Nageswari Kalpatapu

Nkalpatapu@gmail.com

Pneumonia detection

Capstone project

[1. Introduction 2](#_Toc41126003)

[ What is Pneumonia? 2](#_Toc41126004)

[ Pneumonia Detection 2](#_Toc41126005)

[ Business Domain Value 2](#_Toc41126006)

[2. Project Description 3](#_Toc41126007)

[ Understanding Data 3](#_Toc41126008)

[ Visualization of sample images: 5](#_Toc41126009)

[3. Preprocessing data 7](#_Toc41126010)

[ Duplicates 7](#_Toc41126011)

[ Missing Values 7](#_Toc41126012)

[4. Model Training: 7](#_Toc41126013)

[ Model: Resnet (based of Mask R CNN) 7](#_Toc41126014)

[ Conclusion on this model: 11](#_Toc41126015)

[5. Alternative Model: MobileNet 12](#_Toc41126016)

[ Preprocessing data for this model: 12](#_Toc41126017)

[ Model building: 12](#_Toc41126018)

[ Conclusion on Mobilenet model: 13](#_Toc41126019)

|  |
| --- |
| Introduction |

### What is Pneumonia?

Pneumonia is an infection in one or both lungs. Bacteria, viruses, and fungi cause it. The infection causes inflammation in the air sacs in the lungs, which are called alveoli.

Pneumonia accounts for over 15% of all deaths of children under 5 years old internationally.

In 2017, 920,000 children under the age of 5 died from the disease. 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 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 points in time and correlation with clinical symptoms and history are helpful in making the diagnosis.

### Pneumonia Detection

CXRs are the most performed diagnostic imaging study. Several factors such as positioning of the patient and depth of inspiration can alter the appearance of the CXR, complicating interpretation further. In addition, clinicians are faced with the challenge of reading high volumes of images every shift. To detect Pneumonia, we need to detect inflammation of the lungs. In this project, challenge is to build an algorithm to detect a visual signal for pneumonia in medical images. Specifically, algorithm needs to automatically locate lung opacities on chest radiographs.

### Business Domain Value

Automating Pneumonia screening in chest radiographs, providing affected area details through a bounding box.  Project objective is to assist physicians to make better clinical decisions or even replace human judgement in certain functional areas of healthcare.

Guided by relevant clinical questions, powerful AI techniques can unlock clinically relevant information hidden in the massive amount of data, which in turn can assist clinical decision making.

|  |
| --- |
| Project Description |

In this capstone project, the goal is to build a pneumonia detection system, to locate the position of inflammation in an image.

Tissues with sparse material, such as lungs which are full of air, do not absorb the X-rays and appear black in the image. Dense tissues such as bones absorb X-rays and appear white in the image.

While we are theoretically detecting “lung opacities”, there are lung opacities that are not pneumonia related.

In the data, some of these 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.

### Understanding Data

#### Source of Data

This data is collected by RSNA (Radiological Society of North America). They worked with colleagues at the Society for Thoracic Radiology and MD.ai to label pneumonia cases found in the database of chest x-rays made public by the National Institutes of Health (NIH).

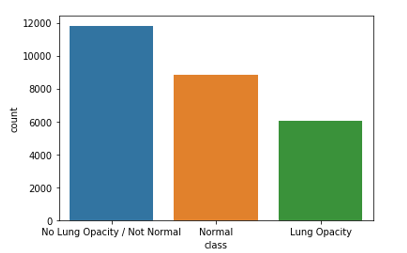
#### Exploring Data

This dataset is already divided into Train and Test sets.

Train set has data related to 26684 unique patients.

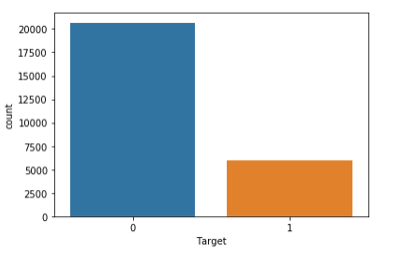
1. stage\_2\_detailed\_class\_info

* PatientId – Unique Id of the patient.
* Class – can take one of “No Lung Opacity / Not Normal”, “Normal”, “Lung Opacity”
* Classes “No Lung Opacity / Not Normal”, “Normal” have 1 row per patient in each of the csv files in train sets.
* Class “Lung Opacity” has one row when there is problem in only 1 lung and has 2 rows per patient when there is problem in both the lungs.



1. stage\_2\_train\_labels

* PatientId – Unique Id of the patient.
* X – Bounding Box xmin
* Y – Bounding Box ymin
* Width – Bounding Box width
* Height – Bounding Box width
* Target – 1 for patients with pneumonia and 0 for without pneumonia.



* For patients who have no pneumonia, there is only 1 row in each of the csv files in training set.
* For patients who have pneumonia, there could be a problem in 1 lung or 2 lungs for a patient. Hence there will be 1 row if there is problem in 1 lung and 2 rows if there is problem in both the lungs in each of the csv files. This needs to be considered when counting positive pneumonia cases.

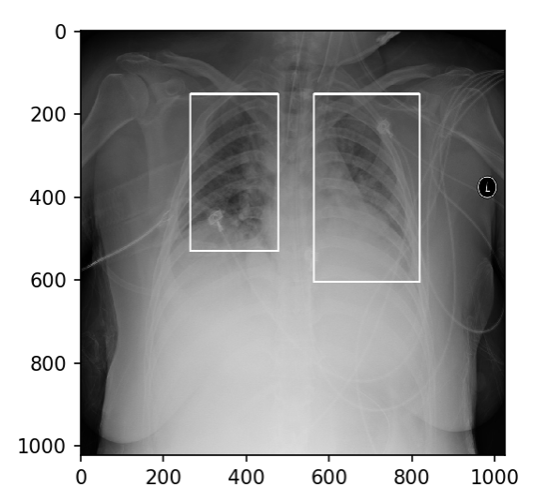
1. Dicom images

* These images are of size 1024 X 1024 each.
* There is 1 X-ray per patient.

### Visualization of sample images:

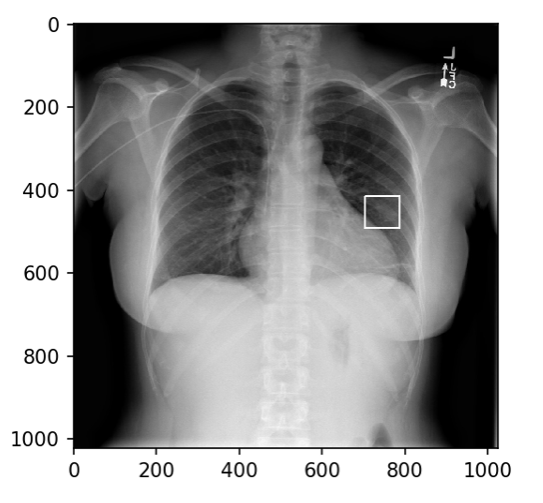
**Sample image with opacity in both the lungs:**

PatientId **-** 00436515-870c-4b36-a041-de91049b9ab4



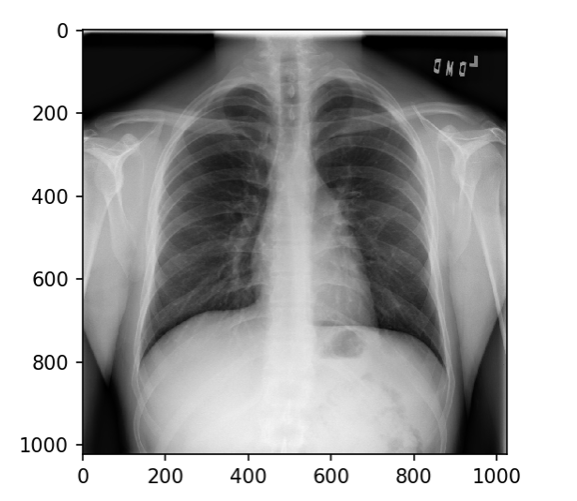
**Sample image with opacity in only 1 lung:**

PatientId **-** 0100515c-5204-4f31-98e0-f35e4b00004a



**Sample image without opacity:**

PatientId - 0004cfab-14fd-4e49-80ba-63a80b6bddd6



# Preprocessing data

### Duplicates

1. **stage\_2\_detailed\_class\_info** table has duplicate values for patients who have lung opacity in both of their lungs. **stage\_2\_train\_labels** table has 2 rows for the same patients who have lung opacity in both the lungs.
2. **stage\_2\_train\_labels** has 2 classes where **stage\_2\_detailed\_class\_info** has 3 classes.

### Missing Values

1. There are 4 fields that have missing values.
   1. **X**
   2. **Y**
   3. **Width**
   4. **Height**
2. These values are missing for “Normal” and “No Lung Opacity / Not Normal” class of patient records. Fill these values with “0”.

# Model Training:

Data is processed based on the model chosen for the problem. The model chosen for this is Mask R CNN model from [MD.ai](https://www.md.ai/) which is popular for models in the medical field.

### Model: Resnet (based of Mask R CNN)

#### About Mask R-CNN

The Mask R-CNN algorithm was introduced by He et al. in their 2017 paper, [Mask R-CNN](https://arxiv.org/abs/1703.06870).

Mask R-CNN builds on the previous object detection work of [R-CNN](https://arxiv.org/abs/1311.2524) (2013), [Fast R-CNN](https://arxiv.org/abs/1504.08083) (2015), and [Faster R-CNN](https://arxiv.org/abs/1506.01497) (2015), all by Girshick et al.

Detector Dataset is inherited from the dataset in Mask R-CNN algorithm.

To utilize this algorithm for the current problem, add the required class to existing classifications.

Train and validation data are prepared using this dataset.

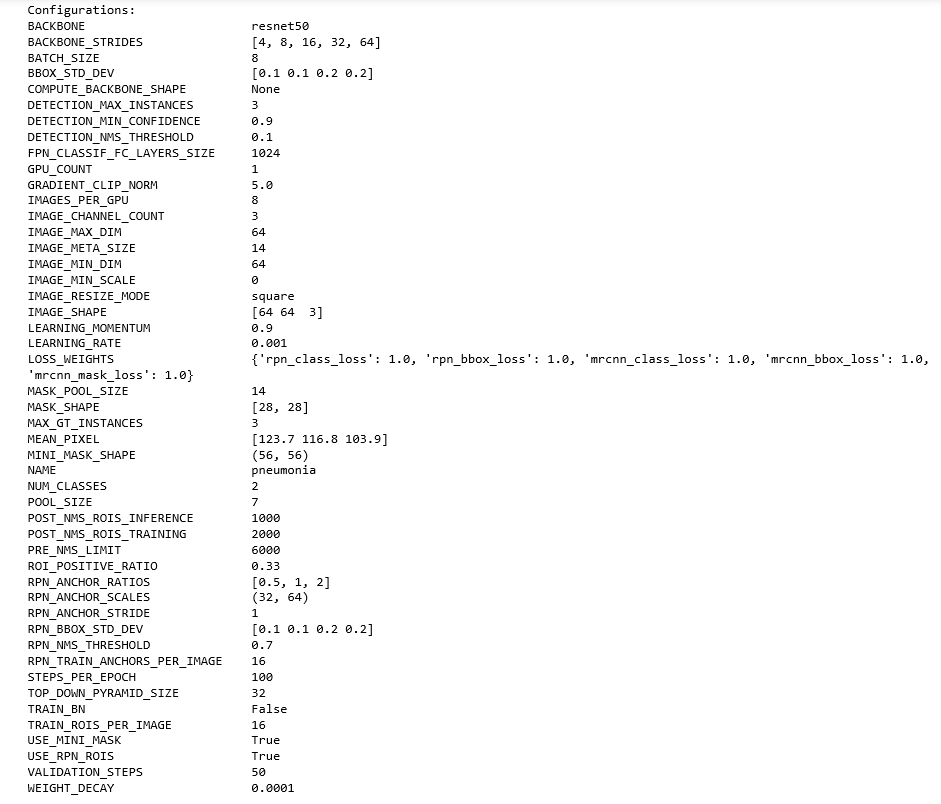
Train and validation data ratio are set to 70:30 ratio.

Images are also augmented based on the recommendations of MD.ai.

There are 2 underlying basic algorithms to choose from. They can be set in the BACKBONE property.

The 2 options are “resnet50” and “resnet101”.

#### Initial Configurations used for model training:



Mask R-CNN also allows to customize the number of layers that can be included in model training.

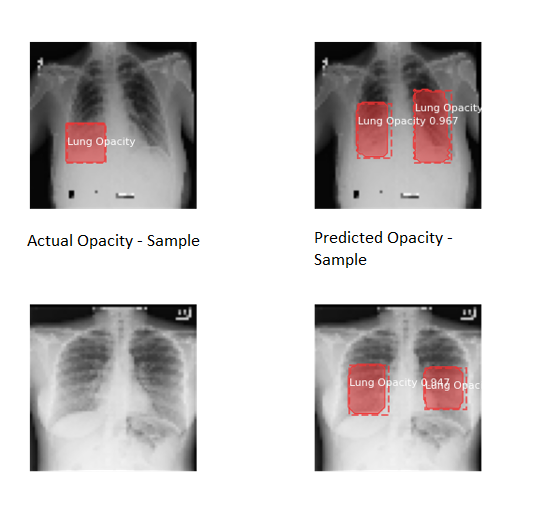
The initial training was done using all layers in the model and it was run for 5 epochs.

**Comparison of the Actual versus Predicted opacity for 2 samples:**

Below is the visualization of the X-ray with actual opacity and predicted opacities.

As can be noticed, only the first sample 1 prediction is close to the actual opacity.

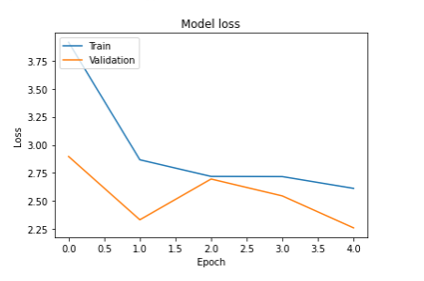
There is no opacity in the second sample, but the algorithm predicted opacity regions in both the lungs.



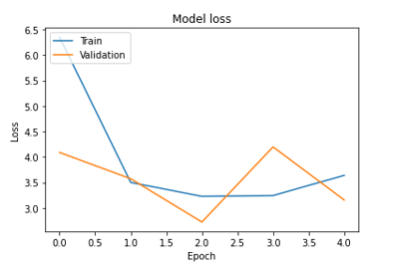
#### Trials for improvement in next iterations:

|  |  |
| --- | --- |
| **Initial plan for next iterations** | **Actual implementation or observation** |
| Classifying the image into one of the 3 classes and then predicting the opacity regions only for the images with lung opacity. | With the given dataset format and model used, it wasn’t possible to classify the images instead of using masks. This was an assumption and could have been avoided on clearly reading the documentation of MD. |
| Adding 3 classes instead of only 1 class to the DetectorDataSet. | Using 3 classes instead of using Target – 0/1 in the DetectorDataSet was giving errors because there are no masks available for “abnormal” class of lungs. |
| Tuning or changing the configurations to use a different BACKBONE, learning rates and other configurations. | Tried Resnet 101, and learning rate 0.005 |
| Changing the number of layers used in the model. | Tried Layers only Head and Layers 3+ |

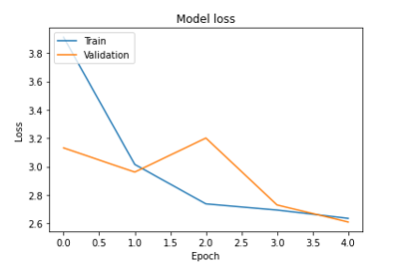
#### Resnet 101 – learning rate – 0.01 - all layers



#### Resnet 101 – head layers



#### Resnet 101 – learning rate 0.001 – all layers



### Conclusion on this model:

1. The loss has not decreased, or the model performance has not shown a good improvement with various options used.
2. Also, was not able to measure accuracy also in model building.

# Alternative Model: MobileNet

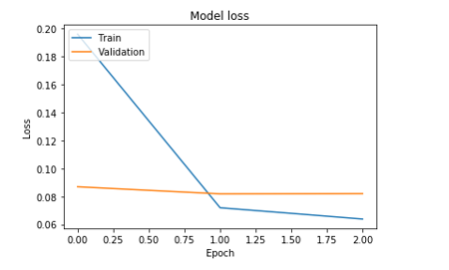
### Preprocessing data for this model:

1. Train and target(masks) arrays are created.
2. Data is split into train and validation sets – 80:20.

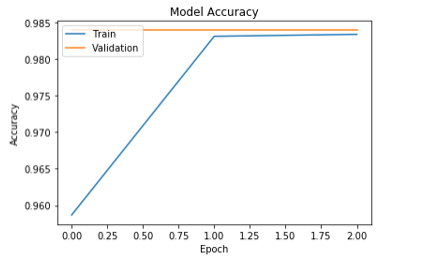
### Model building:

A custom model is built using transfer learning and u-net architecture.

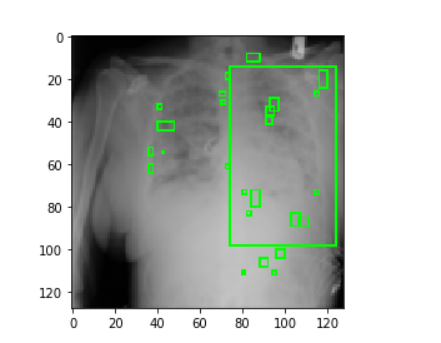
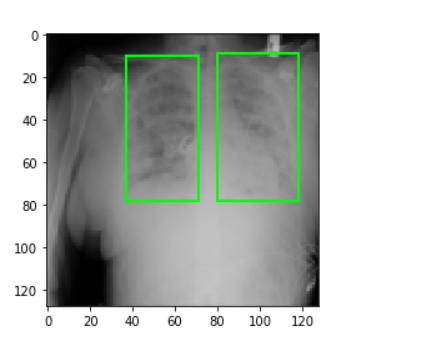
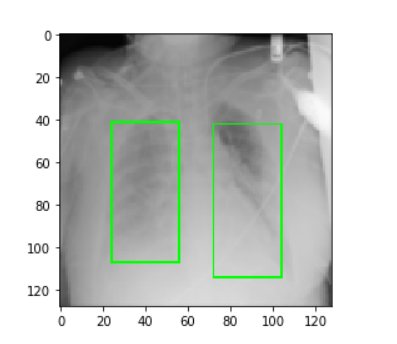
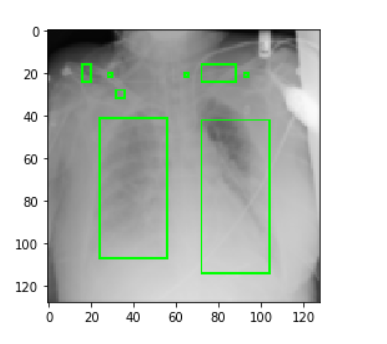
**Loss:**

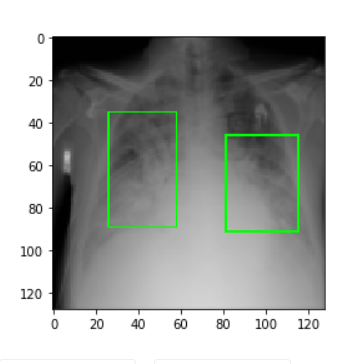
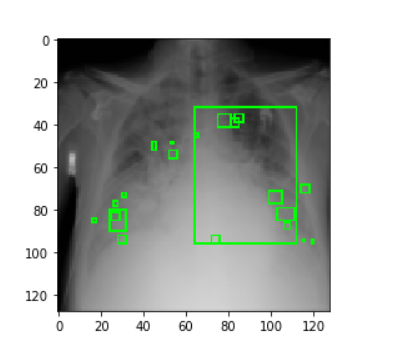


**Accuracy:**



Actual vs Predicted for few Samples:

### Conclusion on Mobilenet model:

This model did way better than Mask R CNN in this case.