Pneumonia Detection Challenge

Team

Subhankar Paul

Kunal Patil

Srinivas Dama

Manesh Nambiar

CAPSTONE Project

ABSTRACT

Machine Learning techniques are a very much active area of research in medical science.With increasing size and complexity of medical data like X-rays deep learning gained huge success in prediction of fatal diseases like pneumonia.

This study proposes convolutional neural network model along with other models to detect pneumonia condition from chest X-ray images samples. We started with KNN and CNN base models to detect pneumonia and thereafter applying different models to evaluate performance across different models.

INTRODUCTION

Pneumonia is a disease that affects many people across the globe. It is an inflammatory condition of the lung affecting primarily the small air sacs known as alveoli. Typically, symptoms include some combination of productive or dry cough, chest pain, fever and difficulty breathing.

History

Pneumonia was regarded by Canadian pathologist William Osler in the 19th century as "the captain of the men of death". With the introduction of antibiotics and vaccines in the 20th century, survival greatly improved. Nevertheless, in developing countries, and also among the very old, the very young and the chronically ill , pneumonia remains a leading cause of death . Pneumonia often shortens suffering among those already close to death and has thus been called "the old man's friend"

Pneumonia is a common illness affecting approximately 450 million people a year and occurring in all parts of the world. It is a major cause of death among all age groups resulting in 4 million deaths (7% of the world's total death) yearly. Rates are greatest in children less than five, and adults older than 75 years. It occurs about five times more frequently in the developing world than in the developed world. Viral pneumonia accounts for about 200 million cases. In the United States, as of 2009, pneumonia is the 8th leading cause of death

In 2008, pneumonia occurred in approximately 156 million children (151 million in the developing world and 5 million in the developed world). In 2010, it resulted in 1.3 million deaths, or 18% of all deaths in those under five years, of which 95% occurred in the developing world.

Countries with the greatest burden of disease include India (43 million), China (21 million) and Pakistan (10 million). It is the leading cause of death among children in low income countries. Many of these deaths occur in the newborn period. The World Health Organization estimates that one in three newborn infant deaths is due to pneumonia. Approximately half of these deaths can be prevented, as they are caused by the bacteria for which an effective vaccine is available. In 2011, pneumonia was the most common reason for admission to the hospital after an emergency department visit in the U.S. for infants and In the United States, as of 2009, pneumonia is the 8th leading cause of death.

For this thesis, it was very important to examine and understand how pneumonia develops. It was also necessary to be able to distinguish between pneumonia and other similar ailments. This is discussed in more detail further on. The Radiological Society of North America, further regarded as RSNA, provided a large dataset on the Kaggle platform. The Dataset itself contained approximately 30,000 files in the DICOM format. Each file contained the data of the patient and his/her chest x-ray scan. There is a dedicated section for the DICOM format, later in the report. Our task initially was to do exploratory data analysis. That meant, excluding unnecessary data and extracting the crucial data that our system needed. (Radiological Society of North America, 2018)

Preliminary Data Research :

pneumonia

There is no doubt that pneumonia is a very serious and life-threatening disease, but to fully create a solution we need to understand a problem at a deeper level. First, let us understand how pneumonia develops and how it affects the lungs.

When humans breathe, air reaches the lungs, as seen in Figure 1, by flowing down the trachea, then it continues through the bronchi and the bronchioles and finishes in the alveoli. The alveoli are tiny little air sucks that are wrapped up in capillaries. This is where the most gas exchange occurs in the lungs. Oxygen leaves the air in the alveoli and enters the bloodstream, while carbon dioxide leaves the

bloodstream and is exhaled out of the lungs

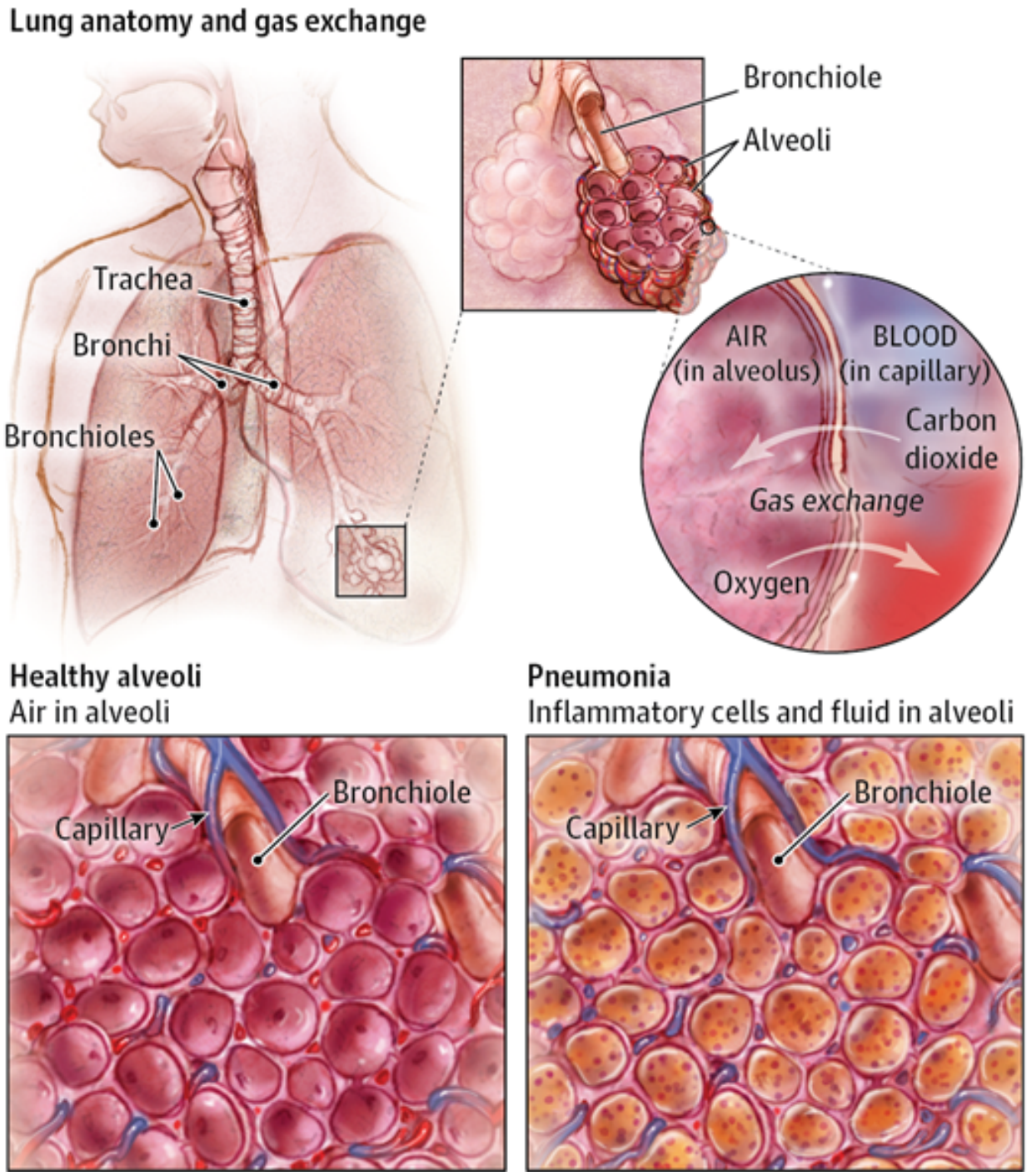
In addition to inhaling air, sometimes other stuff is inhaled, like microbes. However, our immune system is very good at protecting our health in these kinds of situations. For instance, our organisms have mechanical techniques like coughing or special microorganisms like macrophages that are in the alveoli and ready to protect it from anything.

Interesting things happen when some microbes succeed in colonizing the bronchioles or alveoli. When this happens, you have got pneumonia.

Pneumonia is a disease that is caused by bacterial or viral infection of the lungs. This causes the air sacs to fill up with fluid and substantially affects breathing.

Lung Opacity

This is an illustration of the chest anatomy with the lungs highlighted. You can see that there is a mass of tissue surrounding the lungs and between the lungs. These areas contain skin, muscles, fat, bones, and also the heart and big blood vessels.



The infection and the body's immune response, the sacks in the lungs (termed alveoli) are filled with fluids instead of air. The reason that pneumonia associated lung opacities look diffuse on the chest radiograph is because the infection and fluid that accumulate spread within the normal tree of airways in the lung. There is no clear border where the infection stops. That is different from other diseases like tumors, which are totally different from the normal lung, and do not maintain the normal structure of the airways inside the lung.

Opacity is a pretty loose term - "Opacity refers to any area that preferentially attenuates the x-ray beam and therefore appears more opaque than the surrounding area. It is a nonspecific term that does not indicate the size or pathologic nature of the abnormality"

DICOM

One of the ways to diagnose pneumonia is to analyze the Chest X-Ray (further on CXR). The dataset provided in the project has about 30 000 CXR images in dicom format. Now, we need to understand what it is and how to utilize it.

DICOM – Digital Imaging and Communications in Medicine is known as an international standard for medical images and everything that is related to them. DICOM images are known to have high quality, since the diagnosis requires as clear information as possible. Nowadays, this format is implemented in almost all medical domains like radiology, cardiology, radiotherapy devices, ophthalmology and even dentistry.

Since 1993, when DICOM was introduced, it has revolutionized radiology, allowing practitioners to switch from X-Ray films to digital format. (National Electrical Manufacturers Association , 1993) In Figure 2, you can see the specific data inside a DICOM file provided to us by the radiological community in the USA. It contains several important bits of information, such as the unique patient ID, the view position of the body when the scan was taken, the gender and age of the patient. All this information can be used to further explore the possible solutions to the problem.

```
(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: 1.2.276.0.7230010.3.1.4.8323329.28530.151
7874485.775526
(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: 1.2.276.0.7230010.3.1.2.8323329.28530.151
7874485.775525
(0020, 000e) Series Instance UID                 UI: 1.2.276.0.7230010.3.1.3.8323329.28530.151
7874485.775524
(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: ['0.1430000000000002', '0.1430000000000
```

Figure 3 DICOM content (Radiological Society of North America, 2018)

Treatment:

Vaccines are available to prevent pneumonia caused by pneumococcal bacteria or the flu virus, or influenza.

AI for Pneumonia

AI can be applied to various types of healthcare data (structured and unstructured). Popular AI

techniques include machine learning methods for structured data, such as the classical support vector machine and neural network, and the modern deep learning, as well as natural language processing for unstructured data.

China:

Infervision to gather medical data to train machine-learning algorithms in tasks like reading scans more easily than US or European rivals.

Infervision created its main product, software that flags possible lung problems on CT scans, using hundreds of thousands of lung images collected from major Chinese hospitals. The software is in use at hospitals in China, and being evaluated by clinics in Europe, and the US, primarily to detect potentially cancerous lung nodules.

USA:

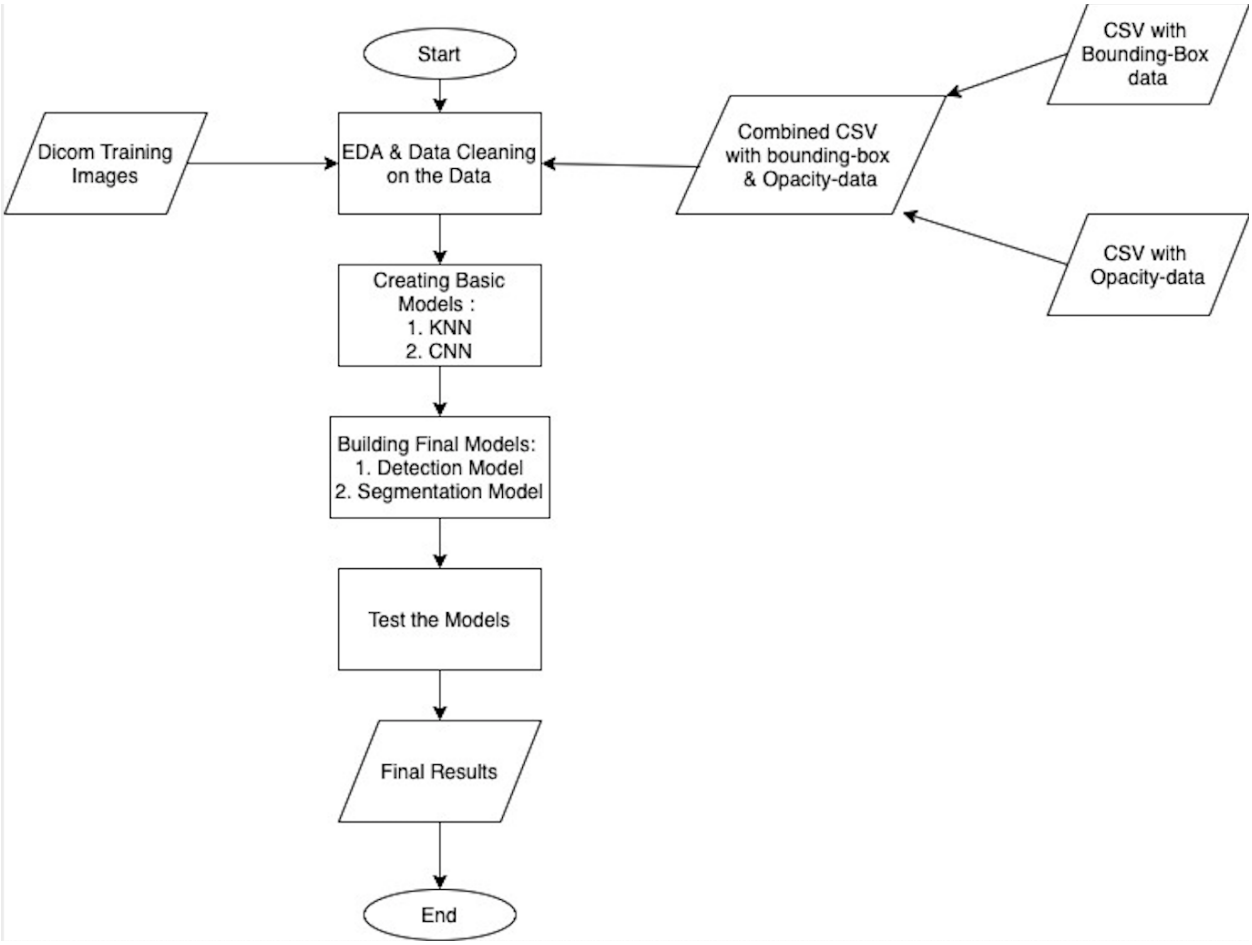
Flagler Hospital in Saint Augustine, Florida, is using artificial intelligence tools to improve the treatment of pneumonia.

The AI tools automatically revealed new, improved care pathways for pneumonia after analyzing thousands of patient records from the hospital and identifying the commonalities for those with the best outcomes

Ayasdi uses a branch of mathematics called topological data analysis to group patients treated similarly and find relationships between those groups. This analysis may result from AI in the form of supervised learning or unsupervised learning.

“Once the data loaded, they use unsupervised learning AI algorithm to generate treatment groups,” “In the case of pneumonia patient data, Ayasdi produced nine treatments groups. Each group was treated similarly and statistics were given to them to understand that group and how it differed from the other groups.”

Flowchart

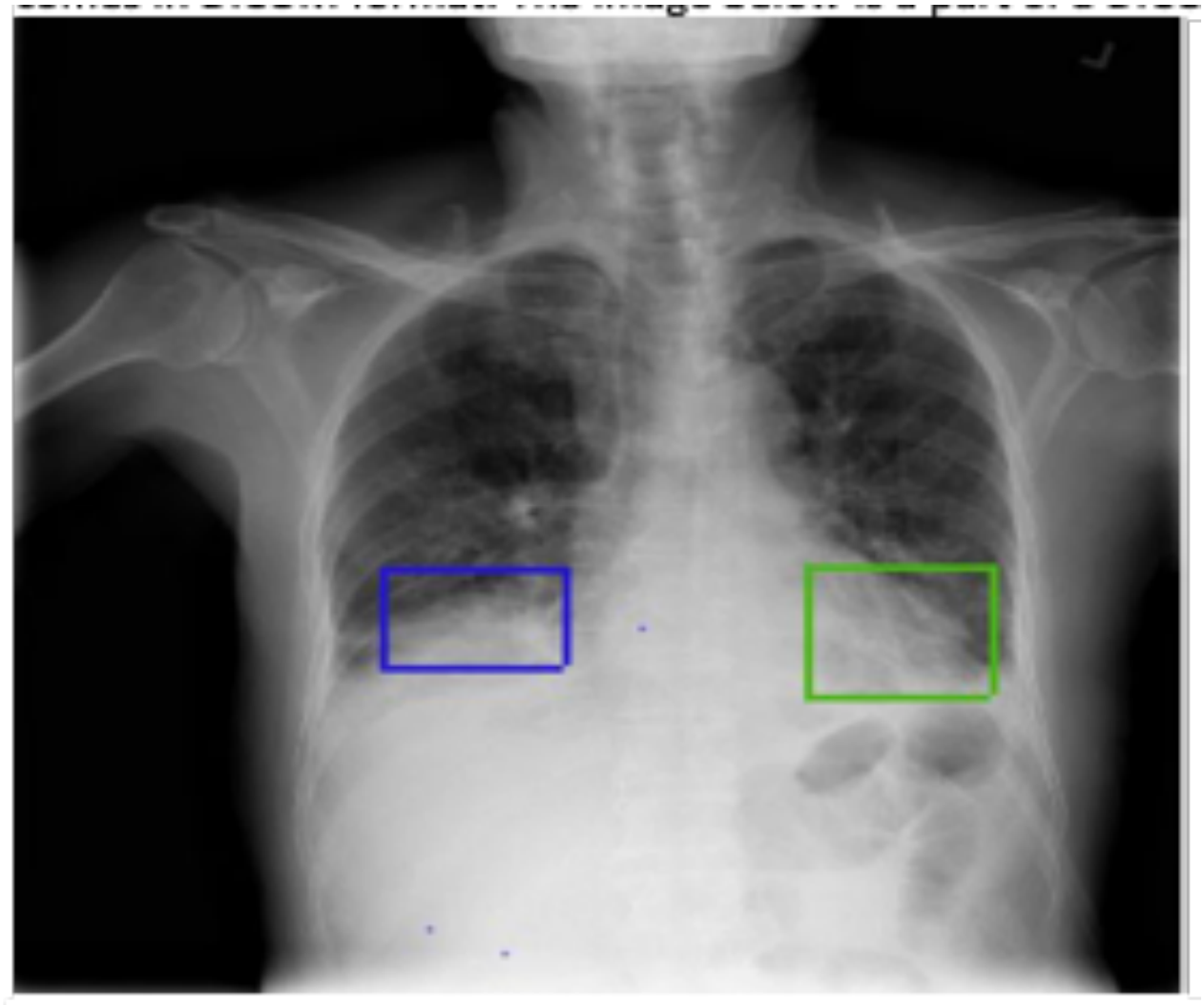


Summary of problem statement, data and findings

EDA and Preprocessing.

The most important part of data science and machine learning is data and data must be neat and clean for models to process it. Data pre-processing methodologies vary with regards to the data types. It can be tabular data, sequential data(text, music etc) or images.

As it has been said earlier, the data comes in DICOM format. The image below is a part of a DICOM file.



pneumonia DICOM

We can see a sample picture of a patient with pneumonia. Knowing how to read CXR helps in understanding the data better. Let us break down the process of X-Ray imaging, X-Ray passes through the body and reaches the detector on the other side. During its journey it encounters matter with different densities, dense materials like bones and tissues absorb the radiation and appear white on CXR, other materials like air do not really absorb X-Ray, thus it reaches the target.

Non-pneumonia DICOM

Briefly speaking:

Air is indicated by the black colour, bones are the white colour, tissues and fluids are grey. Now, we can analyse the pictures above, as we can see the sample patient 1 has pneumonia which is labelled by the bounding boxes, and sample patient 2 has clear and healthy lungs, as illustrated in *Non-pneumonia DICOM*.

In addition to pneumonia, similar opaque regions can be produced by other dense objects, like lung cancer or fluid like water in the lungs. Considering this, there are have 3 classes in the dataset: Lung Opacity, Not Normal and Normal. The target variable can have 2 values, 0 and 1. So it is a binary classification task. However, the positions of

Summary of the Approach to EDA and Pre-processing

Data Observations :

```
In [5]: 1 print(df.iloc[0])
```

```
patientId    0004cfab-14fd-4e49-80ba-63a80b6bddd6
x                                            NaN
y                                            NaN
width                                        NaN
height                                       NaN
Target                                         0
Name: 0, dtype: object
```

```
In [6]: 1 print(df.iloc[5])
```

```
patientId    00436515-870c-4b36-a041-de91049b9ab4
x                                            562
y                                            152
width                                        256
height                                       453
Target                                         1
Name: 5, dtype: object
```

Observations : For entries where the target column contains 0, then the columns x,y,width and height have blank entries. Whereas for the one's with target 1, these columns contain the definition of the bounding box.

Reading the DICOM File :

```
In [7]: 1 patientId = df['patientId'][5]
```

```
In [8]: 1 dcm_file5 = 'stage_2_train_images/%s.dcm' % patientId
```

```
In [9]: 1 print(patientId)
```

```
00436515-870c-4b36-a041-de91049b9ab4
```

```
In [10]: 1 dcm_data5 = pydicom.read_file(dcm_file5)
```

```
In [11]: 1 print(dcm_data5)
```

We can observe that the name of the dicom File is in the format : *<patient_id>.dcm*

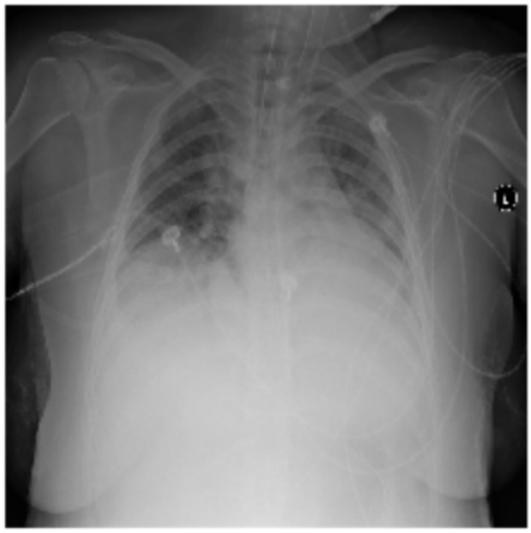
Viewing the actual Image within the DICOM File :

```
1  im = dcm_data5.pixel_array
2  print(im.shape)
```

```
(1024, 1024)
```

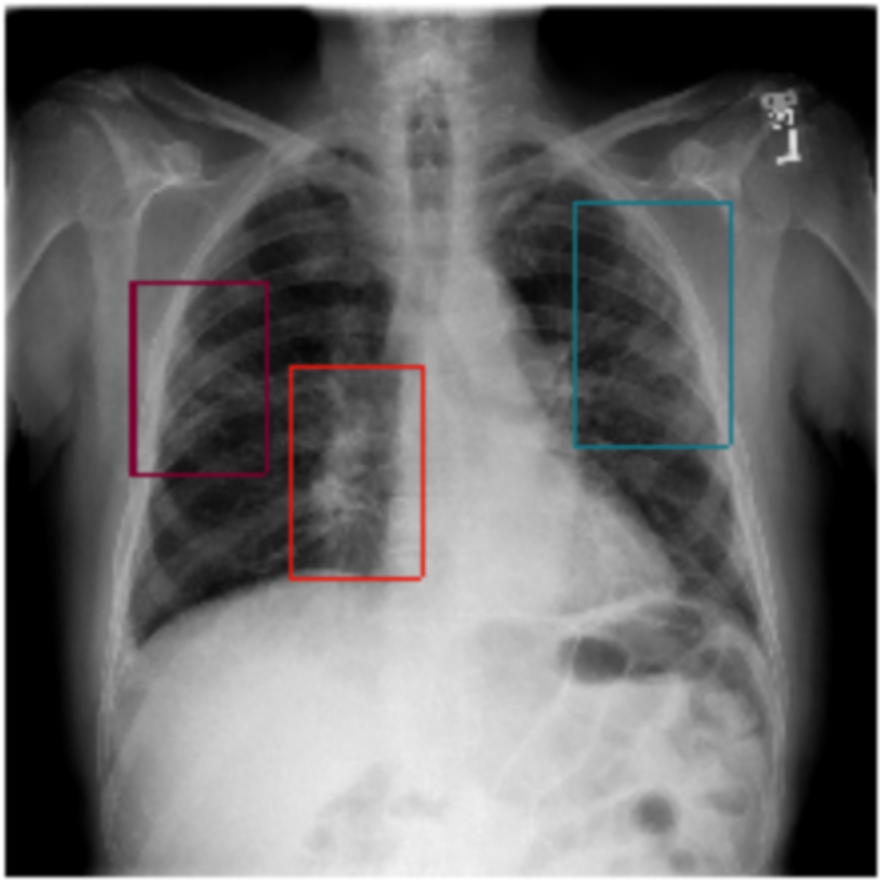
In [16]:
```
1  pylab.imshow(im, cmap=pylab.cm.gist_gray)
2  pylab.axis('off')
```

Out[16]: (-0.5, 1023.5, 1023.5, -0.5)



Bounding boxes must be predicted as well :

We extract this information from the .csv file for all the images which have the target column as 1(ie. Pneumonia is present).

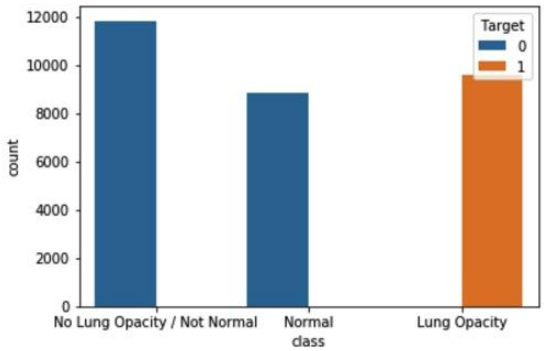


```
In [12]: print(dcm_data)
```

```
(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: 1.2.276.0.7230010.3.1.4.8323329.28530.15
17874485.775526
(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: 1.2.276.0.7230010.3.1.2.8323329.28530.15
17874485.775525
(0020, 000e) Series Instance UID                 UI: 1.2.276.0.7230010.3.1.3.8323329.28530.15
17874485.775524
(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, 0.1430000000000000
2]
(0028, 0100) Bits Allocated                      US: 8
```

Classification of the data With respect to Lung Opacity

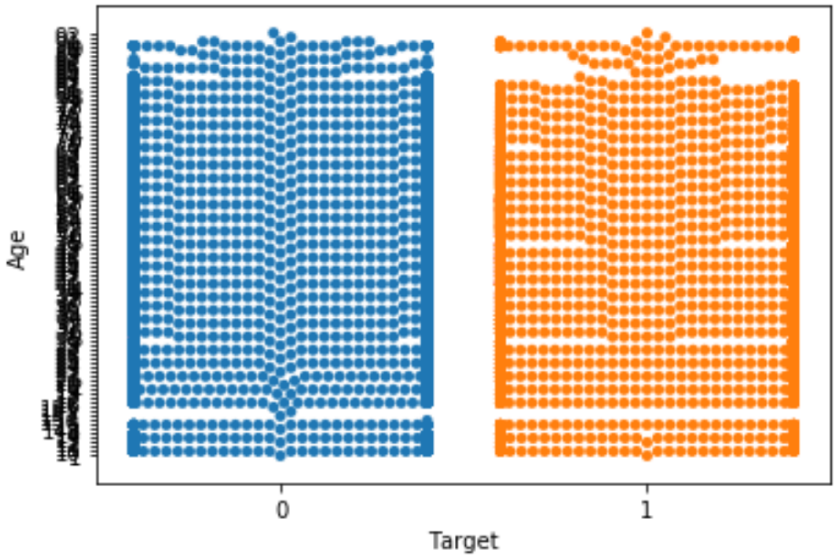
```
In [40]: sns.countplot(x = 'class', hue = 'Target', data = df_marged);
```



Bivariate Data Analysis

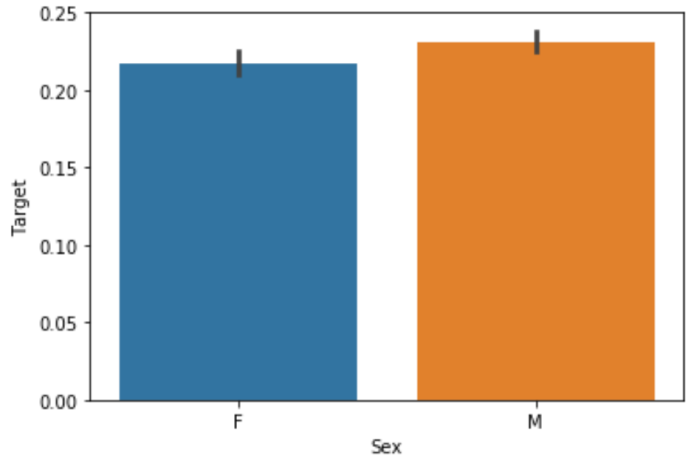
```
In [88]:  1  sns.swarmplot(df_biv['Target'], df_biv['Age'])
```

```
Out[88]: <matplotlib.axes._subplots.AxesSubplot at 0x1a9ce180d0>
```




```
In [83]:  1  sns.barplot(df_biv['Sex'], df_biv['Target'])
```

```
Out[83]: <matplotlib.axes._subplots.AxesSubplot at 0x1aac94cf90>
```



Implementation Section

Code Description :

We have uploaded the Following Files :

1. EDA_and_KNN.ipynb :

   The notebook which contains all the EDA and the KNN model.

2. CNN.ipnb :

   The notebook which contains all the CNN model(Work in Progress)

3. Interim_report.pdf :

   The Interim Project Report.

Deciding on the Models and Model Building :

Approach :

To help understand the best model that would meet our requirements we have created Two basic Models :

1. A KNN based basic model
2. A CNN based basic model

KNN based basic model

Extracting the :

X : Features

y : Target (binary) to build the model.

In [52]:
```
for n, image in enumerate(images_path):
    ds = pydicom.dcmread(os.path.join(folder_path, image))
    pixel_array_numpy = ds.pixel_array
    y[n] = df_merged.loc[df_merged['patientId'] == ds.PatientID].iloc[0].Target
    img = ds.pixel_array
    img = cv2.resize(img, (100,100))
    img = img.flatten()
    x[n] = img
    print(n)
```

Building a KNN model using only images(100x100 array) as inputs

```
In [66]: 1 knn = KNeighborsClassifier(n_neighbors = 4)

In [67]: 1 knn.fit(list(x_train),list(y_train))

Out[67]: KNeighborsClassifier(algorithm='auto', leaf_size=30, metric='minkowski',
                              metric_params=None, n_jobs=None, n_neighbors=4, p=2,
                              weights='uniform')

In [69]: 1 score = knn.score(list(x_test),list(y_test))

In [70]: 1 print(score)

         0.7812890332250811
```

- The available training data was further split into a Train: Test ratio of 70:30
- The KNN based model was build by resizing the image to 100x100 array of pixels.
- The KNN model is used to build a Classification Model which can be used to predict whether a patient has Pneumonia or not based on his Xray Image.
- Using the KNN model we could achieve a prediction score of 78% for predicting the presence of Pneumonia in the test sample.

CNN based basic model

Loading the Data to Drive and Colab:

- Get the Kaggle data to Drive and Colab

- Upload image data to Drive and mount the data
- Import necessary libraries:

```
[46] #IMport Libraries

     import os
     import sys
     import random
     import math
     import numpy as np
     import cv2
     import matplotlib.pyplot as plt
     import json
     from imgaug import augmenters as iaa
     from tqdm import tqdm
     import pandas as pd
     import glob
     from sklearn.model_selection import KFold
```

```
[47] import pydicom
```

```
[ ]  !pip install pydicom

     Requirement already satisfied: pydicom in /usr/local/lib/python3.6/dist-packages (1.4.2)
```

Model to be used for Transfer learning:

-Matterport's Mask RCNN. In case there is tensorflow version dependency issues we will move to https://github.com/tomgross/Mask_RCNN.git

Some setup functions and classes for Mask-RCNN

- dicom_fps is a list of the dicom image path and filenames
- image_annotions is a dictionary of the annotations keyed by the filenames
- parsing the dataset returns a list of the image filenames and the annotations dictionary

```
[55] def get_dicom_fps(dicom_dir):
         dicom_fps = glob.glob(dicom_dir+'/'+'*.dcm')
         return list(set(dicom_fps))

     def parse_dataset(dicom_dir, anns):
         image_fps = get_dicom_fps(dicom_dir)
         image_annotations = {fp: [] for fp in image_fps}
         for index, row in anns.iterrows():
             fp = os.path.join(dicom_dir, row['patientId']+'.dcm')
             image_annotations[fp].append(row)
         return image_fps, image_annotations
```

This is how a sample metadata looks:

```
(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: 1.2.276.0.7230010.3.1.4.8323329.28530.1517874485.775526
(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: 1.2.276.0.7230010.3.1.2.8323329.28530.1517874485.775525
(0020, 000e) Series Instance UID                UI: 1.2.276.0.7230010.3.1.3.8323329.28530.1517874485.775524
(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, 0.14300000000000002]
(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 elements
```

Data Split:

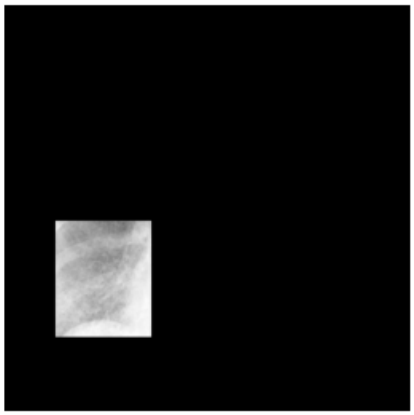
Split the data into training and validation datasets

Note: We have only used only a portion of the images for demonstration purposes. See comments below.

- To use all the images do: image_fps_list = list(image_fps)
- Or change the number of images from 100 to a custom number

Display a random image with bounding boxes:

```
(1024, 1024, 3)
/content/drive/My Drive/stage_2_train_images/IMG/IMG/36f775ee-d044-43a3-bbe5-000ed7b498b8.dcm
[1]
```



Model Training:

While training the model facing compatibility issue:

```
AttributeError: module 'tensorflow' has no attribute 'random_shuffle'
```

SEARCH STACK OVERFLOW

Will either downgrade tensorflow or use updated MASK RCNN

Result

1. Using the KNN as a basic model, we were able to create a model for Pneumonia detection, which gave us about 78% detection Accuracy.
2. Using the CNN based model we could detect the areas within the xray displaying lung opacity.

CONCLUSION

Further Improving Model Performance

Based on the observations that we could make with the basic models that we have build so far, we understand that we would need 2 types of models :

- **A Detection Model**

  The Detection model will primarily be used to detect whether the patient in question is likely to have pneumonia or not.

  Some of the models that could be considered for this include :

  - RCNN
  - Keras-RetinaNet

- **A Segmentation Model**

  The Segmentation model would be used to identify the boulding boxes within the images where lung Opacity is likely to be present.

  Some of the models that could be considered for this include :

  - YOLO
  - DenseNet
  - Faster- RCNN

REFERENCES

1. America, R. S. (2018, August 27). Retrieved from RSNA Pneumonia Detection Challenge: https://www.kaggle.com/c/rsna-pneumonia-detection-challenge

2. Colaboratory. (2017). Retrieved from https://colab.research.google.com

3. Karpathy, A. (2018). CS231n: Convolutional Neural Networks for Visual Recognition. Retrieved from http://cs231n.stanford.edu/