

|  |
| --- |
| **Pneumonia Detection Challenge** |
| **Interim Report** |

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

|  |  |
| --- | --- |
| PRESENTED TO | Great Learning |
| PRESENTED BY | Group 13 – Aug C – Great Learning |
| DATE | 24 July 2020 |
| VERSION | 1.0 |

## Group 13 – Contacts

|  |  |  |
| --- | --- | --- |
| Name | Phone | Email |
| Sonal Sharma | +91 88102 64764 | [Sonal\_1996@yahoo.com](mailto:Sonal_1996@yahoo.com) |
| Chandrashekar C | +91 98867 94586 | [chandrashekar.c@gmail.com](mailto:chandrashekar.c@gmail.com) |
| Susanta Mondal | +91 70440 55951 | Susanta.221285@gmail.com |
| Gopinath Bailur | +91 99400 78996 | [gbailur@gmail.com](mailto:gbailur@gmail.com) |

|  |  |  |  |
| --- | --- | --- | --- |
| CHANGE LOG | | | |
| DATE | VERSION | AUTHOR | CHANGE DESCRIPTION |
| 19 June 2020 | 1.0 | Anuradha Raju | Initial Draft |

**Table of Contents**

[1 Executive Summary 4](#_Toc43477879)

[2 Scope of Work 4](#_Toc43477884)

[3 Pricing 4](#_Toc43477885)

[4 Payment Milestones 6](#_Toc43477886)

[5 Infrastructure 6](#_Toc43477887)

[6 Security Management 6](#_Toc43477888)

[7 Responsibility Matrix 7](#_Toc43477889)

[8 inContact Responsibilities 8](#_Toc43477890)

[9 Servion Responsibilities 8](#_Toc43477891)

[10 Understandings and Dependencies 8](#_Toc43477892)

[11 Escalation Matrix 8](#_Toc43477893)

[12 Terms & Conditions 9](#_Toc43477894)

[11 Signature Block 10](#_Toc43477895)

[I. Exhibit 1 – Servion’s ISMS policy 11](#_Toc43477896)

# Overview

# Pneumonia is a form of acute respiratory infection that affects the lungs. The lungs are made up of small sacs called alveoli, which fill with air when a healthy person breathes. When an individual has pneumonia, the alveoli are filled with pus and fluid, which makes breathing painful and limits oxygen intake.

# Following are some of the key facts about Pnuemonia which needs at most attention to address the problem proactively.

* Pneumonia accounts for 15% of all deaths of children under 5 years old, killing 808 694 children in 2017.
* Pneumonia can be caused by viruses, bacteria, or fungi.
* Pneumonia can be prevented by immunization, adequate nutrition, and by addressing environmental factors.
* Pneumonia caused by bacteria can be treated with antibiotics, but only one third of children with pneumonia receive the antibiotics they need.

# Per WHO, Pneumonia is the single largest infectious cause of death in children worldwide. Pneumonia killed 808 694 children under the age of 5 in 2017, accounting for 15% of all deaths of children under five years old. Pneumonia affects children and families everywhere, but is most prevalent in South Asia and sub-Saharan Africa. Children can be protected from pneumonia, it can be prevented with simple interventions, and treated with low-cost, low-tech medication and care.

# The WHO and UNICEF integrated Global action plan for pneumonia and diarrhoea (GAPPD) aims to accelerate pneumonia control with a combination of interventions to protect, prevent, and treat pneumonia in children with actions to:

* protect children from pneumonia including promoting exclusive breastfeeding and adequate complementary feeding;
* prevent pneumonia with vaccinations, hand washing with soap, reducing household air pollution, HIV prevention and cotrimoxazole prophylaxis for HIV-infected and exposed children;
* treat pneumonia focusing on making sure that every sick child has access to the right kind of care -- either from a community-based health worker, or in a health facility if the disease is severe -- and can get the antibiotics and oxygen they need to get well;

# Based on research we have made an attempt to look at the Chest Radiography to identify the Lung Opacity and quickly help Clinical specialists to take right decisions to drive proactive measure to cure and help avoid the spread of this decease to larger extent.

# Abstract

This project is aimed at detecting Pneumonia by locating the lung opacities on the Chest radiographs. This process can help identify the problem at an early stage as well as helps the Clinical analysis much faster and drives better decision making. This process of Pneumonia detection will be done by looking at several thousand images of Chest radiographs taken from past wherein the analysis and desired results were identified by the specialists. These past datapoints will become the indicators and these images will be processed through the Computer Vision Technology of deep learning to capture every details by which a Deep Learning Algorithm will be built.

The new patients data will be fed to this model which detects the Lung Opacity indication along with its location such that Clinical specialists will be able to confirm diagnosis quickly and can help in taking respective decisions quickly to move forward with the next steps of the treatment.

Deep neural networks models have conventionally been designed and experiments were performed upon them by human experts in a continuing trial and error method. This process demands enormous time, knowhow and resources. To overcome this problem, a novel but simple model is introduced to automatically perform optimal classification tasks with deep neural network architecture .

The Neural network architecture was specifically designed for Pneumonia image classification tasks. The proposed technique is based on the CNN algorithm, utilizing set of neurons to convolve on a given sample images to extract relevant features from them. This is demonstrated through validating the accuracy of the detection along with the objective to reduce the loss while the network is learning the details.

As part of this project, we will be demonstrating the outcome with 4 different model architecture along with their outcome in each of the model and provide the commentary for each of the models developed.

# Project Objective

The objective of this capstone project is to build a Pneumonia detection system to locate the position of the inflammation in a Chest Radiography.

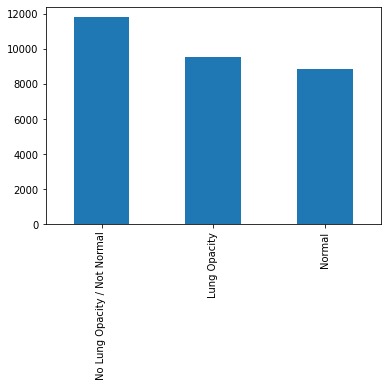
Based on the dataset provided, we will have to do the following steps to work towards building the final model and validate and results.

* Merge the csv files to arrive at the sample dataset to completion and along the line build the merged CSV files along with the right x,y, width and height bounding box coordinates along with the patient ID.
* Validate the images and respective bounding box coordinates.
* Extract all the features from the images and build the csv file for processing further.
* Perform Exploratory Data Analysis to validate all the data points and build insights
* Based on the project objective – isolate the dataset and accordingly images as well which are fully unique by removing the duplicate records in the dataset based on target variable.
* Build the model with different architectures and showcase the accuracy and showcase the right model to approach the problem description

# EDA Inference and Data pre-processing

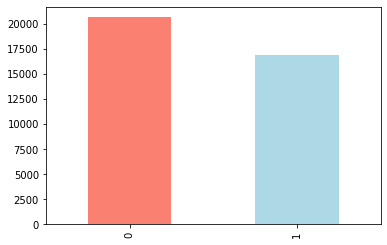
|  |
| --- |
|  |

1. Based on the sample dataset provided – we see that there are 3 class values present..



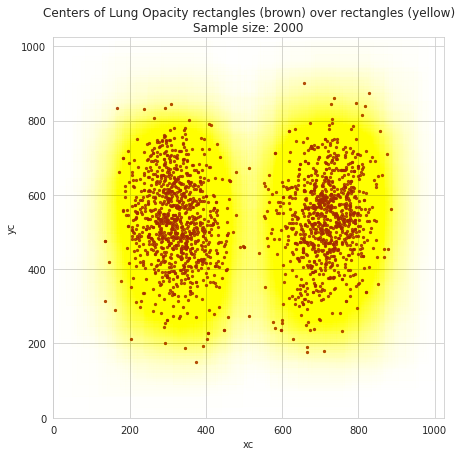
Our objective would be look at images having Lung Opacity and identify the right bounding box to locate the area of inflammation, hence we may need to club them into Lung Opacity and others as another single category

1. Post merging both the documents – class information with the labels document we see the following distribution on the “Target” variable



We observe there are more records not having lung opacity issues, which may lead to bias.

1. We observe the following heat map making the visualization where in the Target = 1



The scatter plot with brown dots shows the areas of inflammation highlighted which needs to be learnt to identify the issues.

1. Following are set of Metadata that we are able to notice from each of the chest X-ray’s provided..

Dataset.file\_meta -------------------------------

(0002, 0000) File Meta Information Group Length UL: 202

(0002, 0001) File Meta Information Version OB: b'\x00\x01'

(0002, 0002) Media Storage SOP Class UID UI: Secondary Capture Image Storage

(0002, 0003) Media Storage SOP Instance UID UI: 1.2.276.0.7230010.3.1.4.8323329.28530.1517874485.775526

(0002, 0010) Transfer Syntax UID UI: JPEG Baseline (Process 1)

(0002, 0012) Implementation Class UID UI: 1.2.276.0.7230010.3.0.3.6.0

(0002, 0013) Implementation Version Name SH: 'OFFIS\_DCMTK\_360'

-------------------------------------------------

(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

1. Of the above metadata – basis the observation of key fields names and its values, we decided to consider following 16 metadata to form the CSV file for validating and processing further..

        idx = (data\_df['patientId']==data\_row\_img\_data.PatientID)

        data\_df.loc[idx,'Modality'] = data\_row\_img\_data.Modality

        data\_df.loc[idx,'PatientAge'] = pd.to\_numeric(data\_row\_img\_data.PatientAge)

        data\_df.loc[idx,'PatientSex'] = data\_row\_img\_data.PatientSex

        data\_df.loc[idx,'BodyPartExamined'] = data\_row\_img\_data.BodyPartExamined

        data\_df.loc[idx,'ViewPosition'] = data\_row\_img\_data.ViewPosition

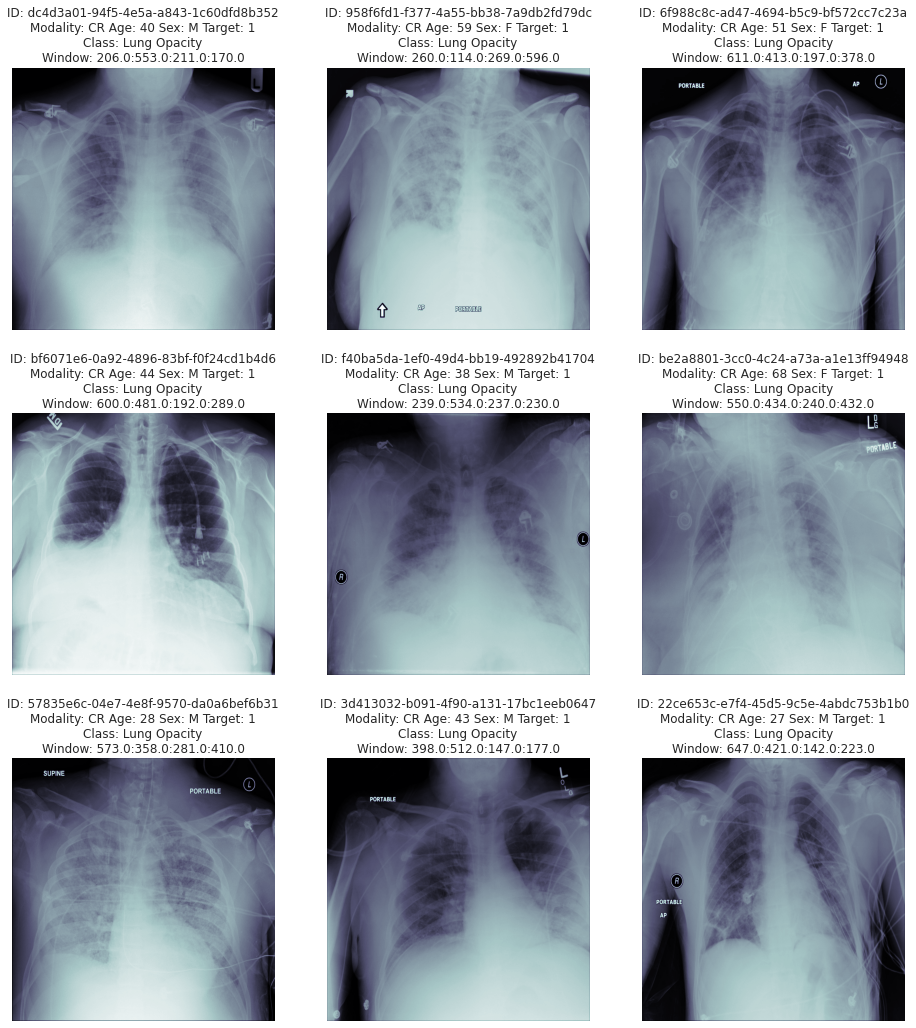
        data\_df.loc[idx,'ConversionType'] = data\_row\_img\_data.ConversionType

        data\_df.loc[idx,'Rows'] = data\_row\_img\_data.Rows

        data\_df.loc[idx,'Columns'] = data\_row\_img\_data.Columns

        data\_df.loc[idx,'PixelSpacing'] = str.format("{:4.3f}",data\_row\_img\_data.PixelSpacing[0])

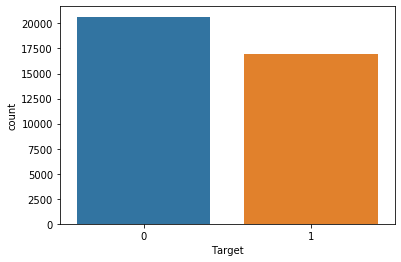
1. Processing the dicom images, we see the following



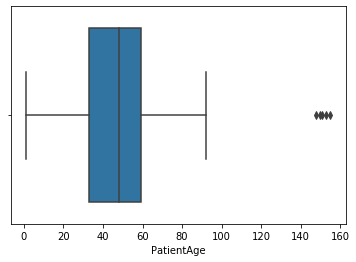
We observe all the images are 1024/1024 shape and it needs to be reshaped before feeding them into the model.

Based on the processed CSV file from the metadata – EDA was performed on this and below are the findings and also pre-processing steps taken to perform the data analysis

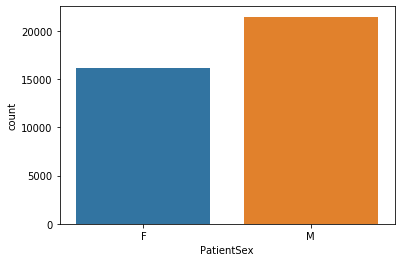
* EDA performed on the data taken out from each of parameters extracted from the images and clubbed against their patient ID's to validate them for its accuracy, impact based on availability of information overall. Below are some of the key findings.
* There are totally 37,629 records available in the dataset where in 16 fields were extracted for validation
* For Pnuemonia detection - "Target" Field provides the classification. 0 : No Lung Opacity, 1 : Lung Opacity



* "class" Field provides 3 groups. However when its seen against Target classification and respective coordinates availability - its doesn't seems to isolate between Normal and Not Normal cases. There are may be other problems, for the purpose of this exercise this observation will be ignored
* "Target" has 20,672 records/images which are Not having Lung Opacity and 16,957 having lung Opacity. We observe 55% of the images doesn't have lung opacity and only 45% having Lung Opacity - This may create imbalance in prediction tending towards not having Lung Opacity - Need to observe this furhter for duplicate records and plan for data augmentation
* "PatientAge" - We observe 5 records having ages above 100 - which should be dropped



* "PatientSex" - We observe 55% of male records in the total count. This may not be a factor for recognition hence decided to continue even though there is imbalance..

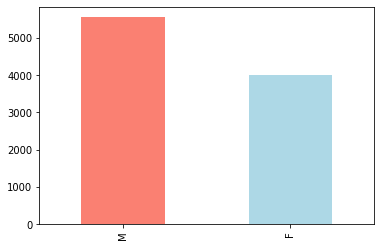


* We observe 9555 records are unique - which needs to be considered for model building
* For processing further – we have taken the ROWS having Target = 1, we found that there are 16K records
* Post processing for Duplicate records – we see that the record count comes to 9555 records..

1 9555

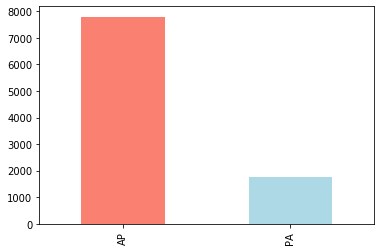
Name: Target, dtype: int64

* Of the 9555 records – below is the distribution on M / F



Since Male records are more – this may create data imbalance even though it may not be crucial at this time this needs to be validated.

* With the column “ViewPosition” we notice the below findings..



### With respect to Column "ViewPosition" correlated with the Target value of 1 - There is high imbalance that we observe with the above information having AP (Anterior-Posterior). One way to interpret this target unbalance is that patients that are imaged in an AP position are those that are more ill, and therefore more likely to have contracted pneumonia. Note that the absolute split between AP and PA images is about 50-50, so the above consideration is extremely significant

* Validating the target with age group – we observe the below pattern.

PatientAge count

0 0-10 219

1 11-20 602

2 21-30 1308

3 31-40 1591

4 41-50 1675

5 51-60 2226

6 61-70 1313

7 71-80 518

8 81-90 100

9 91-100 3

10 more than 100 0

Between age from 21 to 70 – The problem is highly detected and specifically 51 to 60 has more occurrence. We also observe some of the outlier which may needs to be dropped.

# Current Approach on Model building

Deep neural networks models have conventionally been designed and experiments were performed upon them by human experts in a continuing trial and error method. This process demands enormous time, knowhow and resources. To overcome this problem, a novel but simple model is introduced to automatically

YOLO

MobileNet

SSD

Mask R-CNN

# Approach to improve model performance