Pneumonia Detection Using Deep Learning

Milestone Report

Capstone Project 2

Springboard Data Science Career Track

Raviteja Bodla, PharmD

Raviteja.bodla@gmail.com

Introduction:

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. In the United States, pneumonia accounts for over 500,000 visits to emergency departments [1] and over 50,000 deaths in 2015 [2], keeping the ailment on the list of top 10 causes of death in the country.

While common, accurately diagnosing pneumonia is a tall order. It requires review of a chest radiograph (CXR) by highly trained specialists and confirmation through clinical history, vital signs and laboratory exams. Pneumonia usually manifests as an area or areas of increased opacity [3] on CXR. However, 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. Outside of the lungs, fluid in the pleural space (pleural effusion) also appears as increased opacity on CXR. When available, comparison of CXRs of the patient taken at different time points and correlation with clinical symptoms and history are helpful in making the diagnosis.

CXRs are the most commonly performed diagnostic imaging study. A number of factors such as positioning of the patient and depth of inspiration can alter the appearance of the CXR [4], complicating interpretation further. In addition, clinicians are faced with reading high volumes of images every shift.

To improve the efficiency and reach of diagnostic services, I will use the data provided by Radiological Society of North America (RSNA®) and develop deep learning model to detect the presence of pneumonia in the chest X-rays.

Data:

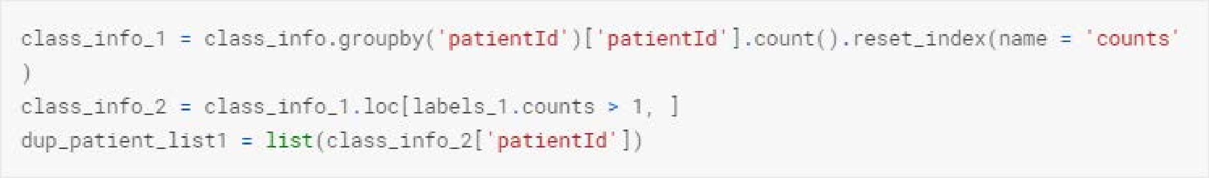
The data used for this project contained three datasets.

1) stage_2_train_labels.csv
2) stage_2_train_images
3) stage_2_detailed_class_info.csv

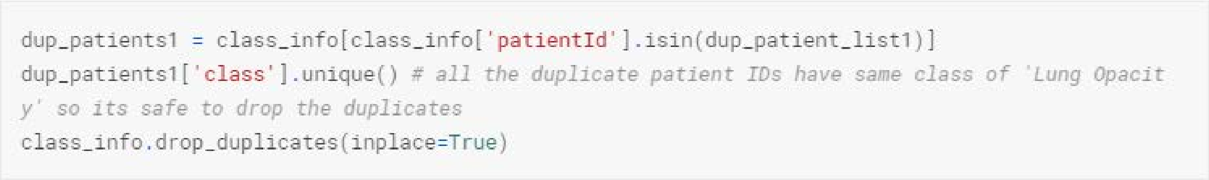
- stage_2_train_labels contains information on patient Id, location of the lung opacity (x,y,height and width) and the target variable (1 or 0). The target variable shows the presence or absence of pneumonia and this is the dependent variable we are trying to predict. In this problem we do not need the position of the lung opacity, so the variables x,y,height and width are removed from the dataset.
- stage_2_detailed_class_info contains information on patient Id and class (No Lung Opacity / Not Normal, Normal, Lung Opacity). If the patient has pneumonia (target = 1) then the class will be lung opacity and if the patient has no pneumonia (target = 0) then the call would be either No Lung Opacity / Not Normal or Normal.
- stage_2_train_images dataset is a DICOM (Digital Imaging and Communications in Medicine) dataset containing the chest x-ray and the metadata regarding the patient.

Data Cleaning:

In both the labels and class_info datasets, there were 30227 rows, however upon exploration, it was found that there are only 26684 unique patient Ids and rest are duplicates. All the duplicates were traced and dropped.

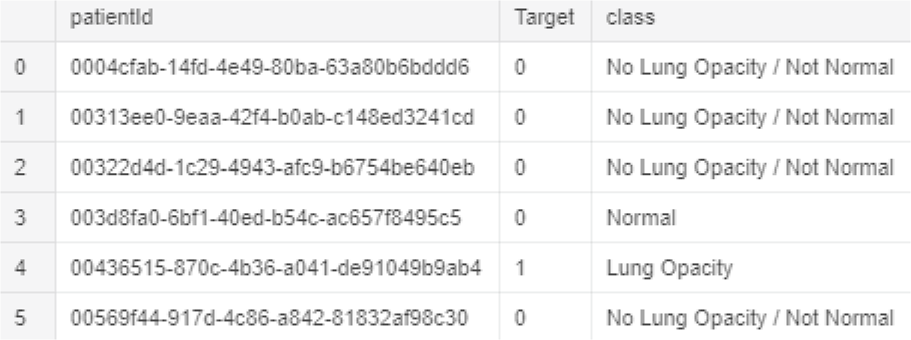


```
class_info_1 = class_info.groupby('patientId')['patientId'].count().reset_index(name = 'counts'
)
class_info_2 = class_info_1.loc[labels_1.counts > 1, ]
dup_patient_list1 = list(class_info_2['patientId'])
```



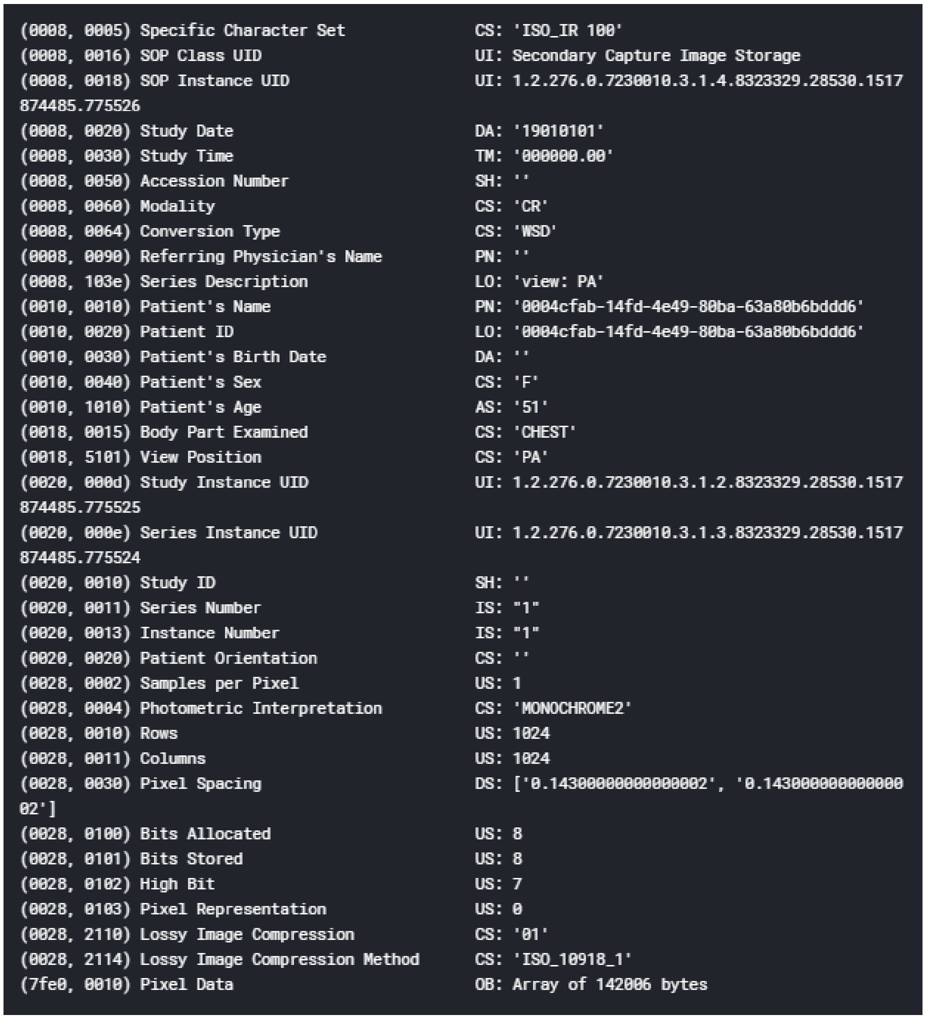
```
dup_patients1 = class_info[class_info['patientId'].isin(dup_patient_list1)]
dup_patients1['class'].unique() # all the duplicate patient IDs have same class of 'Lung Opacit
y' so its safe to drop the duplicates
class_info.drop_duplicates(inplace=True)
```

The labels and class_info datasets were merged into a single dataset.



	patientId	Target	class
0	0004cfab-14fd-4e49-80ba-63a80b6bddd6	0	No Lung Opacity / Not Normal
1	00313ee0-9eaa-42f4-b0ab-c148ed3241cd	0	No Lung Opacity / Not Normal
2	00322d4d-1c29-4943-afc9-b6754be640eb	0	No Lung Opacity / Not Normal
3	003d8fa0-6bf1-40ed-b54c-ac657f8495c5	0	Normal
4	00436515-870c-4b36-a041-de91049b9ab4	1	Lung Opacity
5	00569f44-917d-4c86-a842-81832af98c30	0	No Lung Opacity / Not Normal

The metadata from stage_2_train_images dataset had many variables, but most of the information was redundant and was not useful do only variables 'Modality', 'PatientAge', 'PatientSex', 'BodyPartExamined', 'ViewPosition', 'ConversionType', 'Rows', 'Columns', and 'PixelSpacing' were extracted using user defined function and merged with other data.

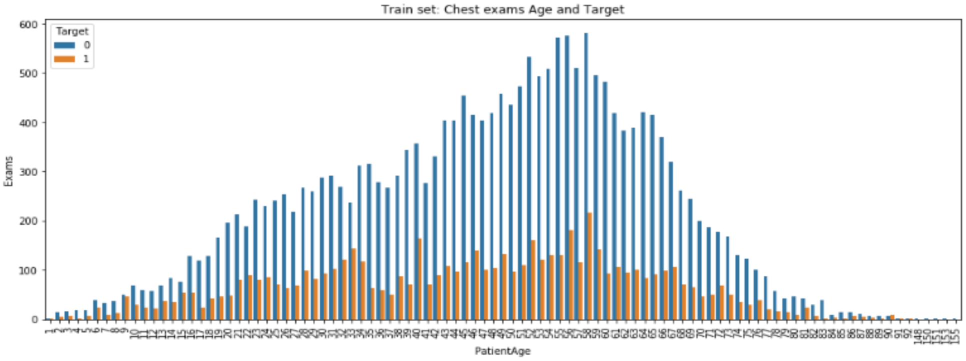


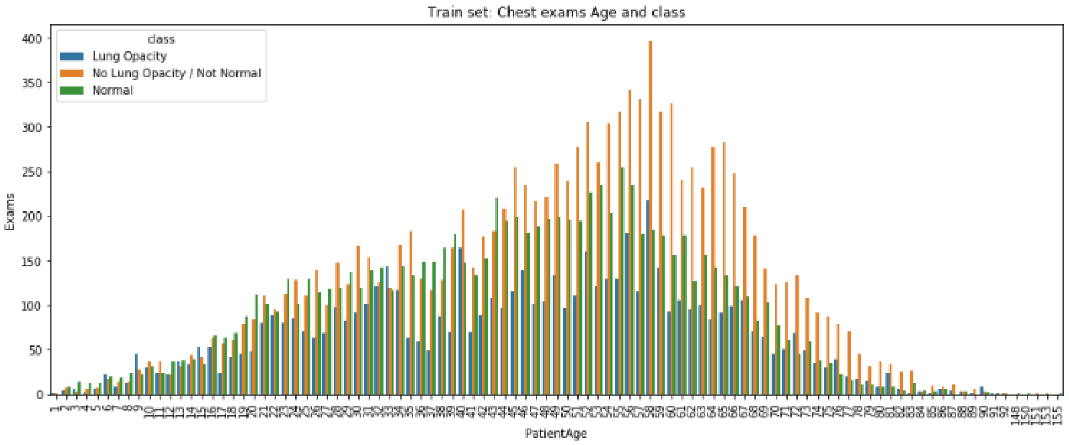
```
(0008, 0005) Specific Character Set              CS: 'ISO_IR 100'
(0008, 0016) SOP Class UID                       UI: Secondary Capture Image Storage
(0008, 0018) SOP Instance UID                    UI: 1.2.276.0.7230010.3.1.4.8323329.28530.1517
874485.775526
(0008, 0020) Study Date                          DA: '19010101'
(0008, 0030) Study Time                          TM: '000000.00'
(0008, 0050) Accession Number                    SH: ''
(0008, 0060) Modality                            CS: 'CR'
(0008, 0064) Conversion Type                     CS: 'WSD'
(0008, 0090) Referring Physician's Name          PN: ''
(0008, 103e) Series Description                  LO: 'view: PA'
(0010, 0010) Patient's Name                      PN: '0004cfab-14fd-4e49-80ba-63a80b6bddd6'
(0010, 0020) Patient ID                          LO: '0004cfab-14fd-4e49-80ba-63a80b6bddd6'
(0010, 0030) Patient's Birth Date                DA: ''
(0010, 0040) Patient's Sex                       CS: 'F'
(0010, 1010) Patient's Age                       AS: '51'
(0018, 0015) Body Part Examined                  CS: 'CHEST'
(0018, 5101) View Position                       CS: 'PA'
(0020, 000d) Study Instance UID                  UI: 1.2.276.0.7230010.3.1.2.8323329.28530.1517
874485.775525
(0020, 000e) Series Instance UID                 UI: 1.2.276.0.7230010.3.1.3.8323329.28530.1517
874485.775524
(0020, 0010) Study ID                            SH: ''
(0020, 0011) Series Number                       IS: "1"
(0020, 0013) Instance Number                     IS: "1"
(0020, 0020) Patient Orientation                 CS: ''
(0028, 0002) Samples per Pixel                   US: 1
(0028, 0004) Photometric Interpretation          CS: 'MONOCHROME2'
(0028, 0010) Rows                                US: 1024
(0028, 0011) Columns                             US: 1024
(0028, 0030) Pixel Spacing                       DS: ['0.14300000000000002', '0.143000000000000
02']
(0028, 0100) Bits Allocated                      US: 8
(0028, 0101) Bits Stored                         US: 8
(0028, 0102) High Bit                            US: 7
(0028, 0103) Pixel Representation                US: 0
(0028, 2110) Lossy Image Compression             CS: '01'
(0028, 2114) Lossy Image Compression Method      CS: 'ISO_10918_1'
(7fe0, 0010) Pixel Data                          OB: Array of 142006 bytes
```

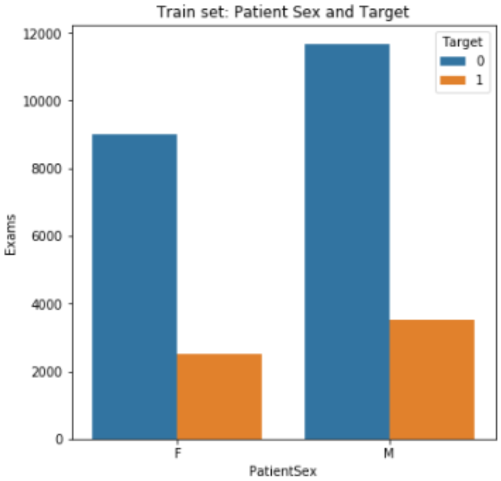
Exploratory Data Analysis:

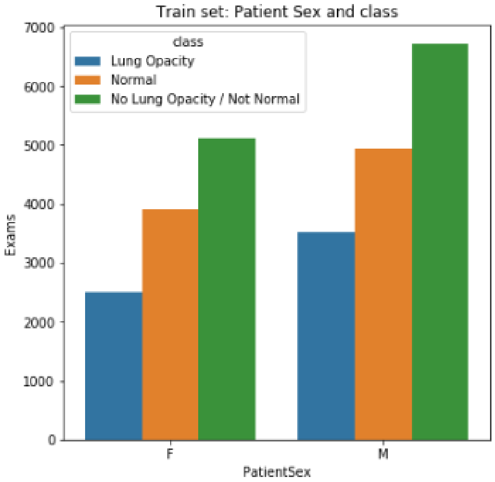
Initial exploratory data analysis showed the following findings.

- There were 20672 patients without pneumonia and 6012 patients with pneumonia.
- Out of 20672 patients without pneumonia, 11821 patients belonged to class No Lung Opacity / Not Normal and 8851 patients belonged to the class Normal.









References:

1) Rui P, Kang K. National Ambulatory Medical Care Survey: 2015 Emergency Department Summary Tables. Table 27. Available from: www.cdc.gov/nchs/data/nhamcs/web_tables/2015_ed_web_tables.pdf
2) Deaths: Final Data for 2015. Supplemental Tables. Tables I-21, I-22. Available from: www.cdc.gov/nchs/data/nvsr/nvsr66/nvsr66_06_tables.pdf
3) Franquet T. Imaging of community-acquired pneumonia. J Thorac Imaging 2018 (epub ahead of print). PMID 30036297
4) Kelly B. The Chest Radiograph. Ulster Med J 2012;81(3):143-148