## Summary of the problem statement

Pneumonia is in the list of top 10 causes for death in the US. It accounts for 15% of all deaths in children under the age of 5 globally. Accurately diagnosing Pneumonia is an elaborate process. It requires review of Chest Radiograph by trained specialists and other detailed examination.

The specialists are burdened with reviewing a high volume of Chest X-rays. Using AI to screen the radiographs for opacities indicating possibility of pneumonia, would be a solution to expedite the review process for these specialists.

**Pneumonia Detection Model**

The goal of this project is to build a computer vision model, which can be used in the diagnosis of Pneumonia. The output of the model should isolate inflammation in the Chest X-Ray images with bounding boxes and classify the images with detectable opacities as possibly having Pneumonia, for further investigation by domain expert.

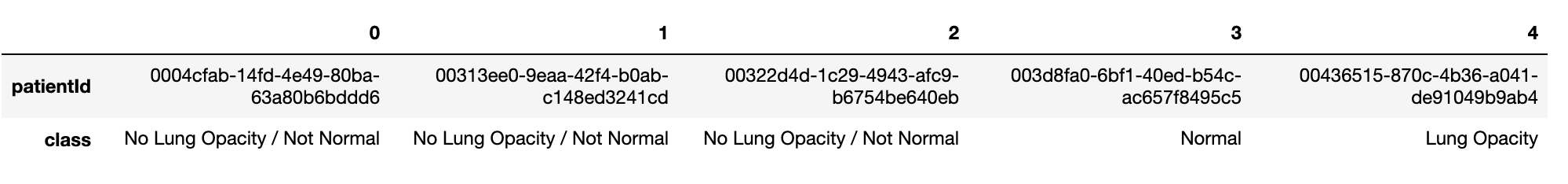
## The Dataset

The dataset contains images with details in DICOM® format. DICOM® (Digital Imaging and Communications in Medicine) is the international standard to transmit, store, retrieve, print, process, and display medical imaging information. The data set that has been shared has 26,684 training and 3,000 test X-ray images. The images are annotated with bounding boxes to highlight the region in the X-ray that is indicative of possible Pneumonia.

The data downloaded for analysis has the following files:

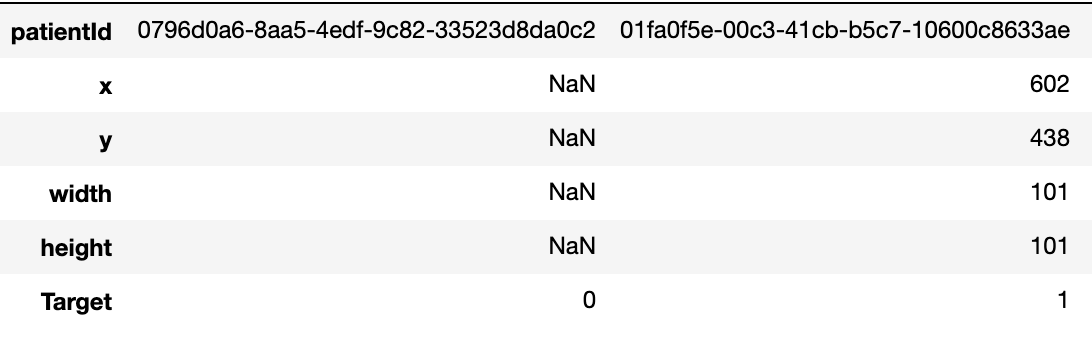
1. **stage\_2\_detailed\_class\_info.csv**

This file has all the patient\_Ids mapped on to the recorded outcome of their X-Ray.



1. **stage\_2\_train\_labels.csv**

This file has data where all the patient\_Ids are mapped to the regions of opacity identified in the X-Ray images. One patient can have more than one opaque region. The opacity is identified with the (x,y) originating coordinates pixel on the image of size (1024,1024) and the width and heights as pixels there of. It also includes Target.



1. **stage\_2\_train\_images - DICOM files**

These files are special image files as mentioned above and along with the actual images also have meta-data about the patient. The files include all the possible outcomes. These are the images we will train our models on to predict whether the X-Ray is indicative of presence of lung opacity, which is indicative of Pneumonia.

1. **stage\_2\_test\_images - DICOM files**

These have the same information as the training images. These have been separated out for convenience of testing.

**Exploratory Data Analysis:**

The following observations where made with the available data set. This analysis has formed the basis for the data pre-processing and the model built.

### The classification categories of the X-Ray images based on opacity

**Based on the opacity of lungs in the X-Ray of the chest, they are classified into 3 categories:**

* Normal**(Healthy lungs)**
* No Lung Opacity/Not Normal**(unhealthy lungs but not indicative of Pneumonia)**
* Lung Opacity (Indicative of pneumonia)

1. **Classification of X-Ray images based on the Pneumonia Diagnosis**

Based on whether the patient was diagnosed with Pneumonia or not, the Target column in the dataset has been set 1 or 0, where 1 indicates that the patient was diagnosed with Pneumonia and 0 indicates that patient was not diagnosed with pneumonia.

1. **Relation between the two classifications**

All the patient records classified as ‘Normal’ and ‘No Lung Opacity/Not Normal’ are classified as Target 0, as they are not diagnosed for Pneumonia. All the patient records classified as ‘Lung Opacity’ are classified as Target ‘1’. This is an important consideration for our data pre-processing. Since the objective of this project is only to develop a model for diagnosis of Pneumonia, we will take mainly into consideration only the Target and not the other classification. In the training data set we have 20672 patients who don’t have pneumonia and 6012 patients who have pneumonia.

1. **The opacity bounding boxes**

The stage\_2\_train\_labels.csv contains the coordinates of the opacity, which can be presented as boxes on the images to indicate the region of detection. Each patient can more than one lung opacity. In the given dataset for 13 patients, there were 16 bounding boxes, for 119 patients there were 9 bounding boxes, for 3266 patients there were 4 bounding boxes and for 2614 patients there was just one bounding box. The bounding box highlights the lung opacity.

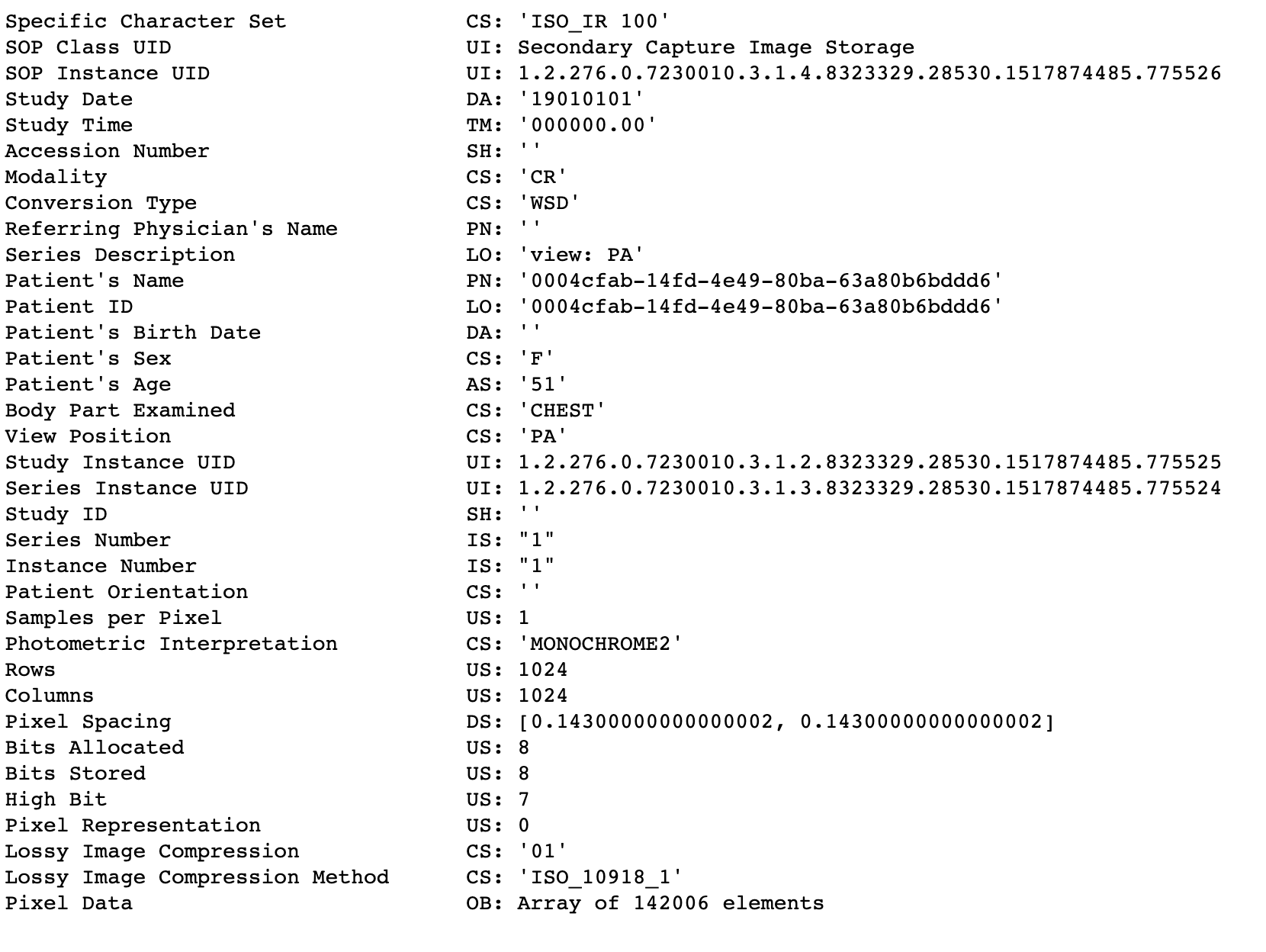
Subject matter expert opinion states that the more the number of detected opacities, the faster the infection is spreading. It also means that the bigger the size of the opacity the more severe the infection is.

1. **The DICOM Images**

The DICOM Images have the X-Ray images in the following format and have additional meta data about the patient whose X-Ray is diagnosed.

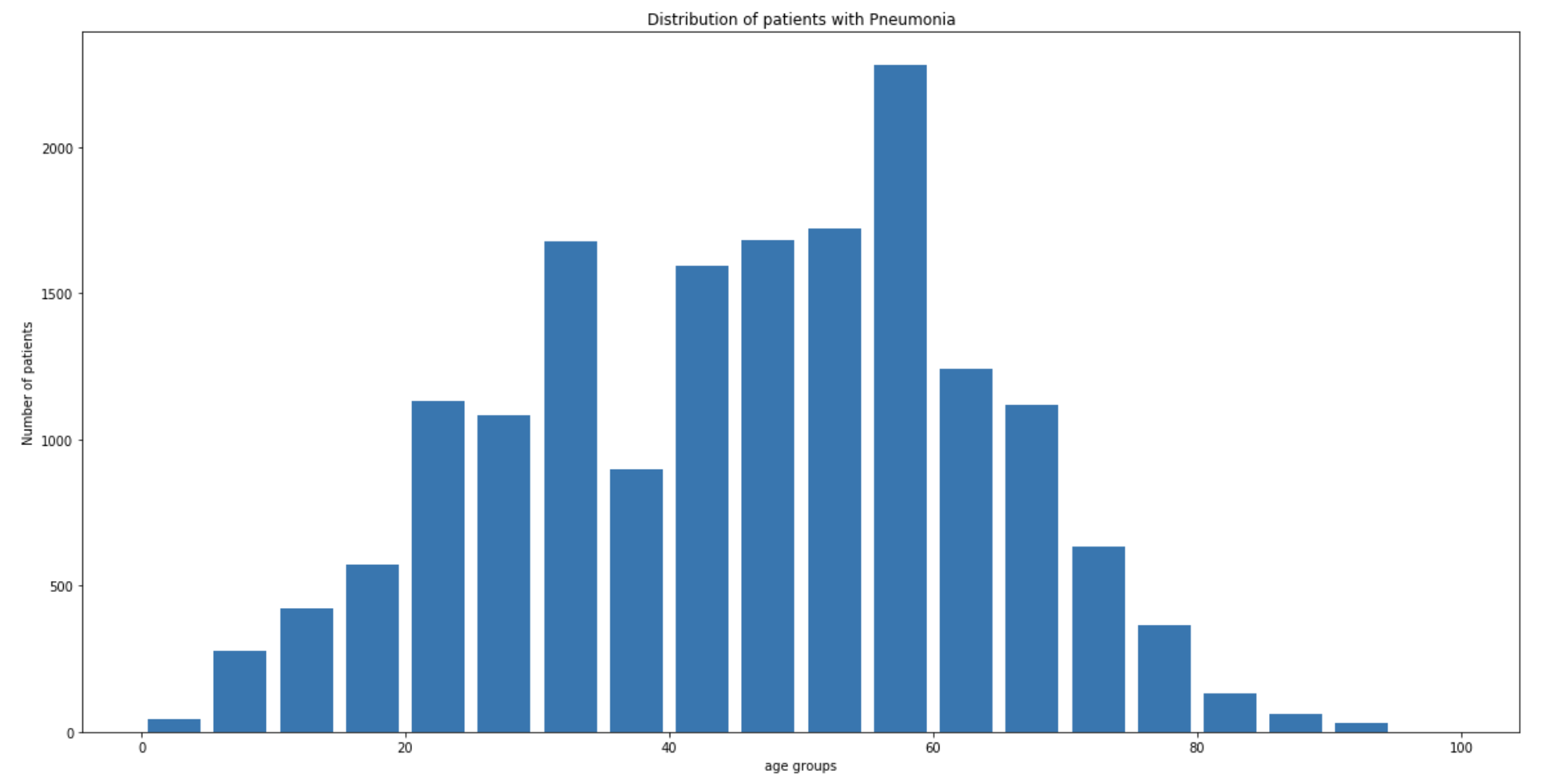
|  |  |
| --- | --- |
| X-Ray Image of patient with pneumonia 1 | X-Ray image showing infected region 1 |

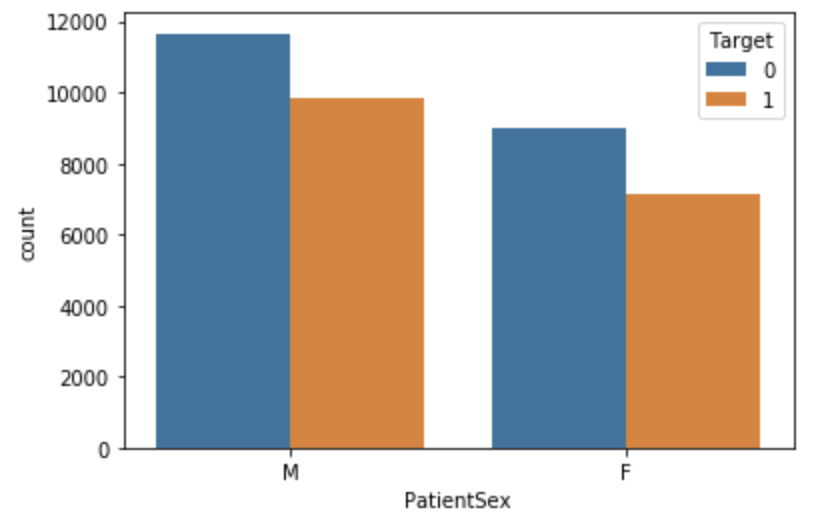
The DICOM images contain the following meta data about the patients and the image.



1. **Influence of metadata on outcome**

Of the metadata available, the most relevant ones on discussion with subject matter expert seems to be the age and gender. Younger children and older people are more prone to pneumonia, However the dataset is not directly indicative of this. This could be because of the demographic of the region where it was collected or it could be indicative of the age group that is more prone to infection due to exposure.

 The data indicates that men are more prone to pneumonia than women. Some scholarly [article](mailto:https://www.thoracic.org/patients/patient-resources/resources/top-pneumonia-facts.pdf#https://www.thoracic.org/patients/patient-resources/resources/top-pneumonia-facts.pdf) is in support of this inference but the data is not sufficient to confirm the same.



**Balancing the data**

We created a merged dataset with all the imperative information, which we used in our analysis. Since we decided to do a binary classification of the data, we decided to balance the data to train the model to have equal number of cases(6000) of Pneumonia and non-pneumonia cases to train the model.

**Model building:**

## The outcome of the model is two fold. First the model should be able to categorize whether the image under investigation is Pneumonic or not and secondly, of the images categorised as Pneumonic, the model should identify the infected areas or the opacities and draw bounding boxes. The first sequential model we developed is a 3-Convultion layers model, which does a binary classification of the image with 85% accuracy and an F1 score of 74. We chose this model of the many models we tried as the accuracy seemed better and F1 score was promising.

**Other Models and Data Balancing considered**

The below table summarizes the few models and data set and balancing we considered before deciding to finalize on the 3 Conv layer shown above for initial binary classification. Our decision was based on the Validation dataset accuracy and sensitivity of the data based on test data. These tests have been submitted on a separate notebook along with the submission.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **S.No** | **Model** | **Data Balanced** | **Data Balance** | **Train Acc** | **Val Acc** | **F1 Score** | **Consideration** |
| 1 | 5 Conv layers | Yes | Dropped 'No Lung Opacity/Not Normal' | 90 | 91 | 92 | Might be over-fitting. |
| 2 | 5 Conv layers | No |  | 80 | 81 | 78 | Not Balanced. Sensitivity might be low |
| 3 | 5 Conv layers | Yes | Randomly selected 6000 Target 0 cases | 86 | 85 | 76 | Worth considering as final model. But too many layers |
| 4 | 3 Conv layers | Yes | Dropped 'Not Normal' | 93 | 91 | 90 | Might be over-fitting. |
| 5 | 3 Conv layers | Yes | Randomly selected 6000 Target 0 cases | 85 | 84 | 75 | Possibly accuracy will increase after more epoch. Good result on Test data as shown in additional notebook attached. Worth considering as final model. |

**Model to predict the opacities:**

We evaluated 3 pre-trained Computer vision models to predict based on the X-Ray image and output the bounding boxes.

1. YOLO – Darknet – We used the YOLO Darknet [model](https://docs.google.com/document/d/1TVtdqzqhUG_G7hOZc1tpTGatT1As-J-_oQXwhhfVeqc/edit#bookmark=id.30j0zll) to predict the bounding boxes in the X-Ray images.  The model training is time- consuming in the Colab basic version and often gets timed out. We used a pre-trained model and ran the validation and tested the prediction.

The output of the YOLO model for the X-Ray we tested had two opacities detected. In the table below you can see the original image, the YOLO predicted output of bounding boxes, the actual bounding boxes and the overlap of the bounding boxes between the actual and predicted.

|  |  |
| --- | --- |
|  |  |
|  |  |

1. ChexNet
2. Mask RCNN