Capstone project Pneumonia detection
(Batch 21B, Group 6)

Team:

1) Ravi A
2) KamalKishor Pande
3) Manas Haldar
4) Rahul Gupta
5) Akbar Boghani

Table of Contents

1 INTRODUCTION 3

2 PNEUMONIA KEY FACTS[1] 3

3 HOW PNEUMONIA IS DETECTED? 4

4 WHY PNEUMONIA DETECTION IS CHALLENGING? 5

5 DATABASES 6

6 EDA 6

6.1 NULL AND DUPLICATE VALUES 7
6.2 EXAMINING THE IMAGES 9
6.3 GENDER V/S TARGET VALUES 10
6.4 GENDER WISE LUNG CONDITION 11
6.5 AGE WISE DISTRIBUTION PLOT 12
6.6 CHECKING FOR OUTLIERS 12

7 MODEL BUILDING. 13

7.1 VARIOUS MODEL ITERATIONS TRIED 13

8 CONCLUSIONS 14

9 CHALLENGES 14

10 NEXT STEPS 14

11 LIST OF FIGURES 15

12 REFERENCES 15

1 Introduction

This document is a summary of the research papers on Pneumonia detection and our approach to building, testing and refining the Machine learning model for Pneumonia detection.

The project is to build a machine learning model to detect pneumonia based on the X-ray Images.

Objective:

- To undertake a multi-faceted project that demonstrates your understanding and mastery of the key conceptual and technological aspects of Deep Learning.
- To develop an understanding of how challenging human-level problems can be approached and solved using a combination of tools and techniques.
- To understand current scenarios in deep learning, understand the practicalities and the trade-offs that need to be made when solving a problem in real life.

2 Pneumonia Key facts[1]

1. Pneumonia is an infection of the lung. The lungs fill with fluid and make breathing difficult.
2. Pneumonia is the world's leading cause of death among children under 5 years of age, accounting for 16% of all deaths of children under 5 years
3. In the US, pneumonia is less often fatal for children, but it is still a big problem. Pneumonia is the #1 most common reason for US children to be hospitalized.
4. For US adults, pneumonia is the most common cause of hospital admissions other than women giving birth. About 1 million adults in the US seek care in a hospital due to pneumonia every year, and 50,000 die from this disease.
5. While young healthy adults have less risk of pneumonia than the age extremes, it is always a threat.
6. Older people have higher risk of getting pneumonia, and are more likely to die from it if they do. It greater risk of death compared to any of the other top 10 reasons for hospitalization.
7. Pneumonia is the most common cause of sepsis and septic shock, causing 50% of all episodes.
8. Pneumonia can develop in patients already in the hospital for other reasons. Hospital-acquired pneumonia has a higher mortality rate than any other hospital-acquired infection.
9. Pneumonia can be caused by lots of different types of microbes, and no single one is responsible. For most pneumonia patients, the microbe causing the infection is never identified.
10. In the US and the rest of the world, viral pneumonias are the leading cause of hospitalization of infants.
11. Antibiotics can be effective for many of the bacteria that cause pneumonia. For viral causes of pneumonia, antibiotics are ineffective and

should not be used. There are few or no treatments for most viral causes of pneumonia.

12. Antibiotic resistance is growing amongst the bacteria that cause pneumonia. This often arises from the overuse and misuse of antibiotics in and out of the hospital. New and more effective antibiotics are urgently needed.
13. Our changing interactions with the microbial world mean constantly developing new pneumonia risks.
14. Patients with pneumonia may need to be hospitalized or even go to the intensive care unit (ICU). After developing pneumonia, it often takes 6-8 weeks until a patient returns to their normal level of functioning and wellbeing.
15. While successful pneumonia treatment often leads to full recovery, it can have longer term consequences.
16. Pneumonia is a huge burden on our healthcare systems.
17. In the US, pneumonia was one of the top ten most expensive conditions seen during inpatient hospitalizations. In 2013, pneumonia had an aggregate cost of nearly $9.5 billion for 960,000 hospital stays.
18. The death rate from pneumonia in the US has had little improvement since antibiotics became widespread more than half a century ago. We are not yet winning the battle against pneumonia.
19. Pneumonia does not have an effective advocacy strategy. It is not the subject of fund-raising walks or runs.

3 How Pneumonia is detected?

The pneumonia detection is commonly performed by examining chest X-Ray radiograph (CXR). It is indicated as an area or areas of increased opacity on CXR. the diagnosis is further confirmed through clinical history, vital signs and laboratory exams. [3]

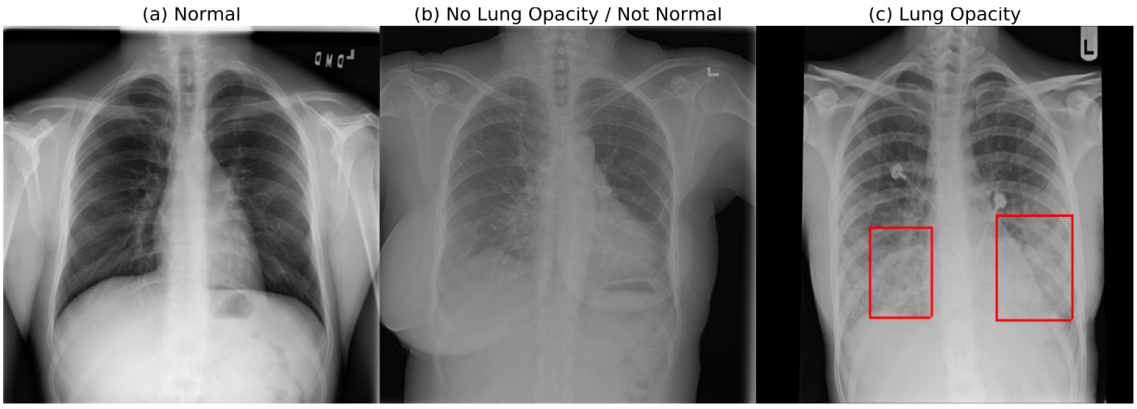


Figure 1. Examples of the chest X-Ray images for (a) ”Normal”, (b) ”No Lung Opacity / Not Normal”, and (c) ”Lung Opacity” cases. The lung opacities regions are shown on (c) with red bounding boxes.

Figure 01

4 Why Pneumonia Detection is Challenging?

The diagnosis of pneumonia on CXR is complicated because of a number of other conditions in the lungs, such as fluid overload, bleeding, volume loss, lung cancer, post-radiation or surgical changes. When available, comparison of CXRs of the patient taken at different time points and correlation with clinical symptoms and history is helpful in making the diagnosis.

A number of factors such as positioning of the patient and depth of inspiration can alter the appearance of the CXR. In addition, clinicians are faced with reading high volumes of images every shift.

Apart from Pneumonia, there are 7 other thorasic diseases observed in the Chest X-ray. These are

1. Atelectasis: complete or partial collapse of the entire lung or area (lobe) of the lung
2. Cardiomegaly: An enlarged heart (cardiomegaly) isn't a disease, but rather a sign of another condition.
3. Effusion: fluid leaking into the pleural space. This is from increased pressure in the blood vessels or a low blood protein count. Heart failure is the most common cause
4. Infiltration : A finding indicating the presence of an inflammatory or neoplastic cellular infiltrate in the lung parenchyma.
5. Mass: A lung mass is defined as an abnormal spot or area in the lungs larger than 3 centimeters (cm), about 1.5 inches, in size.
6. Nodule: Spots smaller than 3 cm in diameter are considered lung nodules.
7. Pneumothorax : A pneumothorax occurs when air leaks into the space between your lung and chest wall. This air pushes on the outside of your lung and makes it collapse

Following figure indicates the X-ray images of eight lung conditions [2}

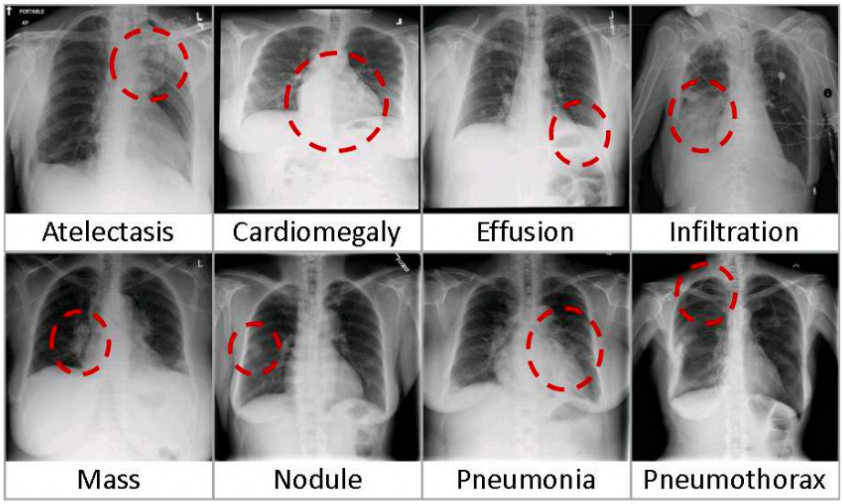


Figure 1. *Eight common thoracic diseases observed in chest X-rays that validate a challenging task of fully-automated diagnosis.*

Figure 02

These eight diseases may not occur in isolation, there is a possibility that one condition leads to another. Following figure indicates the co-occurrence of these eight diseases [2]

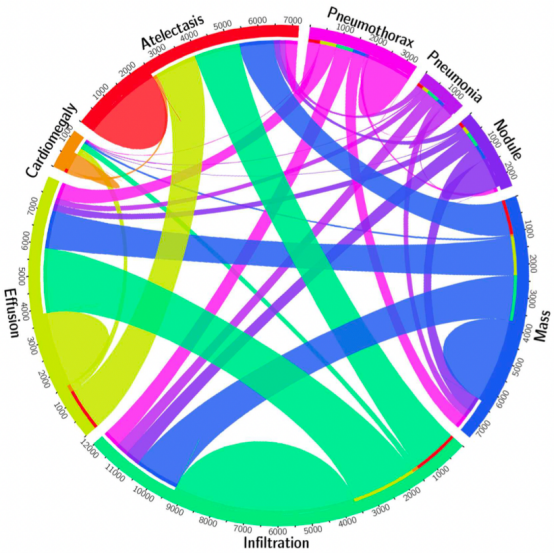


Figure 2. *The circular diagram shows the proportions of images with multi-labels in each of 8 pathology classes and the labels' co-occurrence statistics.*

Figure 3

5 Databases

We will be using the database from the source.

https://www.kaggle.com/c/rsna-pneumonia-detection-challenge/data

The RSNA is an international society of radiologists, medical physicists and other medical professionals with more than 54,000 members from 146 countries across the globe. They see the potential for ML to automate initial detection (imaging screening) of potential pneumonia cases in order to prioritize and expedite their review.

Challenge participants may be invited to present their AI models and methodologies during an award ceremony at the RSNA Annual Meeting which will be held in Chicago, Illinois, USA, from November 25-30, 2018.

6 EDA

Data: Following information is available

1) There are image files in DICOM format.
There are two sets of images. Number of training images are and test images 26000 are 3000

2) stage_2_train_labels.csv. this has patient id and the object information (x, y, height, width) and label that indicates whether patient has pneumonia, or NO?
This file has 30227 records and 6 attributes.

```
#   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   patientId   30227 non-null  object
 1   x           9555 non-null   float64
 2   y           9555 non-null   float64
 3   width       9555 non-null   float64
 4   height      9555 non-null   float64
 5   Target      30227 non-null  int64
```

Target column has two distinct values ‘0’ and ‘1’. Patients with Target Value '0' are 'non-Pneumonia' Patients. This constitutes 31.6 %

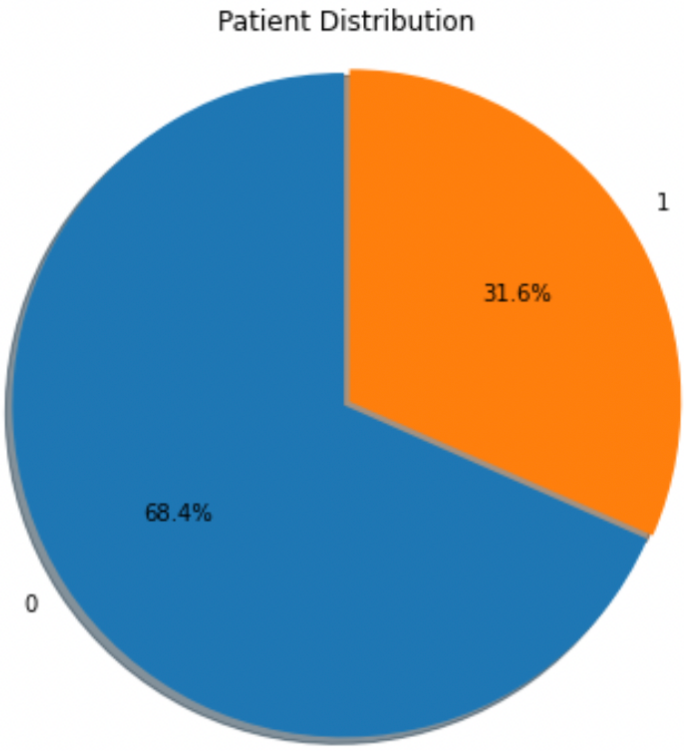


Figure 04

6.1 Null and duplicate values

There are no null values. There are 3543 duplicate records. This may be because there may be multiple X-ray images from different angle OR at different time period. We will retain this information.

3) stage_2_detailed_class_info.csv. this file has the patient id and the lung condition whether it is normal and/ or opaque.

The file has 30227 rows and two columns patient id and class

```
Data columns (total 2 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   patientId  30227 non-null  object
 1   class      30227 non-null  object
dtypes: object(2)
memory usage: 472.4+ KB
```

The data across three classes are fairly distributed.

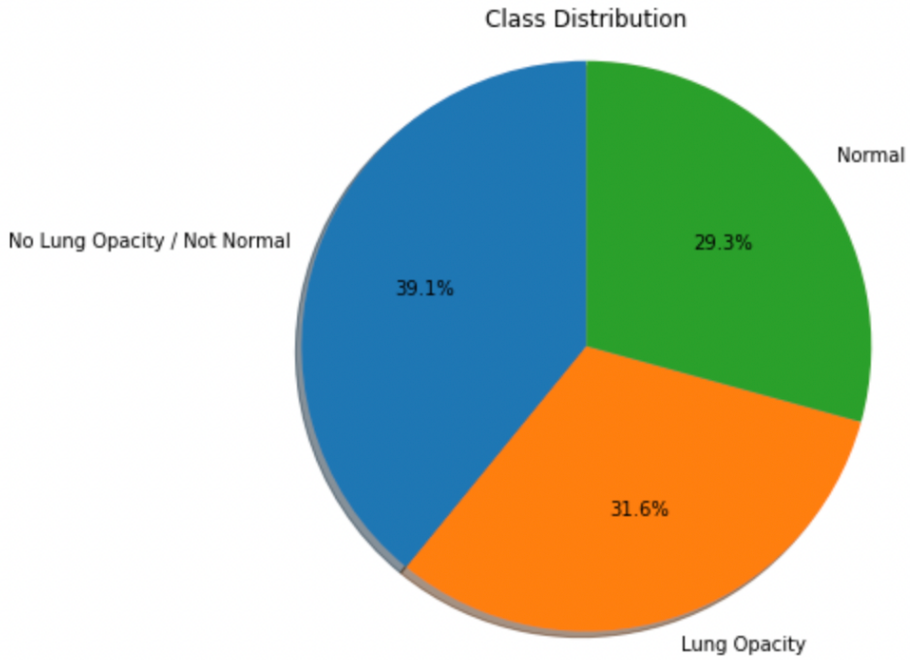


Figure 05

```
No Lung Opacity / Not Normal    11821
Lung Opacity                     9555
Normal                           8851
```

Above two files has the common key patientid. We have merged the datasets based on the patientid and created a new dataframe.

A DICOM file consists of a header and image data sets packed into a single file. The information within the header is organized as a constant and standardized series of tags. By extracting data from these tags one can access important information regarding the patient demographics, study parameters, etc.

We have extracted sex and age from the images files and populated the same in the dataframe. The new dataframe has the following structure now.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 30227 entries, 0 to 30226
Data columns (total 9 columns):
 #   Column     Non-Null Count  Dtype
```

```
---   ------      --------------  -----
 0    patientId   30227 non-null  object
 1    x           9555 non-null   float64
 2    y           9555 non-null   float64
 3    width       9555 non-null   float64
 4    height      9555 non-null   float64
 5    Target      30227 non-null  int64
 6    class       30227 non-null  object
 7    sex         30227 non-null  object
 8    age         30227 non-null  object
dtypes: float64(4), int64(1), object(4)
```

6.2 Examining the images

We have examined 9 X-ray images each for the patient with pneumonia and no-pneumonia.

X-ray Images of patients with pneumonia:

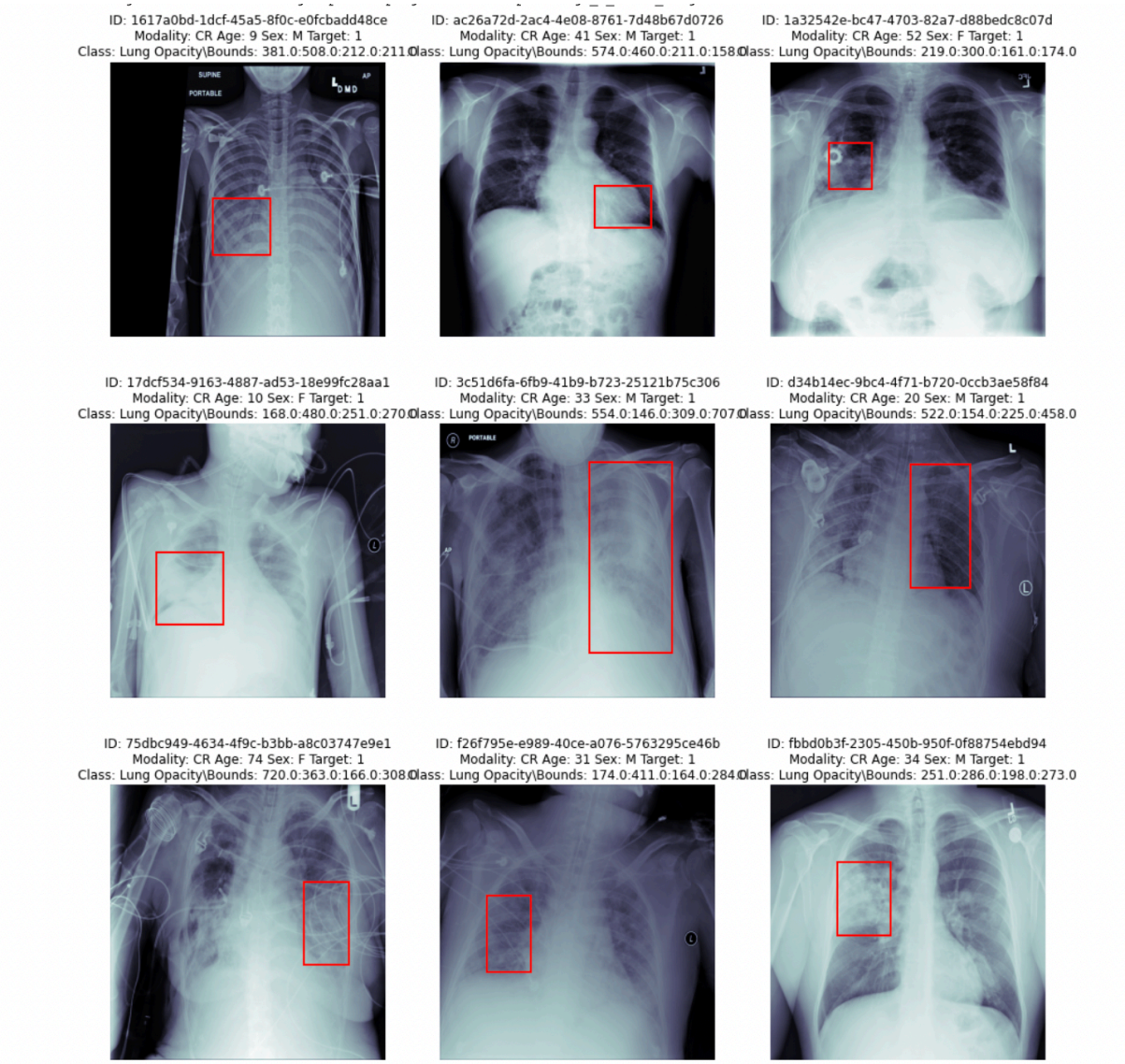


Figure 06

The box indicates the areas with opacity.

X-ray Images of patients with no-pneumonia:

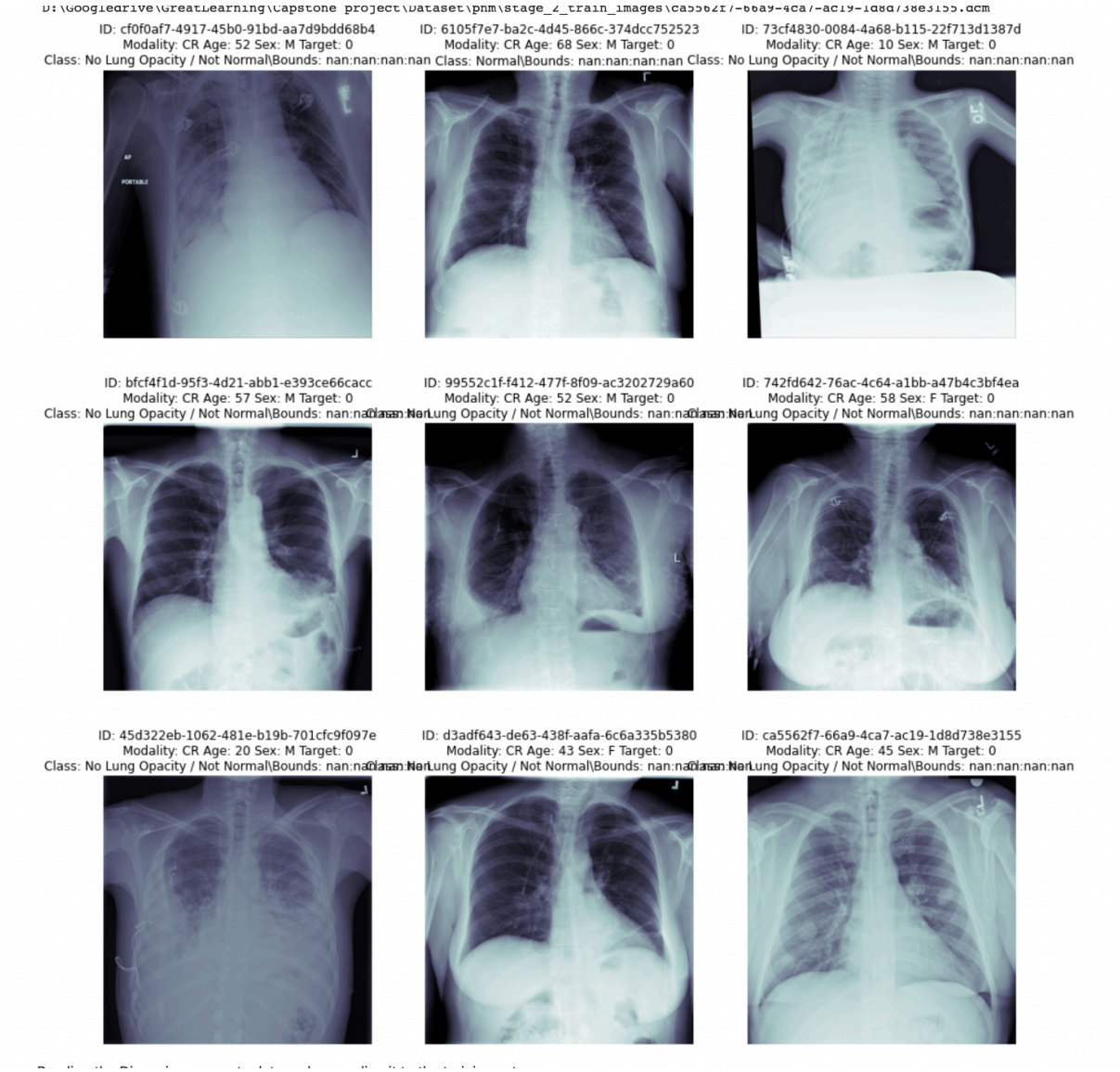


Figure 07

6.3 Gender v/s target values

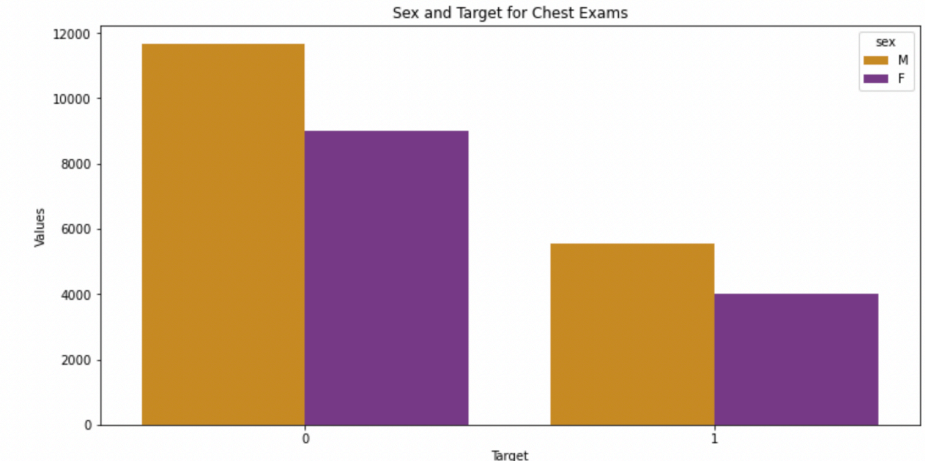


Figure 08

Pneumonia occurrence is higher in male

6.4 Gender wise lung condition

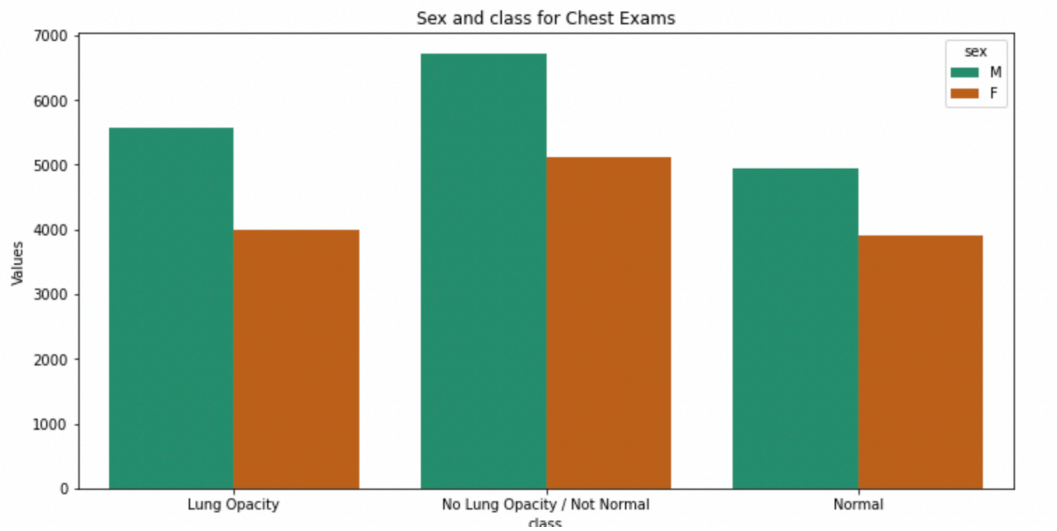


Figure 09

The histogram is quite similar to the previous one (gender vs target values)

6.5 Age wise distribution plot

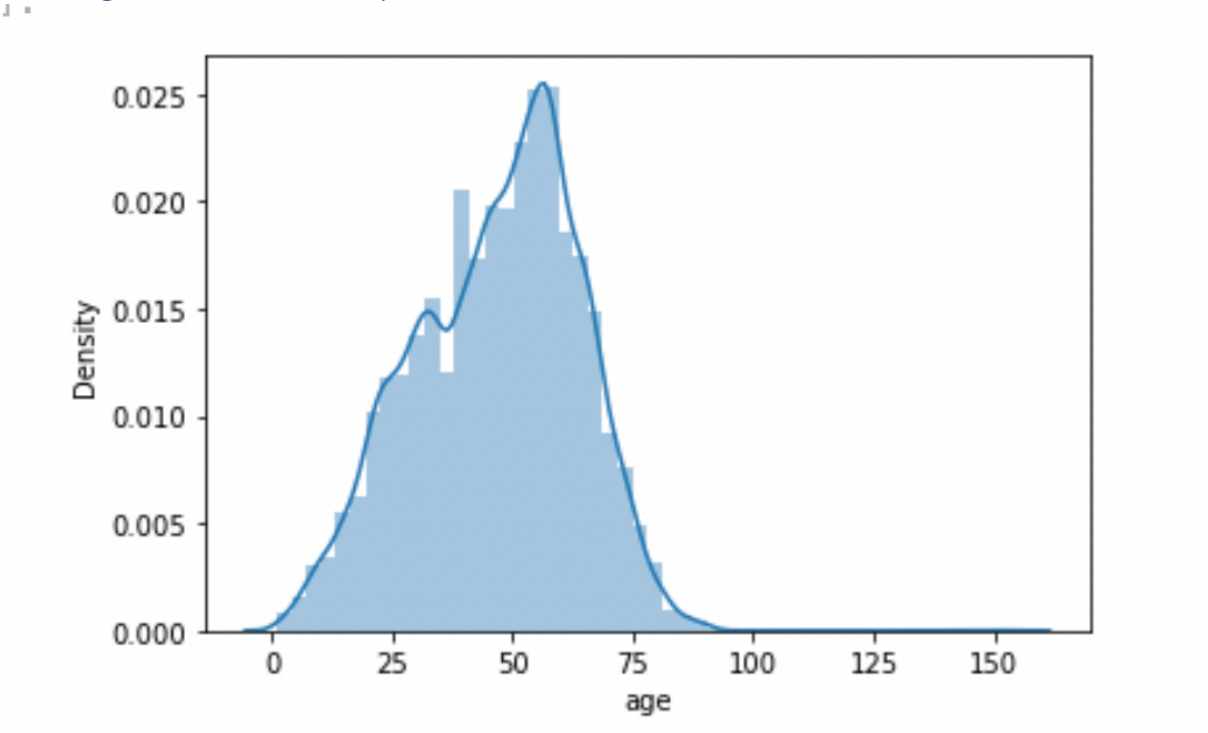
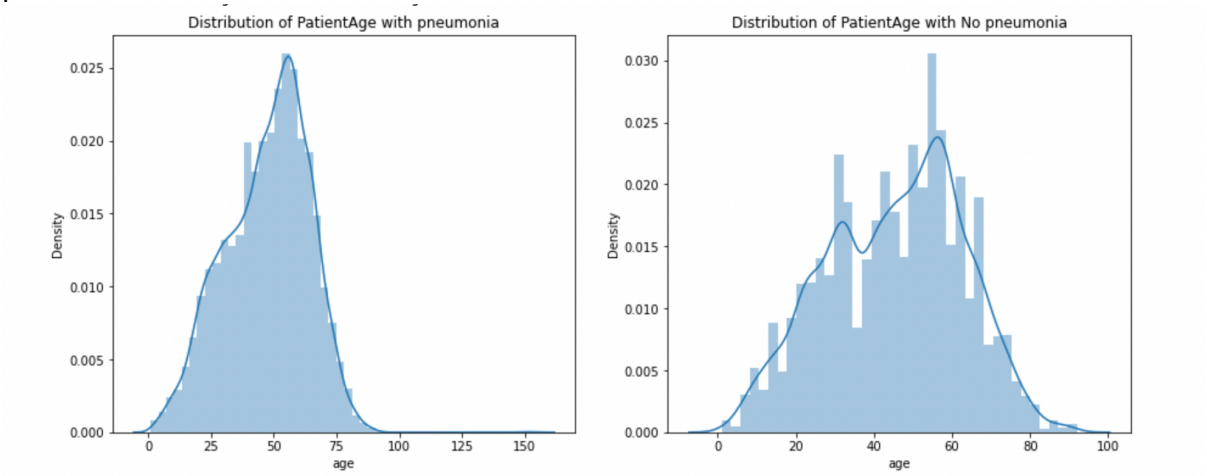


Figure 10

The plot is normal distribution with peak around the age of 55 years.

Following figure shows the separate distribution for the pneumonia and non-pneumonia patients.



Distribution of Patients who have pneumonia is slightly skewed. However, the Distribution of Patients who have No pneumonia is has two peaks around age 30 and 55

Figure 11

6.6 Checking for outliers

The box plot below is for the lung condition and the age. There are outliers for two classes, ‘normal’ and ‘no lung opacity’.

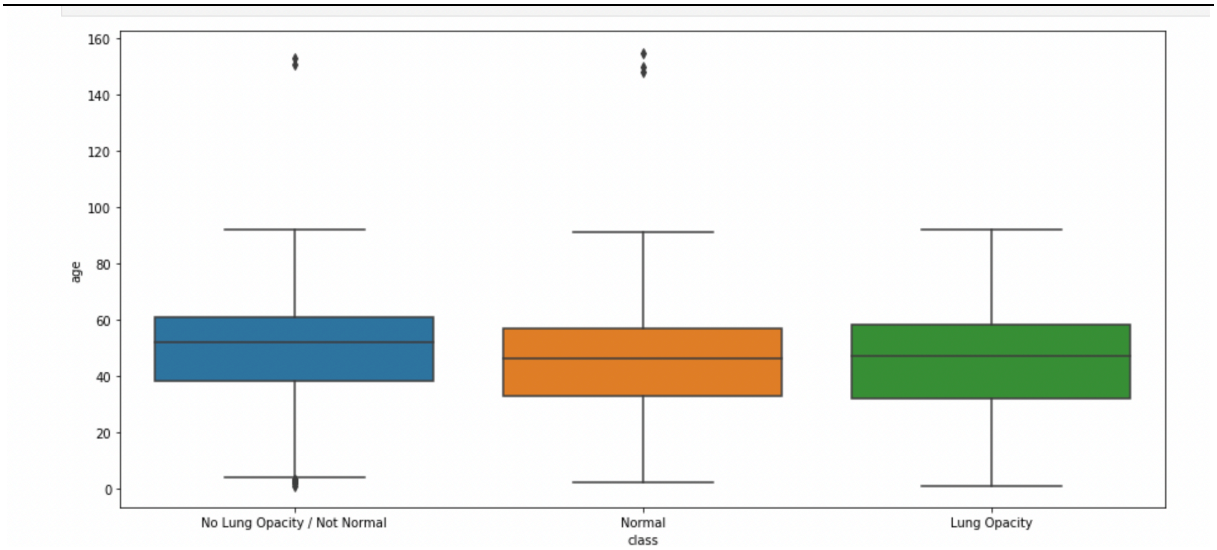


Figure 12

There are only 5 patients with age > 100. the same is observed in the box plot. lets ignore these 5 records as of now

7 Model Building.

This is a classification problem. We are using a CNN architecture to train the model with the provided training data. The model architecture has the following elements:
- 6 convolution layers
- 3 maxpool layers
- 4 dropout layers
- 1 output layer

We chose a CNN model ass this is a classification problem based out of image input data. CNN architecture works well with images as opposed to ANNs.

We begin with a random model, and try different iterations to achieve a better model architecture by changing the input filters, changing the kernel sizes across layers in the corresponding models, changing paddings, trying out different dropout percentages and running the model over different epochs and batch sizes.

The goal is to achieve an optimal architecture for the model that will train well with the given data and perform reasonably well with the test data.

After running through a few iterations, we have conceptualized a crude architecture for the model and we would like to enhance it further by tuning the hyperparameters.

7.1 Various model iterations tried

Model Run	Optimizer	Epochs	Batchsize	Decay	Row	Epsilon	LR	Trainable parameters
CNN1	RMSprop	10	100	0	0.9	1.00E-08	0.001	3,38,723

CNN2	RMSprop	15	80	0	0.9	1.00E-08	0.0001	3,38,723
CNN3	RMSprop	20	100	0	0.9	1.00E-07	0.0001	8,69,283
CNN4	RMSprop	10	100	0	0.9	1.00E-08	0.001	8,69,283
CNN5	Adam	20	100	0	0.9	1.00E-07	0.0001	2,76,083
CNN6	Adam	20	100	0	0.9	1.00E-07	0.001	2,76,083

Figure: 13

The outcome of each iteration is shown in the table below.

	Method	accuracy	Test Score	1_precision	1_recall	1_f1-score	1_support
0	CNN1	0.435714	0.413333	0.398625	0.958678	0.563107	121
0	CNN2	0.533571	0.423333	0.398551	0.454545	0.424710	121
0	CNN3	0.459286	0.423333	0.429688	0.454545	0.441767	121
0	CNN4	0.423571	0.446667	0.431138	0.595041	0.500000	121
0	CNN5	0.510000	0.573333	0.581633	0.471074	0.520548	121
0	CNN6	0.395000	0.400000	0.398625	0.958678	0.563107	121

Figure 14

Accuracy is between 39.5 and 53%. Recall is varying between 45% and 95%

8 Conclusions

1) We have executed 6 iterations with various combination optimizer, learning rate, epochs and batch size as mentioned in previous section.
2) Accuracy is between 39.5 and 53%. However, there is a huge variation in recall which we need to analyze further
3) iteration 2 and 5 are shows improvement in performance over the epochs, accuracy can further improve by increasing no of epochs.

9 Challenges

The dataset provided is huge and we were not able to load the data completely on the local machines. We have tried google colab as well, but we faced performance issue and timeout issues as well. So we decided to build model on sample data of 2000 images. However, we have ensured that the data selected has the comparable distribution with original dataset

10 Next steps

1) Next phase we will choose one model from the six iterations and tune it further
2) We will also experiment with other parameters like early stop, decay, row, epsilon and train on larger dataset

3) We will deploy transfer learning and compare the results with the previous models
4) Focus on most relevant performance parameters such as recall.
5) Train the model on the larger dataset.
6) Finally, we will improve the code by building more functions

11 List of Figures

Figure no	Description
Figure 1	Example of chest X-Ray with three
Figure 2	Chest X-ray with eight lung conditions
Figure 3	Proportion of images for eight different conditions
Figure 4	Patient distribution in the given dataset
Figure 5	Data distribution across three classes
Figure 6	Nine X-ray images for patients with Pneumonia
Figure 7	Nine X-ray images for patients with No Pneumonia
Figure 8	Gender v/s target values
Figure 9	Gender v/s lung condition
Figure 10	Age wise distribution plot
Figure 11	Age wise distribution plot for two target values
Figure 12	Box plot showing outliers for three lung conditions
Figure 13	List of various model iterations
Figure 14	Outcome of each model iteration

12 References

1) https://www.thoracic.org/patients/patient-resources/resources/top-pneumonia-facts.pdf
2) ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases (Xiaosong Wang[1], Yifan Peng [2], Le Lu [1], Zhiyong Lu [2], Mohammadhadi Bagheri [1], Ronald M. Summers [1])
3) Deep Learning for Automatic Pneumonia Detection (Tatiana Gabruseva, Dmytro Poplavskiy , Alexandr Kalinin)
4) CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning