: 4](#_Toc45481206)

[Exploratory Data Analysis (EDA) or Data Understanding 5](#_Toc45481207)

[Exploring the Class data 5](#_Toc45481208)

[Exploring the Target data 6](#_Toc45481209)

[Exploring the Discom Meta data 7](#_Toc45481210)

[Exploring the Discom image data 11](#_Toc45481211)

[Our Approach 12](#_Toc45481212)

[MobileNet 12](#_Toc45481213)

[Baseline Model 12](#_Toc45481214)

# Project Description:

**Pneumonia** is a one of the common acute respiratory lung disease. Pneumonia is a bacterial or virus-related or fungal infection. This can affect one or both sides of the lungs that causes the tiny air sacs (alveoli), of the lungs to fill up with fluid or pus. Symptoms can be mild or severe and could include cough, fever, and trouble breathing. The extent of seriousness of pneumonia depends on factor, such as the type of germ causing the lung infection, age, and overall health.

The WHO estimate that around 1.4 millions of children lost their life due to the failure of detecting this lung disease at its early stage. In 2015, 920,000 children under the age of 5 died from the disease. In the United States, pneumonia accounts for over 500,000 visits to emergency departments and over 50,000 deaths in 2015, keeping the ailment on the list of top 10 causes of death in the country.

To diagnose pneumonia, doctor / highly trained specialists review medical history, perform a physical exam, and other demand diagnostic tests. Chest X-ray (CXR) diagnosis is performed to find the infection in the patient's lungs and depth of it’s spread. This information helps doctor / specialist determine the type of pneumonia.

The diagnosis of pneumonia on CXR is complicated because of a number of other conditions in the lungs such as fluid overload (pulmonary edema), bleeding, volume loss (atelectasis or collapse), lung cancer, or post-radiation or surgical changes. In addition to the above complexity, clinicians are faced with reading high volumes of images every shift.

# Business Value:

* Automating Pneumonia screening in chest radiographs, providing affected area details through bounding box.
* Machine – human collaborated effort by assisting physicians to make better clinical decisions or even replace human judgment in certain functional areas of healthcare.
* Combined Machine- Human decision making helps in improved efficiency and additional capacity of clinical decision making.

# Dataset

This capstone project leverages the data from Kaggle repository titled RSNA Pneumonia Detection. The dataset composes of three class as No Lung Opacity / Not Normal, Normal & Lung Opacity and two target classes which are normal lung and pneumonia lung.

The dataset consists of two main zip files (i.e., training and validation) and two csv file containing class and target details of patient with pneumonia (1) and normal (0). A total of 26,684 X-ray images comprising of anterior-posterior (AP) and PA - Posterior to Anterior chests.

The dataset composes of two classes which are normal lung and pneumonia lung as can be seen in the figure below.

|  |  |
| --- | --- |
| Sample image with no pneumonia and No Lung Opacity / Not Normal | Sample image with no pneumonia and Normal |
|  |  |

|  |
| --- |
| Sample image with pneumonia and Lung Opacity |
|  |

# Data Description:

## stage\_2\_detailed\_class\_info:

Data set contains information of 26684 Patients with classification on the Inflammation as No Lung Opacity / Not Normal, Normal and Lung Opacity

### Data Fields:

* PatientId - Each patientId corresponds to a unique image.
* Class – the tertiary classification, indicating whether the No Lung Opacity / Not Normal, Normal and Lung Opacity

## stage\_2\_train\_labels:

Data set contains 30227 rows and 26684 unique Patients and bounding box / target information. This row include multiple bounding boxes.

* For patientIds with no predicted pneumonia , bounding boxes are not available
* For patientIds with a single predicted, one row of bounding box are available
* For patientIds with multiple predicted, multiple rows of bounding boxes are available

### Data Fields:

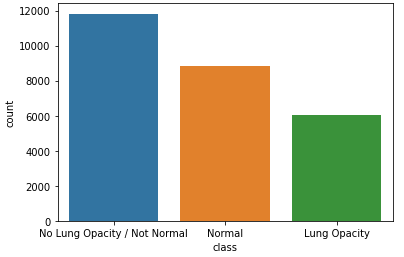
* PatientId - Each patientId corresponds to a unique image.
* x - x coordinate of the bounding box.
* y - y coordinate of the bounding box.
* width - the width of the bounding box.
* height - the height of the bounding box.
* Target - the binary Target, indicating whether this sample has evidence of pneumonia.

# Exploratory Data Analysis (EDA) or Data Understanding

We would explore the data class data and training label data

## Exploring the Class data

The class data contains 30227 rows with 26684 unique patient id. This means there are 3543 duplicate entries. This duplicate entries have been removed.



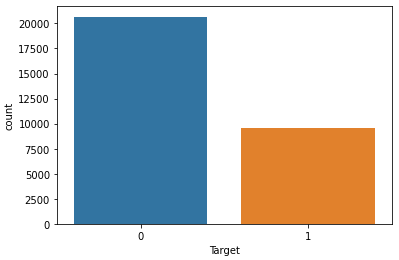
The plot on class shows the split among the different classes of data namely No Lung Opacity / Not Normal, Normal & Lung Opacity

* ~6000 Lung Opacity class
* ~8500 Normal class
* ~12000 No Lung Opacity / Not Normal Class

The distribution of class is biased towards No Lung Opacity / Not Normal & Normal Class

## Exploring the Target data

The target data contains 30227 rows with 26684 unique patient id. There are 9555 rows with pneumonia positive. Out of the 6012 pneumonia positive cases 2614 patients have one of the lung infected. 3266 patients have at have 2 infection. 119 patients have at have 3 infection and 13 have 4 infections.



The plot Target class shows if the patient has Pneumonia or not

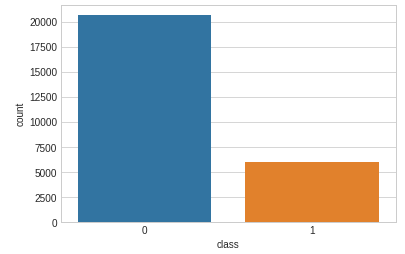
* ~9500 with Pneumonia (~6000 unique patients with Pneumonia)
* ~20000 without Pneumonia

The distribution of Target is biased towards Non Pneumonia cases

After combining the class and target it can be observed that the Lung Opacity cases are Pneumonia positive / true cases other cases are non-Pneumonia cases

## Exploring the Discom Meta data

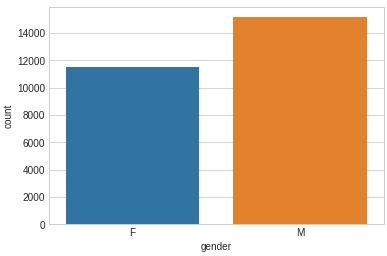
The meta data containing Patient ID, Patient's Birth Date, Patient's Sex, Patient's Age, Body Part Examined, View Position, Rows and Columns was extracted from discom file. This extracted meta data was combined with target for the target data (Pneumonia / Non- Pneumonia) and analyzed.



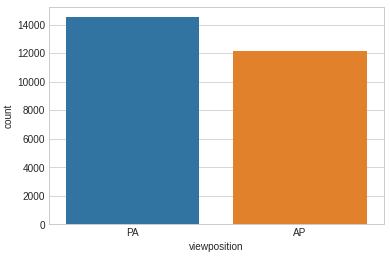
The plot on the Target class distribution based on the unique patient with Pneumonia or not, is as shown below

* ~6000 with Pneumonia
* ~20000 without Pneumonia

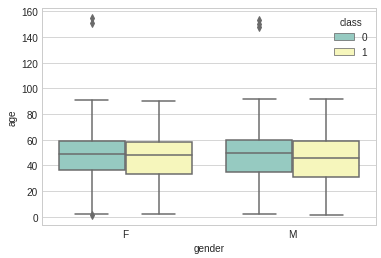
The distribution of Target class is biased towards Non Pneumonia cases



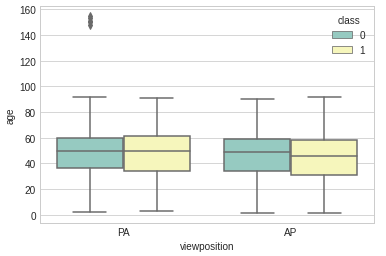
From the gender distribution plot it can be noted thete the data is biased with more male records.



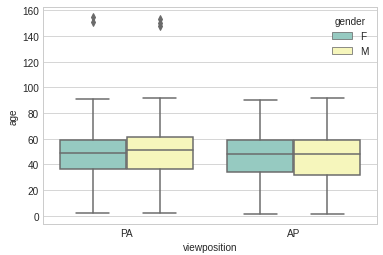
The plot on the position shows there are more PA position images compared to the AP



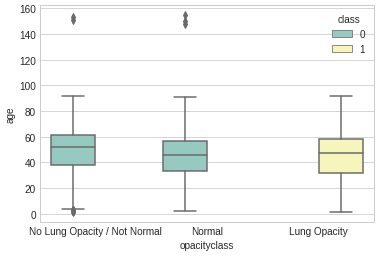
Plot on the age and sex shows the age distribution is between late 30's and 60 for both male and female



From the plot , there is equal proportion of split between pneumonic and non-pneumonic on the position(AP - Anterior to Posterior or PA - Posterior to Anterior) though the target is biased (pneumonic and non-pneumonic)



From the plot , there is equal proportion of split between male and female on the position (AP - Anterior to Posterior or PA - Posterior to Anterior)



From the plot, there is a clear depiction of the data on the lung opacity which are classified as pneumonia

## Exploring the Discom image data

|  |
| --- |
| Sample image with no pneumonia and No Lung Opacity / Not Normal |
|  |

|  |  |
| --- | --- |
| Sample image with no pneumonia and Normal | Sample image with pneumonia and Lung Opacity |
|  |  |

# Our Approach

For this project, we focus on binary classification of images, attempting to classify a particular X-Ray image as having pneumonia or not. Followed by detecting to identify the inflammation of the lungs. Inflammation are highlighted with a bounding boxes around the area of the lung. Image without bounding boxes are negative and contain no definitive evidence of pneumonia. Images with bounding boxes indicate evidence of pneumonia.

There is a strong class imbalance in the dataset, with only about 25% of images labeled as having pneumonia. Images from the dataset are 1024x1024. To begin, we resized each image using filter. Our Deep Learning model (MobileNet) uses 224x224 resolution. We have also standardize the data so that each feature (each pixel) has zero mean and approximately unit variance.

We would also be making an attempt in following models:

* Mask RCNN
* SSD
* YOLO

# MobileNet

MobileNet is an efficient architecture that uses depth wise separable convolutions to construct lightweight deep convolutional neural networks and provides an efficient model for mobile and embedded vision applications. The structure of MobileNet is based on depth wise separable filters, as shown in below.

<< MobileNet Architecture>>

# Baseline Model

As a baseline, we use MobileNet model to classify whether or not a given X-ray contains pneumonia. MobileNet works well as a baseline as it is relatively easy to implement.

<<Base line loss value>>

References:

1. <https://www.kaggle.com/c/rsna-pneumonia-detection-challenge/overview/acknowledgements>
2. <https://www.nhlbi.nih.gov/health-topics/pneumonia#:~:text=Chest%20x%20ray%20to%20look,is%20actively%20fighting%20an%20infection.>
3. <https://biomedpharmajournal.org/vol13no1/a-deep-learning-based-approach-towards-the-automatic-diagnosis-of-pneumonia-from-chest-radio-graphs/>
4. <https://www.hindawi.com/journals/misy/2020/7602384/>