RSNA PNEUMONIA DETECTION CHALLENGE – FINAL PROJECT REPORT

In Support of Capstone Project of Post Graduate Programme in Artificial Intelligence and Machine Learning

Team:

Ronald Chiwariro | Sujatha Mohanram

Vikas Kumar Dewangan

Julius Wosowei | Mahesh Pachhapurkar

Guide: Girijesh Prasad

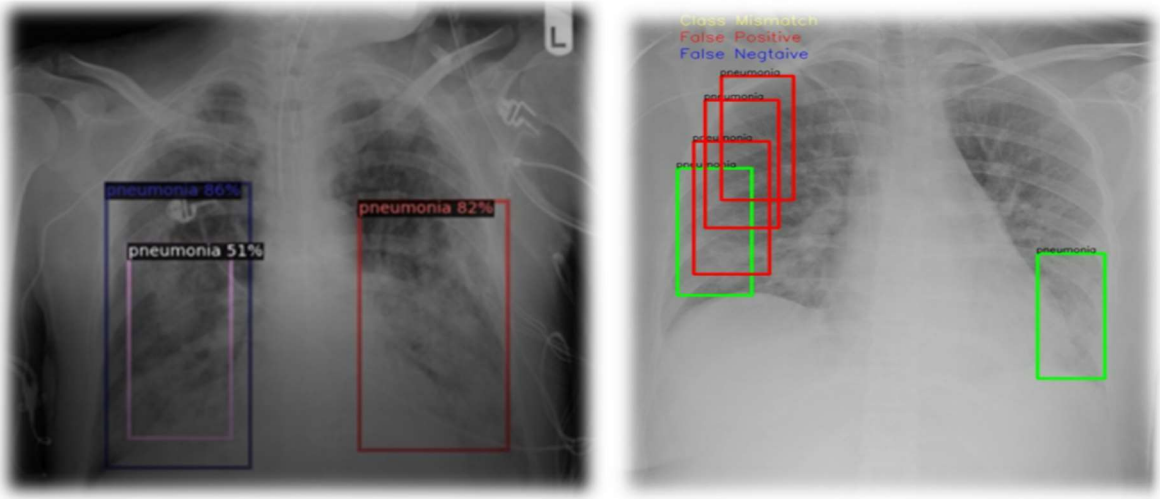


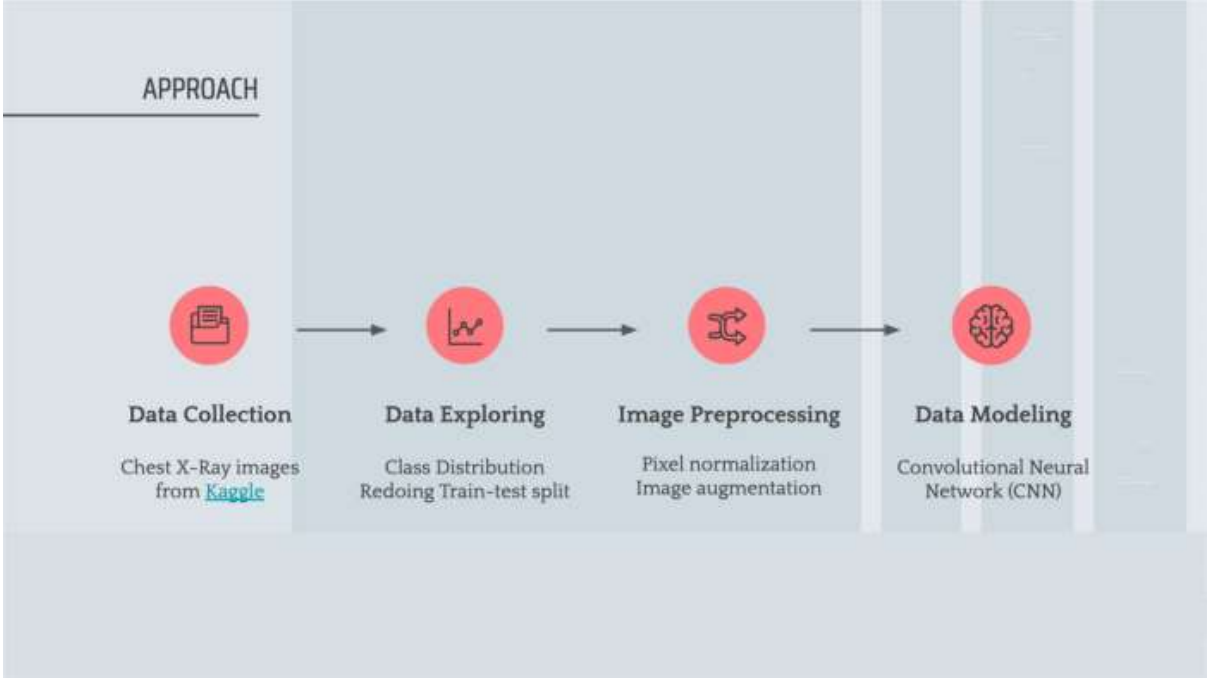
Table of Contents

1. Summary of the problem statement, data and findings

Pneumonia is a respiratory infection caused by bacteria or viruses; it affects many individuals, especially in developing and underdeveloped nations, where high levels of pollution, unhygienic living conditions, and overcrowding are relatively common, together with inadequate medical infrastructure. The problem to be addressed is to diagnose Pneumonia using Chest X-ray imaging and location of the Lung inflammation, which is the most frequently used method. However, the examination of chest X-rays is a challenging task and is prone to subjective variability. In this study, we developed a computer-aided diagnosis system for automatic pneumonia detection and detection of inflammation using chest X-ray images. Deep transfer learning was employed using VGG19 [2], VGG16 [2], Resnet50 [3], InceptionNet [4] and YOLO [5]. The scores of four standard evaluation metrics; precision, recall, f1-score, the area under the curve and bounding boxes confidence levels used to assess the Model performance. In the classification of pneumonia and non-pneumonia cases, VGG16 outperformed the other classifier-based Models and YOLO was applied for object detection to locate the position of inflammation. In general, pneumonia cases tend to be high in males of 40 and above age group and all cases were classified correctly with the bounding boxes showing high confidence levels.

2. Overview of the final process

Below is the high-level approach on how the deep learning model was created. First, the data was collected from Kaggle, then, exploratory data analysis was performed on it. These were followed by pre-processing of the input images through rescaling the pixels and image augmentations. Finally, different deep learning models have been experimented on the data.



2.1 Dataset

The proposed approach used the RSNA Pneumonia Detection Challenge [1] dataset from Kaggle. It is a dataset of chest X-Rays with annotations, which shows which part of the lung has symptoms of pneumonia. The size of the entire dataset is approximately 4 GB, with 3 subfolders, namely *stage_2_detailed_class_info, stage_2_train_labels, stage_2_train_images* and *stage_2_test_images*. The proposed method achieved a validation accuracy rate of 98.81%.

2.2 Introduction

Pneumonia is an acute pulmonary infection that can be caused by bacteria, viruses, or fungi and infects the lungs, causing inflammation of the air sacs and pleural effusion, a condition in which the lung is filled with fluid. Early diagnosis and management can play a pivotal role in preventing the disease from becoming fatal. Radiological examination of the lungs using computed tomography (CT), magnetic resonance imaging (MRI), or radiography (X-rays) is frequently used for diagnosis. X-ray imaging constitutes a non-invasive and relatively inexpensive examination of the lungs.

The white spots in the pneumonic X-ray called infiltrates, distinguish a pneumonic from a health condition. However, chest X-ray examinations for pneumonia detection are prone to subjective variability. Thus, an automated system for the detection of pneumonia is required. In this study, we developed a computer-aided diagnosis system that uses an ensemble of deep transfer learning models for the accurate classification of chest X-ray images. Deep learning is an important artificial intelligence tool, which plays a crucial role in solving many complex Computer Vision problems.

Deep learning models, specifically convolutional neural networks (CNNs), are used extensively for various image classification problems. However, such models perform optimally only when they are provided with a large amount of data. For biomedical image classification problems, such a vast amount of labelled data is difficult to acquire because it requires that expert doctors classify each image, which is an expensive and time-consuming task. Transfer learning is a work-around to surmount this obstacle. In this technique, to solve a problem that involves a small dataset, a model trained on a large dataset is re-used and the network weights determined in this model are applied. CNN models trained on a large dataset such as ImageNet [7], which consists of more than 14 million images, are frequently used for biomedical image classification tasks.

Exploratory Data Analysis (EDA) was conducted to check on various statistical properties of the data. Findings of this stage will be explained later in detail. Upon completion of the EDA the images were pre-processed by resizing, scaling, augmentation and annotation and finally converted into jpg format. The pre-processed data was subjected to a variety of classification and object detection models out of which VGG19, VGG16, Resnet 50, InceptionNet and YOLO were chosen for further tuning and final solution build-up.

3.3 Step-by-step walk through the solution

3.1 EDA and Pre-processing

First, we give the details of the data/files given for this problem.

1. stage_2_train_labels.csv - the training set. It contains patientIds, the location of bounding boxes and target information.
   - patientId_ - Each patientId corresponds to a unique image.
   - x_ - the upper-left x coordinate of the bounding box.
   - y_ - the upper-left y coordinate of the bounding box.
   - width_ - the width of the bounding box.
   - height_ - the height of the bounding box.
   - Target_ - the binary Target, indicating whether this sample has evidence of pneumonia.
2. stage_2_detailed_class_info.csv - provides detailed information about the type of positive or negative class for each patientId. Class has three values depending on what is the current state of the patient's lung: 'No Lung Opacity / Not Normal', 'Normal' and 'Lung Opacity'.
3. DICOM files (*.dcm): These are special format, in which medical images of the patients are given. They contain a combination of header metadata as well as underlying raw image arrays for pixel data.

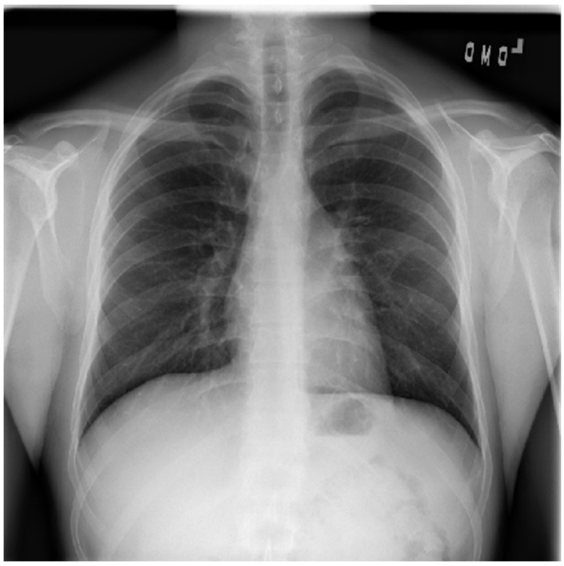
In this project, we must predict whether pneumonia exists in a given image. This is done by predicting bounding boxes around areas of the lung. Samples without bounding boxes are negative and contain no definitive evidence of pneumonia. Samples with bounding boxes indicate evidence of pneumonia. When making predictions, the model should predict as many bounding boxes as necessary, in the format: confidence x-min, y-min, width and height. There will be only ONE predicted row per image. This row may include multiple bounding boxes.
A properly formatted row may look like any of the following.

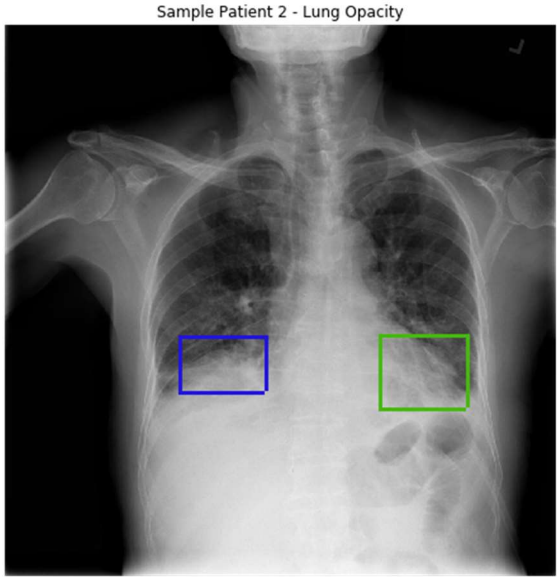
- For patientIds with no predicted pneumonia / bounding boxes: 0004cfab-14fd-4e49-80ba-63a80b6bddd6,
- For patientIds with a single predicted bounding box: 0004cfab-14fd-4e49-80ba-63a80b6bddd6,0.5 0 0 100 100
- For patientIds with multiple predicted bounding boxes: 0004cfab-14fd-4e49-80ba-63a80b6bddd6,0.5 0 0 100 100 0.5 0 0 100 100, etc.

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 labelled “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 often times this finding may mimic the appearance of true pneumonia.
It's important to note that the various shades of gray in the chest X-Ray refer to the following:

- *Black* = Air
- *White* = Bone
- *Grey* = Tissue or Fluid

In a normal image, we see the lungs as black, but they have different projections on them - mainly the rib cage bones, main airways, blood vessels and the heart. In the case of pneumonia, a haziness (also referred to as *consolidation*) is present in the chest x-ray image.
Images with no lung opacity and no pneumonia are images where the patient can have rounded hazy boundaries or masses (probably because of lung nodules or masses which can be because of cancer). There are other exceptional cases as well where there can be no lung opacity but no pneumonia either. Some of these cases include *pneumonectomy* (lung removed by surgery), *enlarged heart*, *pleural effusion*, etc.
Reference: https://www.kaggle.com/zahaviguy/what-are-lung-opacities





The training images (stage_2_train_images.zip) and annotations (stage_2_train_labels.csv) were downloaded from kaggle.

	patientId	x	y	width	height	Target
0	0004cfab-14fd-4e49-80ba-63a80b6bddd6	NaN	NaN	NaN	NaN	0
1	00313ee0-9eaa-42f4-b0ab-c148ed3241cd	NaN	NaN	NaN	NaN	0
2	00322d4d-1c29-4943-afc9-b6754be640eb	NaN	NaN	NaN	NaN	0
3	003d8fa0-6bf1-40ed-b54c-ac657f8495c5	NaN	NaN	NaN	NaN	0
4	00436515-870c-4b36-a041-de91049b9ab4	264.0	152.0	213.0	379.0	1

First row in the above image corresponds to patient with ID '0004cfab-14fd-4e49–80ba-63a80b6bddd6'. 'Target' column is 0 for this row. It means this patient doesn't have pneumonia. On other hand, patient in the last row has pneumonia because of the area on the corresponding chest radiograph (xmin = 264, ymin = 152, width = 213, height = 379) with opacity.

Distribution of Data based on the Target

The data set has 30227 entries out of which 26684 are unique (i.e., 68.4% of which correspon d to patient ‘Not having Pneumonia’ and 31.6% entries corresponds to ‘Positive case’ of Pneu m-onia.)

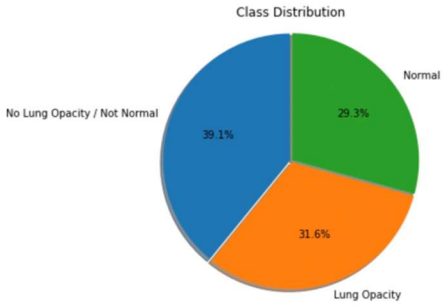
Target Distribution

Observation:

The positive cases are under-represented class.

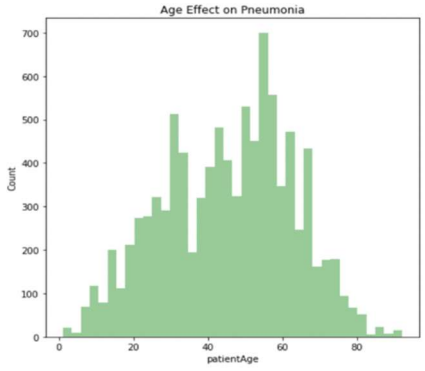
We will use the methods to ensure this imbalance is addressed while training the model.

Class Distribution



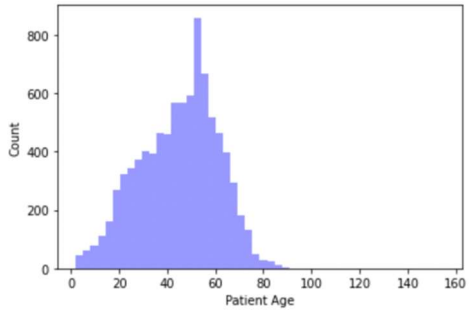
The data set has 31.6% patients with lung Opacity, 39.1% with No lung Opacity but not normal and 29.3% with normal condition.

Age Effect on Pneumonia



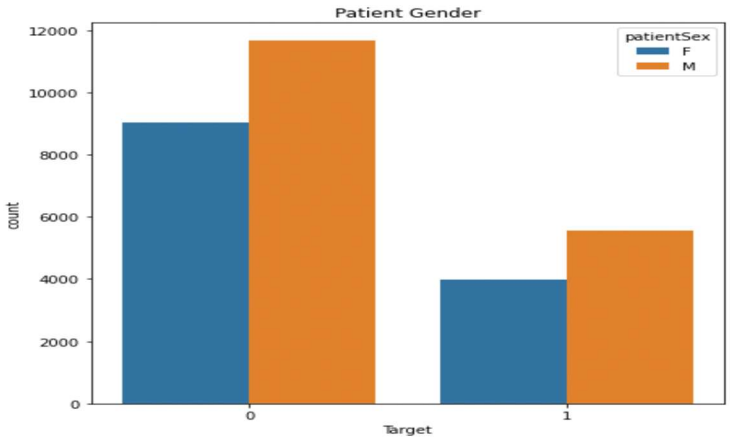
The graph shows the relation between pneumonia and patient's age.

There are more numbers of pneumonia patients between ages 40 to 60.



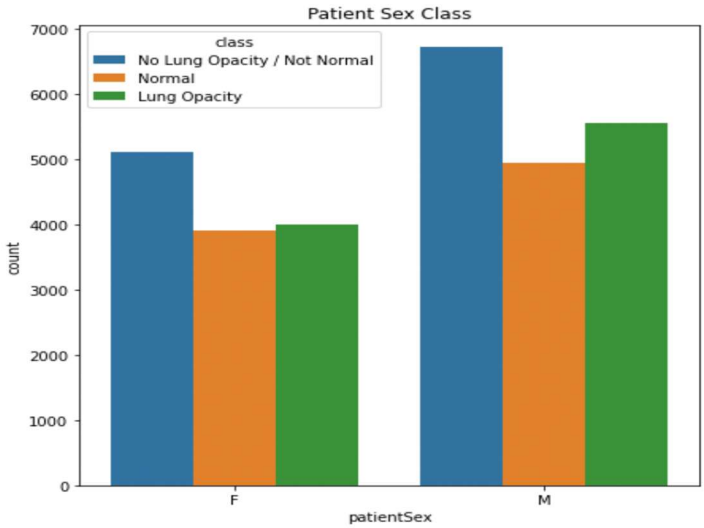
Ages from 40 to 60 show more numbers with normal patients

Distribution based on Patient Gender



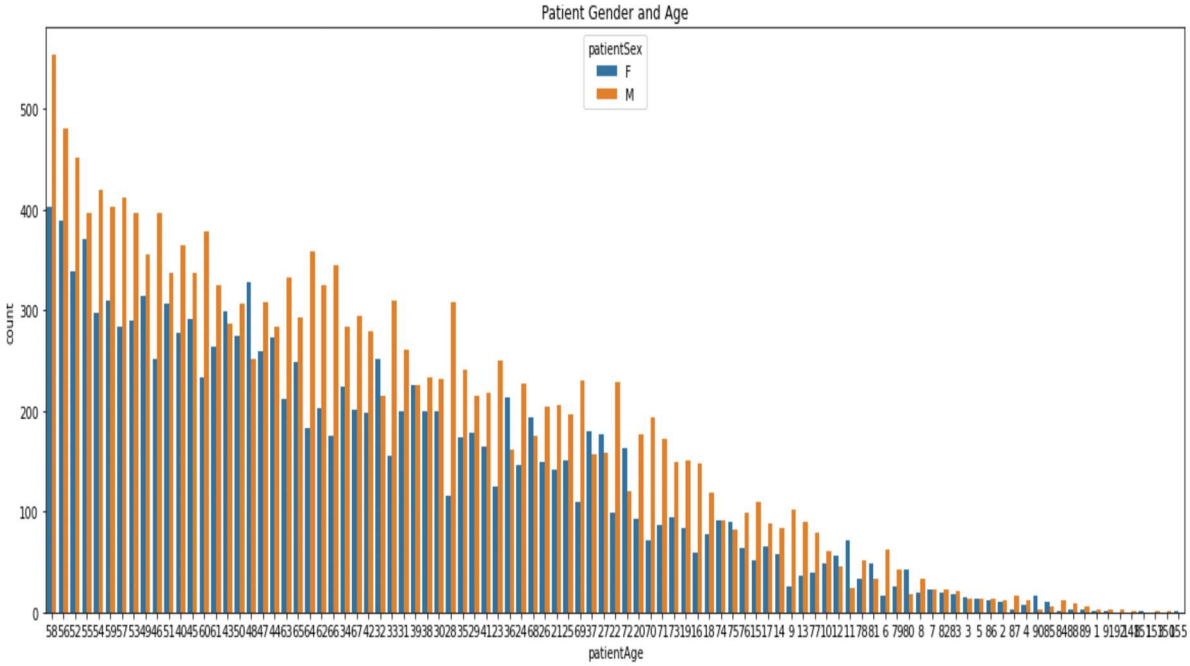
Patients without pneumonia in both genders are higher. The number of males is more in both cases.

Different classes as per Patient Gender



Male patients have more records for different classes

Impact on Age and Gender

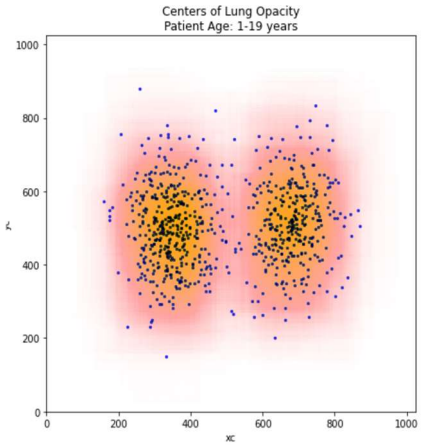


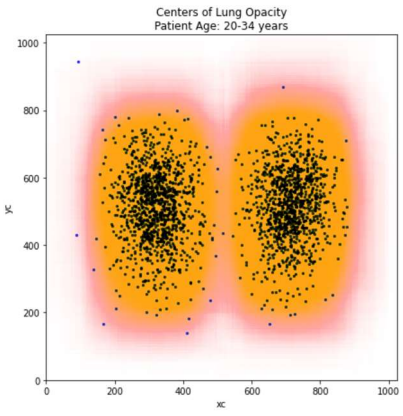
Observation - Male patients dominate in most age groups

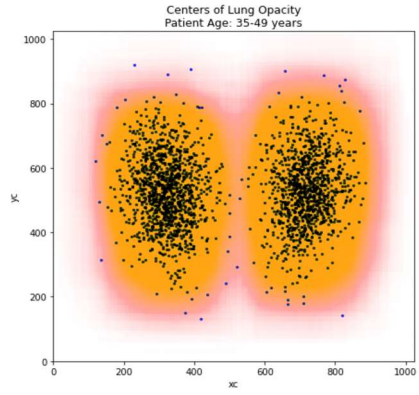
Pneumonia presence based on age

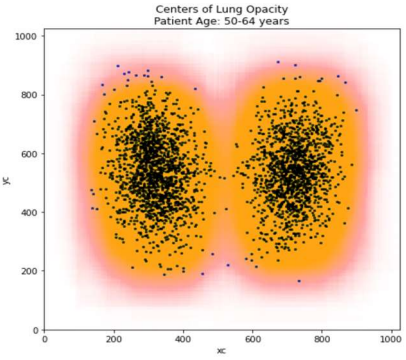
Age and Pneumonia relation

The following images will show how the Pneumonia is spread in the lung region based on the age group.

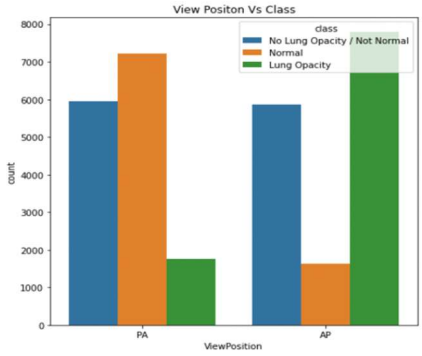




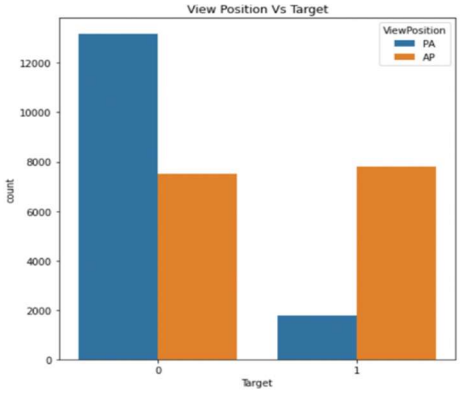




View position on different features



Patients from Anterior Posterior (AP) Projection view position have more number of records than the Posterior-Anterior (PA) projection.



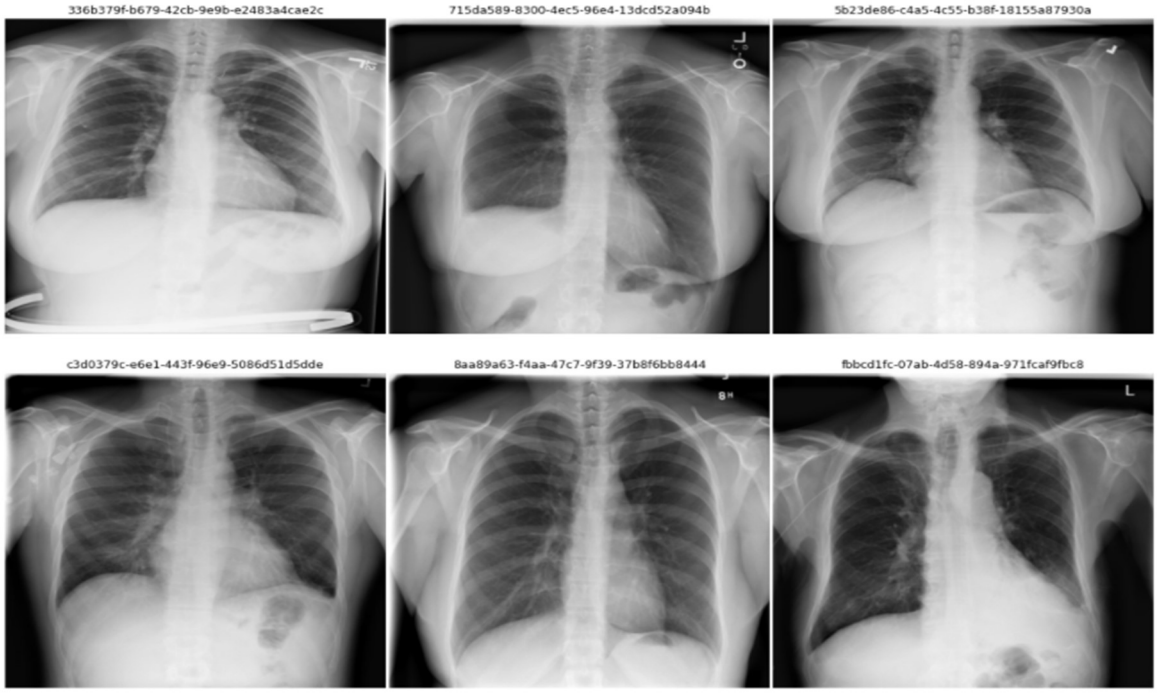
Patients with AP view position has more number of records than PA.

X-Ray view Positions

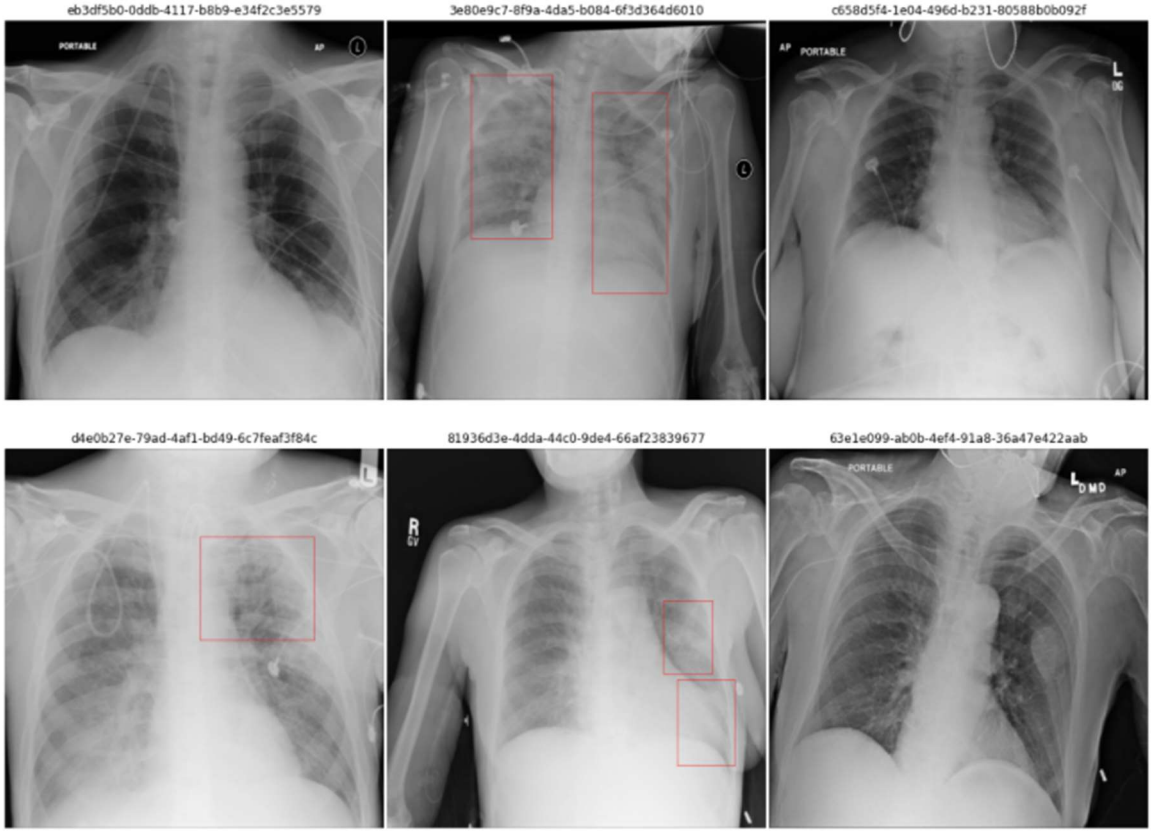
Posterior/Anterior (PA): Here the chest radiograph is acquired by passing the X-Ray beam from the patient's posterior (back) part of the chest to the anterior (front) part. These are of higher quality and assess the heart size more accurately

Anterior/Posterior (AP): At times it is not possible for radiographers to acquire a PA chest X-ray. This is usually because the patient is too unwell to stand. In these images the size of Heart is exaggerated.

Sample of PA X-rays

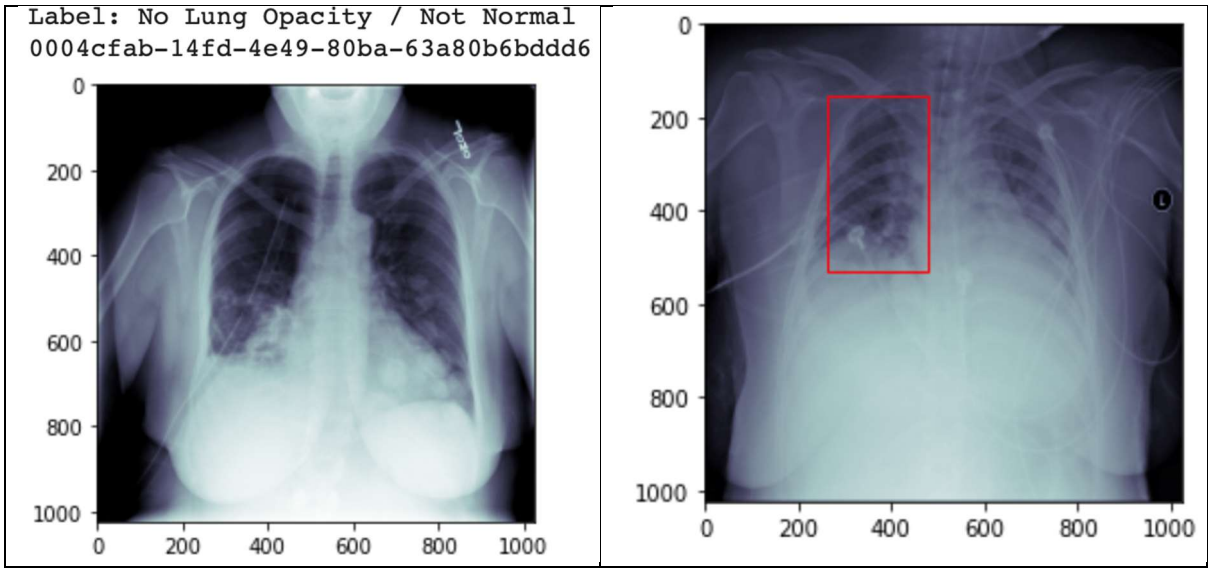


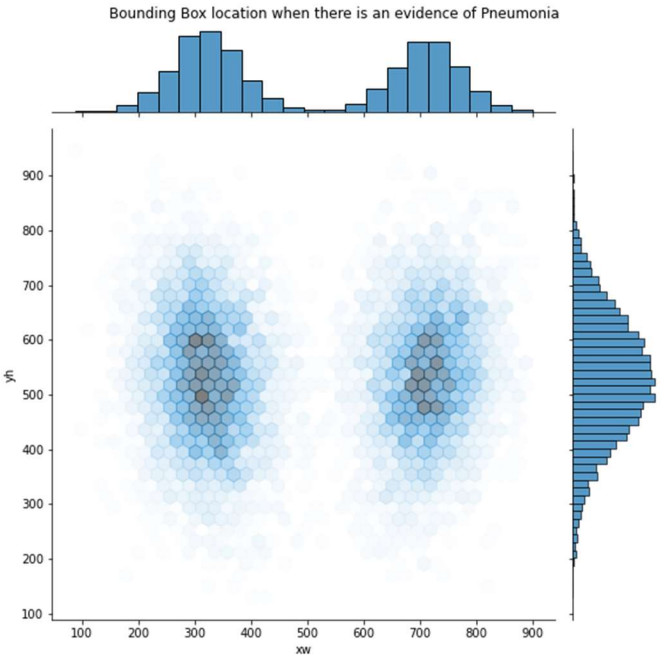
Sample of AP X-rays



The above image grid shows X-rays taken in PA and AP positions for different patients with and without signs of pneumonia.

Images of Normal cases and cases where Pneumonia was detected





The diagram shows the distribution of centre of bounding boxes for patients having pneumonia.

The insights gained from the EDA process provided a detailed understanding of the dataset in terms of the features, targets, and classes. This necessitated the pre-processing stage for the images. During this stage, the images were subjected to augmentation and annotation. This further enables the data to be passed on to the Models for training after being split into train and test sets as well as passing through the Image Data Generator.

4. Model evaluation

Model Evaluation is an integral part of the model development process. It helps to find the best model that represents our data and how well the chosen model will work in the future. The various models were tried and evaluated before the best model was finalized. The methods and benchmarks for the evaluation has been detailed below.

4.1 Classification Parameters

Categorical cross entropy is a loss function that is used in this multi-class classification task. It is designed to quantify the difference between two probability distributions. Mathematically,

$$\text{Loss} = -I = 1\sum tyi \cdot logy\hat{}i$$

Were;

- \hat(y)_iy^i is the i^{th} scalar value in the model output,
- y_iyi is the corresponding target value, and it is summed over the output size is the number of scalar values in the model output.

A Classification report is used to measure the quality of predictions from a classification algorithm. It is used to evaluate the model. The report shows the main classification metrics precision, recall and f1-score on a per-class basis.

Accuracy is defined as the percentage of correct predictions for the test data. It can be calculated easily by dividing the number of correct predictions by the number of total predictions.

$$Accuracy = Correct\ Predictions\ /\ Total\ Predictions$$

Precision is defined as the fraction of relevant examples (true positives) among all the examples which were predicted to belong in a certain class.

$$Precision = True\ Positives\ /\ (True\ Positives + False\ Positives)$$

Recall is defined as the fraction of examples which were predicted to belong to a class with respect to all of the examples that truly belong in the class.

$$Recall = True\ Positives\ /\ (True\ Positives + False\ Negatives)$$

The F1 score is a weighted harmonic mean of precision and recall such that the best score is 1.0 and the worst is 0.0.

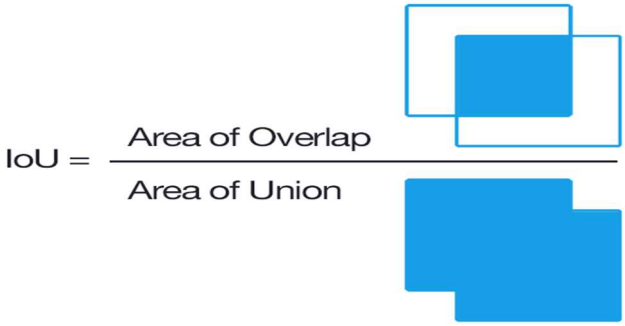
$$F1\ Score = 2*(Recall * Precision)\ /\ (Recall + Precision)$$

4.2 Object Detection Parameter

Mean Average Precision (mAP) is used to determine the accuracy of a set of object detections from a model when compared to ground-truth object annotations of a dataset.

Intersection over Union (IoU) is used when calculating mAP. It is a number from 0 to 1 that specifies the amount of overlap between the predicted and ground truth bounding box. IoU of

0 means that there is no overlap between the boxes and IoU of 1 means that the union of the boxes is the same as their overlap indicating that they are completely overlapping.

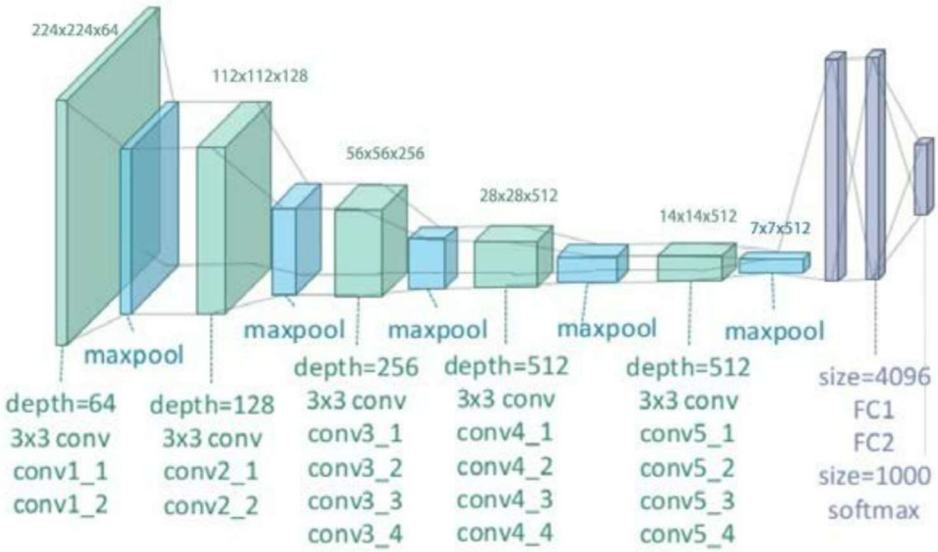


4.3 Models Used

Models used on the dataset are VGG19, VGG16, ResNet50, InceptionNet v3 and YOLO v3. The models were trained for 5 epochs on the same 'training set'. The results from the Models are as shown below:

1. VGG19

VGG19 is a variant of VGG model which in short consists of 19 layers (16 convolution layers, 3 Fully connected layer, 5 MaxPool layers and 1 SoftMax layer). There are other variants of VGG like VGG11, VGG16 and others. VGG19 has 19.6 billion FLOPs. The input to VGG based convNet is a 224*224 RGB image. Pre-processing layer takes the RGB image with pixel values in the range of 0–255 and subtracts the mean image values which is calculated over the entire ImageNet training set. VGG19 Variation has 19 weight layers consisting of 16 convolutional layers with 3 fully connected layers and 5 pooling layers.

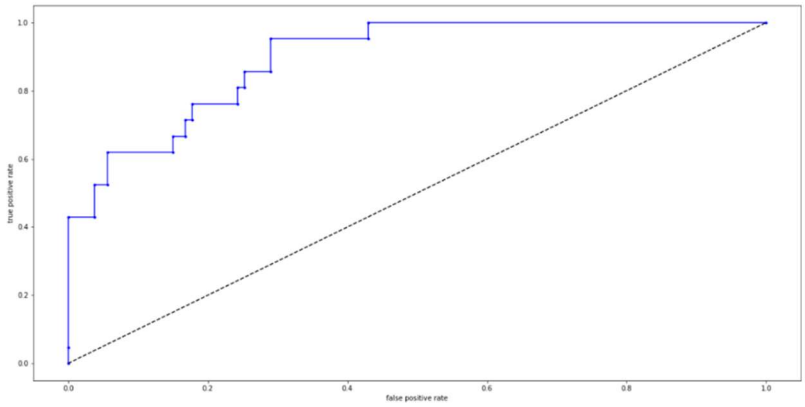


VGG-19 Architecture

Classification Report

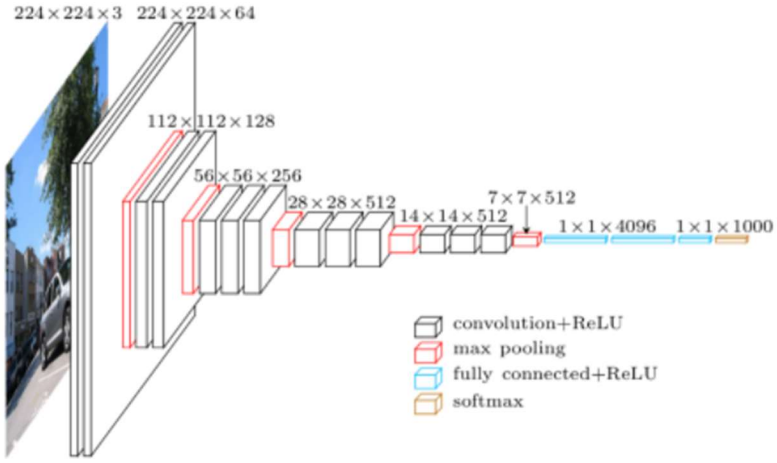
	precision	recall	f1-score	support
0.0	0.95	0.81	0.87	107
1.0	0.44	0.76	0.56	21
accuracy			0.80	128
macro avg	0.70	0.79	0.72	128
weighted avg	0.86	0.80	0.82	128

Area Under Curve



2. VGG16

VGG16 is a CNN architecture which was used to win ILSVR (Imagenet) competition in 2014. It is one of the excellent vision model architectures till date. Most unique thing about VGG16 is that instead of having many hyper-parameters they focused on having convolution layers of 3x3 filter with a stride 1 and always used same padding and maxpool layer of 2x2 filter of stride 2. It follows this arrangement of convolution and max pool layers consistently throughout the whole architecture. In the end it has 2 fully connected layers followed by a softmax for output. The 16 in VGG16 refers to it having 16 layers that have weights. This network is a pretty large network, and it has about approximately 138 million parameters.

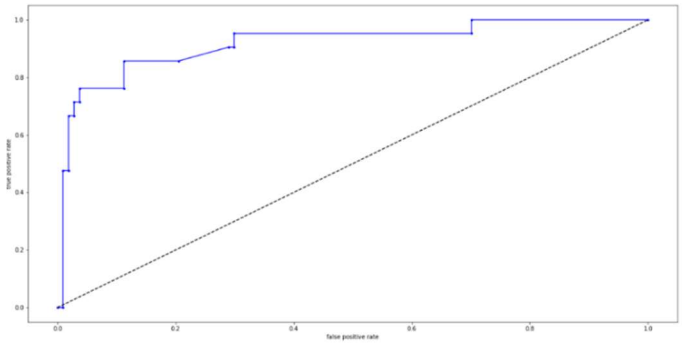


VGG-16 Architecture

Classification Report

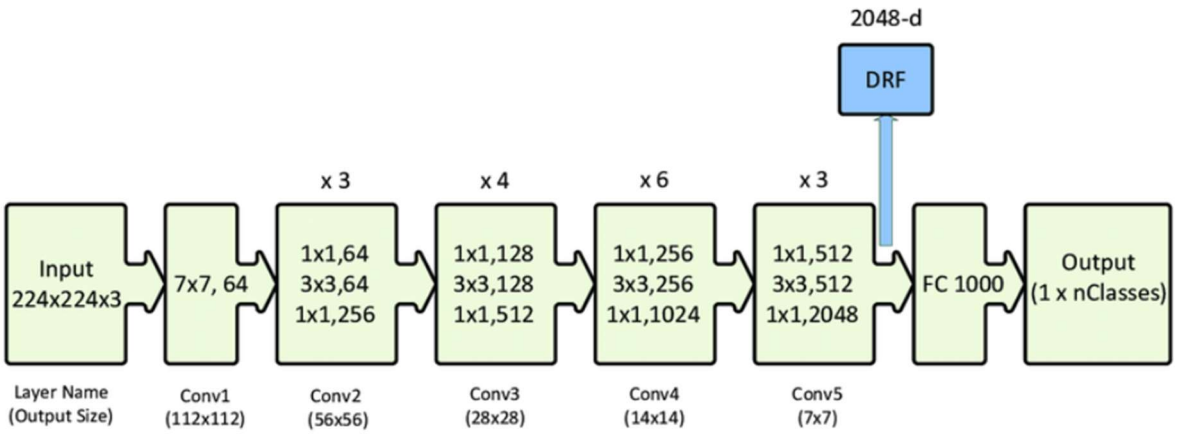
	precision	recall	f1-score	support
0.0	0.97	0.88	0.92	107
1.0	0.58	0.86	0.69	21
accuracy			0.88	128
macro avg	0.77	0.87	0.81	128
weighted avg	0.91	0.88	0.88	128

Area Under Curve



3. ResNet50

ResNet-50 is a convolutional neural network that is 50 layers deep. Note that there is only one 3x3 convolution rather than two. 1x1 convolutions are used to map in lower dimension and then perform 3x3 convolution and then remap them to higher dimensions. This way the training time will be less. The other part to note is in ResNet 50 when there is dimension change then the authors used 1x1 convolutions at x to make the dimension same.

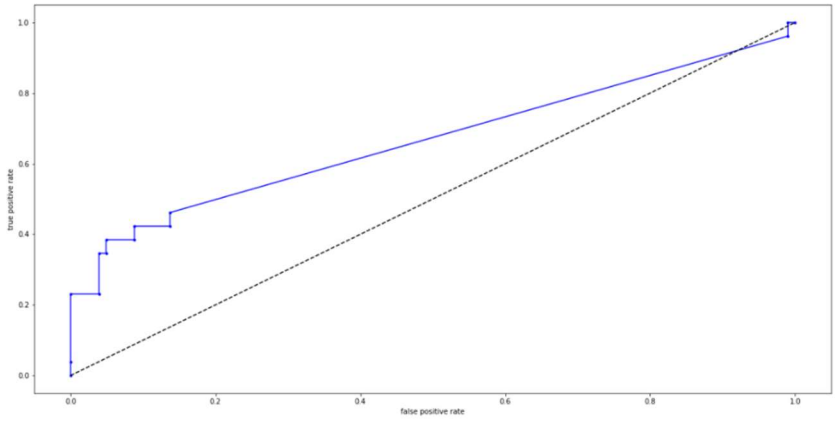


ResNet50 Architecture

Classification Report

	precision	recall	f1-score	support
0.0	0.86	0.86	0.86	102
1.0	0.46	0.46	0.46	26
accuracy			0.78	128
macro avg	0.66	0.66	0.66	128
weighted avg	0.78	0.78	0.78	128

Area Under Curve

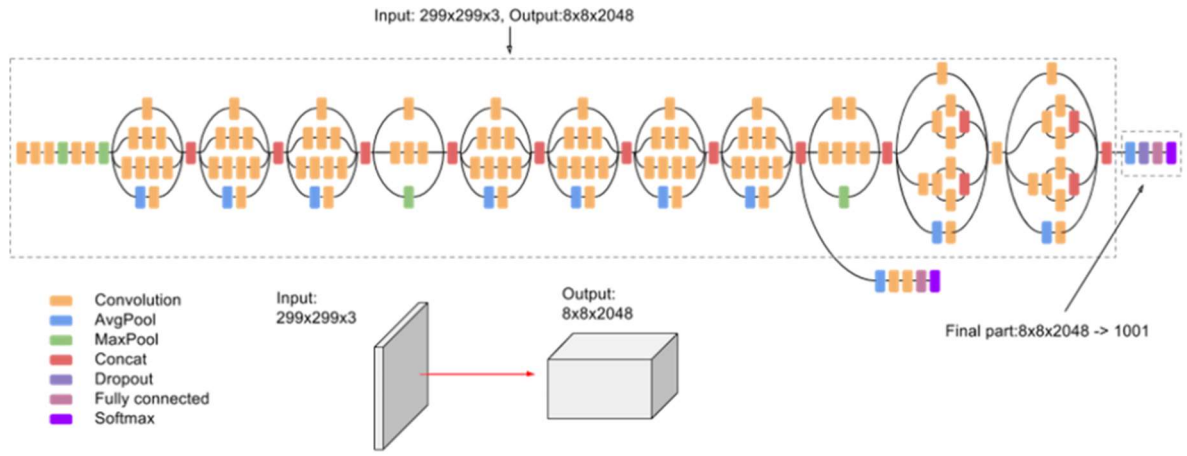


4. Inception v3

Inception-v3 is a convolutional neural network architecture from the Inception family that makes several improvements including using Label Smoothing, Factorized 7 x 7 convolutions, and the use of an auxiliary classifier to propagate label information lower down the network (along with the use of batch normalization for layers in the side head).

The architecture of an Inception v3 network is progressively built, step-by-step, as explained below:

1. Factorized Convolutions: this helps to reduce the computational efficiency as it reduces the number of parameters involved in a network. It also keeps a check on the network efficiency.
2. Smaller convolutions: replacing bigger convolutions with smaller convolutions leads to faster training. Say a 5 × 5 filter has 25 parameters; two 3 × 3 filters replacing a 5 × 5 convolution has only 18 (3*3 + 3*3) parameters instead.

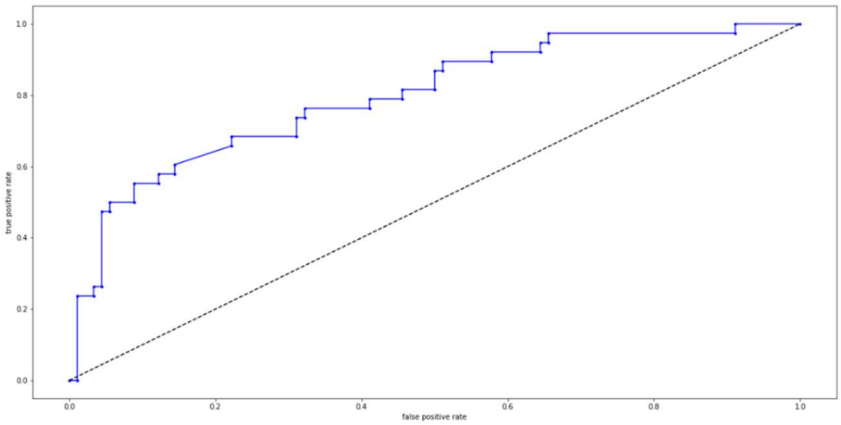


Inception v3 Architecture

Classification Report

	precision	recall	f1-score	support
0.0	0.84	0.78	0.81	90
1.0	0.56	0.66	0.60	38
accuracy			0.74	128
macro avg	0.70	0.72	0.71	128
weighted avg	0.76	0.74	0.75	128

Area Under Curve



5. YOLO v3

In YOLO, single neural network predicts bounding boxes and class probabilities directly from full images in one evaluation. Since the whole detection pipeline is a single network, it can be optimized end-to-end directly on detection performance.

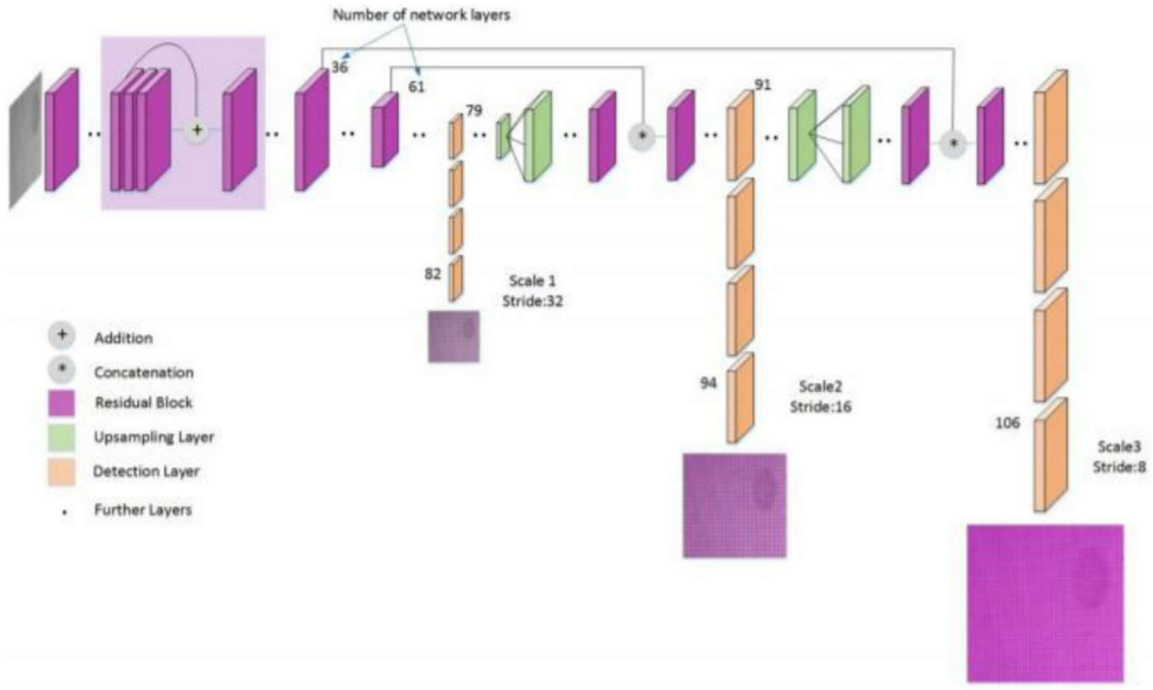
The inputs are a batch of images of shape (m, 416, 416, 3). YOLO v3 passes this image to a convolutional neural network (CNN). The last two dimensions of the above output are flattened to get an output volume of (19, 19, 425):

Here, each cell of a 19 x 19 grid returns 425 numbers.

- 425 = 5 * 85, where 5 is the number of anchor boxes per grid.
- 85 = 5 + 80, where 5 is (pc, bx, by, bh, bw) and 80 is the number of classes we want to detect.

The output is a list of bounding boxes along with the recognized classes. Each bounding box is represented by 6 numbers (pc, bx, by, bh, bw, c). If we expand c into an 80-dimensional vector, each bounding box is represented by 85 numbers.

Finally, we do the IoU (Intersection over Union) and Non-Max Suppression to avoid selecting overlapping boxes.



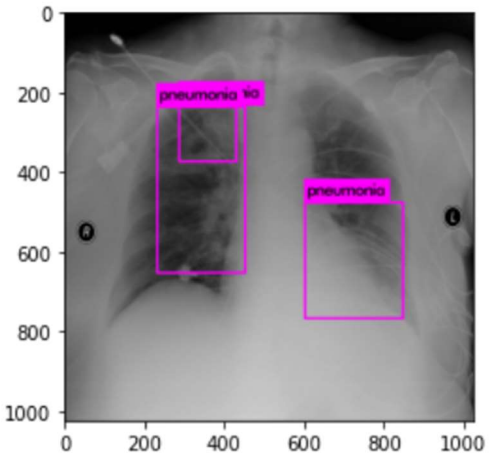
YOLO v3 Architecture

Prediction 1

```
#Plot the prediction with the position of Lung Inflammation

plt.imshow(cv2.imread("predictions.jpg"))
```

```
<matplotlib.image.AxesImage at 0x7fba6798d250>
```



- Detections of inflammations on test images.
- Default Threshold of 0.005% was used during testing.

Prediction 2

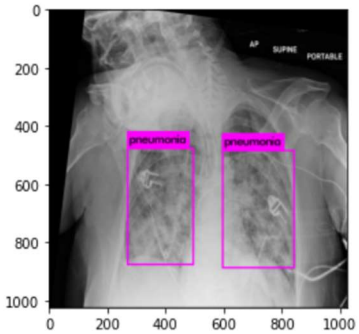
```
Loading weights from ./backup/rsna_yolov3_15300.weights...Done!
./images/395738ff-53d2-42e6-aa6d-a01e4a07f01f.jpg: Predicted in 22.544326 seconds.
pneumonia: 99%
pneumonia: 96%
```

- Boxes show confidence of:
  - 99%
  - 96%

```
#Plot the prediction with the position of Lung Inflammation

plt.imshow(cv2.imread("predictions.jpg"))
```

```
<matplotlib.image.AxesImage at 0x7fba67a881d0>
```



- Detections of inflammations on test images were correctly done with a high confidence level.
- Threshold of 25% was applied as a filter during testing.

4.4 Summary of Model Results

Results:

- The four Models namely, **VGG19, VGG16, ResNet50**, **InceptionNet v3** and **YOLO v3** were trained on the full dataset and their performances based on Validation Accuracy and Area Under Curve (AUC) were as follows:

    1. **VGG19:** Validation Accuracy - **80%** and AUC-ROC – **82.5%**.
    2. **VGG16:** Validation Accuracy - **88%** and AUC-ROC – **91.8%**.
    3. **ResNet50:** Validation Accuracy - **73%** and AUC-ROC – **80.9%**.
    4. **InceptionNet v3:** Validation Accuracy - **79%** and AUC-ROC – **87%**.
    5. **YOLO v3:** Inflammation-detection bounding boxes with high confidence.

- Therefore, it can be observed that the **VGG16 Model** has the highest Validation Accuracy and Area Under Curve for **Classification**.
- **YOLO v3** is the applied to locate position of the inflammation.

Limitations:

- The training time was high even though Google Collab Pro (TPU and GPU) were used, and we were forced to reduce the number of training epochs to avoid the continuous usage timeouts before finishing execution of all Models.
- Accuracy can be improved by further hyperparameter tuning and allowing the Models to train for more epochs, computing resources permitting.
- However, sampling of the dataset may also help mitigate this challenge although it may also have an effect on the Model Accuracy.

5. Comparison to benchmark

Benchmarking is used to measure performance using a specific indicator resulting in a metric that is then compared to others. In other words, it is the comparison of a given model's inputs and outputs to estimates from alternative internal or external data or models.

At the outset, we began with Customized CNN model, but it was yielding very low Accuracy and Recall scores around 20%; hence it was discarded for pretrained models which were then used. To enhance the score and become more effective in detecting pneumonia, we used augmentation techniques and experimented on multiple models such as VGG16 and 19, Resnet 50, Inception v3 and YOLO to determine Pneumonia with higher accuracy and for detection of inflammation.

The benchmark was improved by hyperparameter tuning on the number of epochs, learning rate and batch sizes. Due to limitations of resources training could not be done on more than 20 epochs due to timeouts in Google collab. Further increasing the epochs showed great improvement in the classification models.

6. Visualizations

During the EDA process we discovered that there are total of 32227 patient records out of which 26684 unique patients. Overall, the distribution of data is imbalanced with Target class being only 31.6% of the whole dataset. This tends to result in bias. We have addressed such

data issues using augmentation techniques and sampling methods to equally represent data classes.

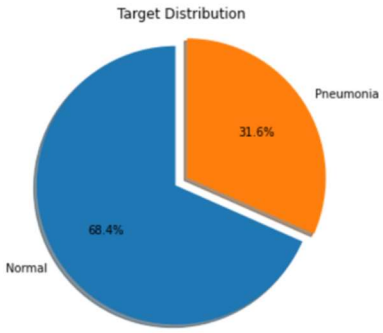
Findings from the Dicom images dataset are:

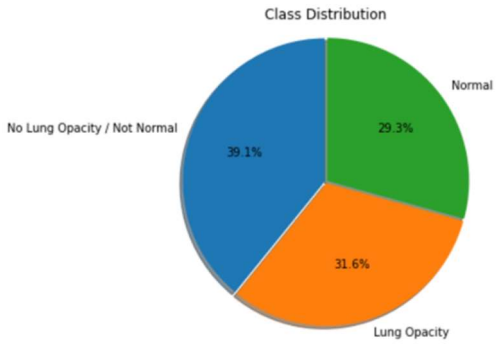
Statistics for patients with multiple bounding boxes are as follows:

a. 3266 patients have 2 bounding boxes defined.
b. 119 patients have 3 bounding boxes defined.
c. 13 patients have 4 bounding boxes defined.

We will run a binary classification to predict patients with pneumonia. Target – 1 indicates patients with pneumonia and target – 0 indicates normal patients or patients having lung opacity but not pneumonia. Further statistics are as follows:

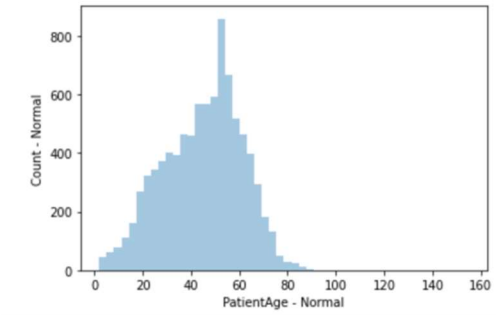
a. 8,851 (29.3%) patients are healthy/ Normal.
b. 9,555 (31.6%) patients have Lung Opacity.
c. 11,821 (39.1%) patients have No Lung Opacity but may have other lung abnormality.

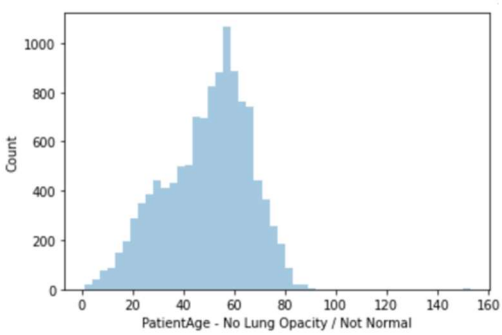


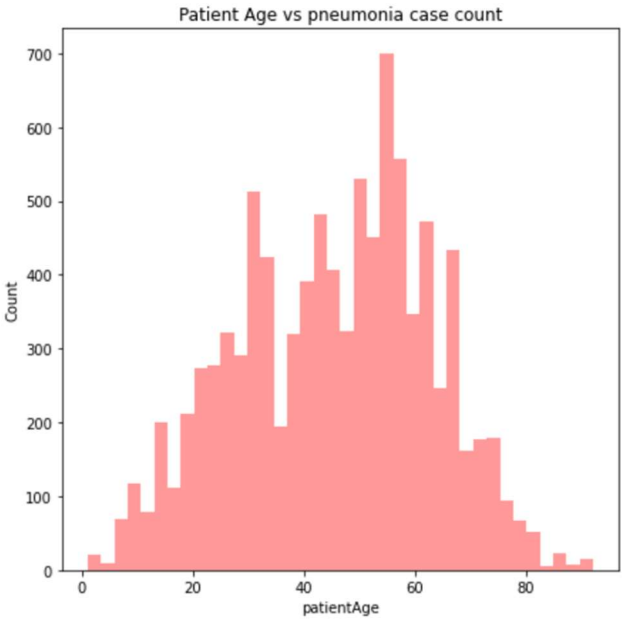


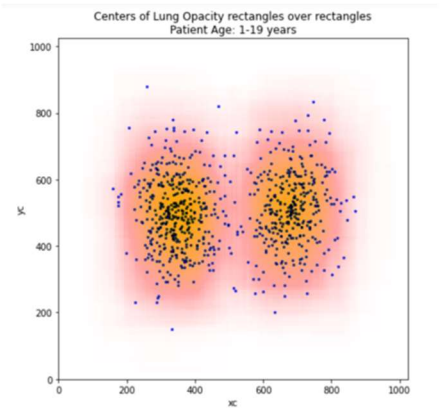
	class	Target	Patient Count
0	Lung Opacity	1	9555
1	No Lung Opacity / Not Normal	0	11821
2	Normal	0	8851

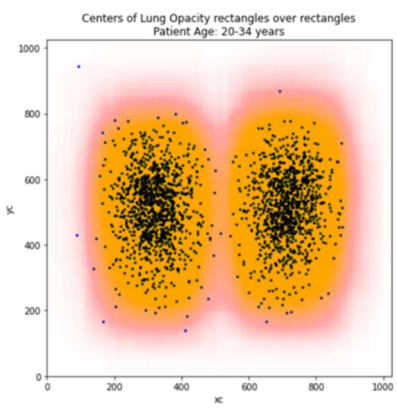
Patient distribution shows most pneumonia patients or patients are found between age 40-60 dominated by male patients. Proportion of pneumonia and normal patients is equal for female patients. However, for males, proportion of pneumonia patients is more than for Normal patients.

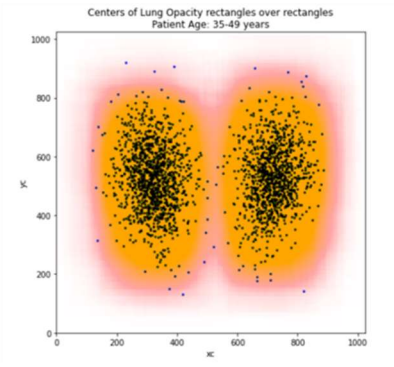

Count - Normal
PatientAge - Normal

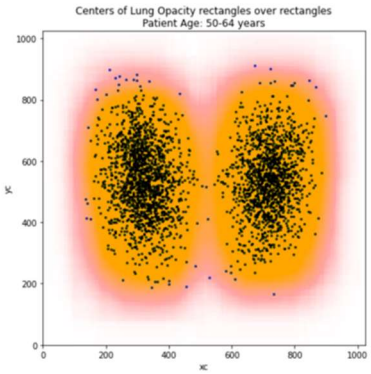

Count
PatientAge - No Lung Opacity / Not Normal

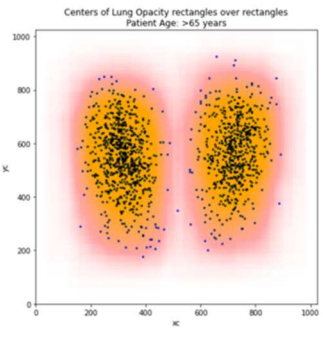

Patient Age vs pneumonia case count
Count
patientAge

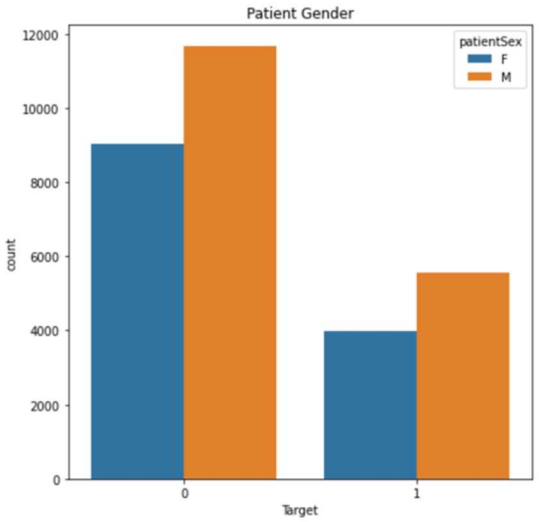

Centers of Lung Opacity rectangles over rectangles
Patient Age: 1-19 years
yc
xc

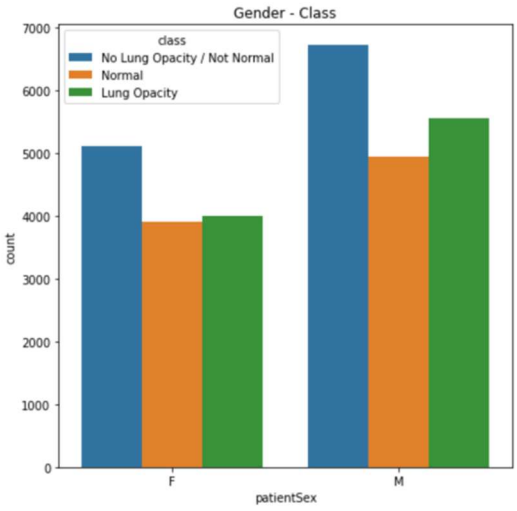

Centers of Lung Opacity rectangles over rectangles
Patient Age: 20-34 years
yc
xc


Centers of Lung Opacity rectangles over rectangles
Patient Age: 35-49 years
yc
xc

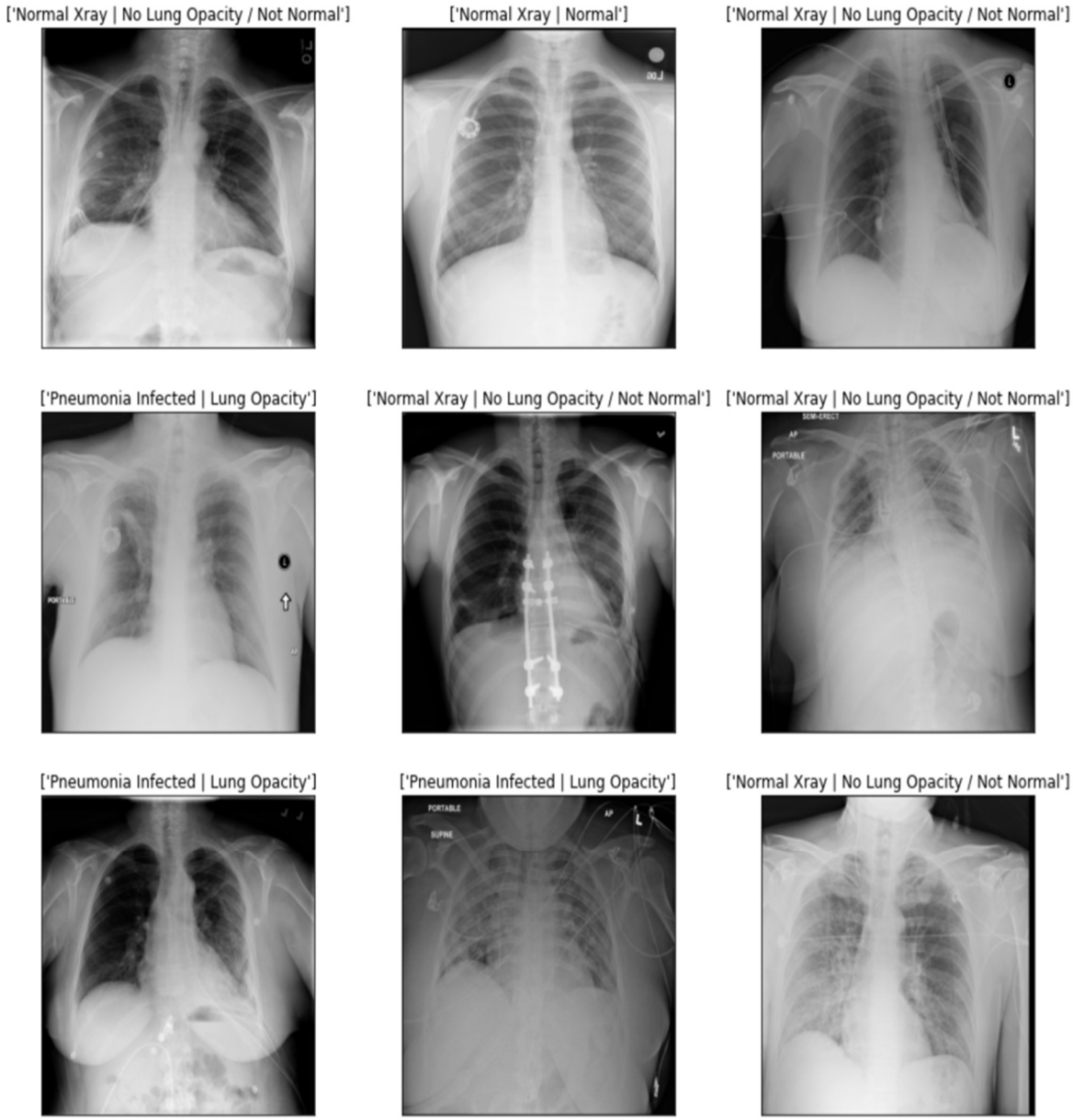

Centers of Lung Opacity rectangles over rectangles
Patient Age: 50-64 years
yc
xc


Centers of Lung Opacity rectangles over rectangles
Patient Age: >65 years
yc
xc

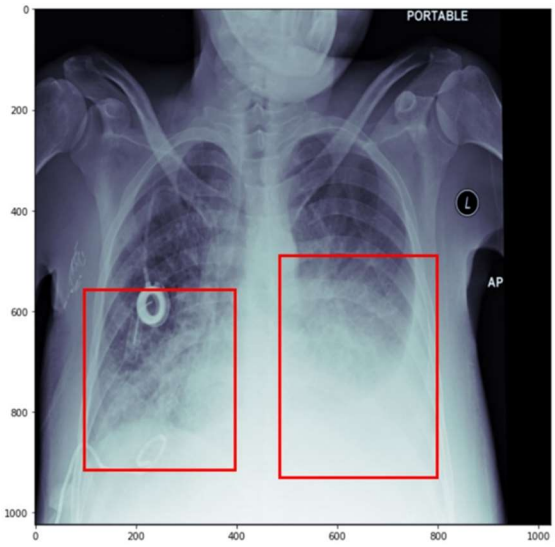


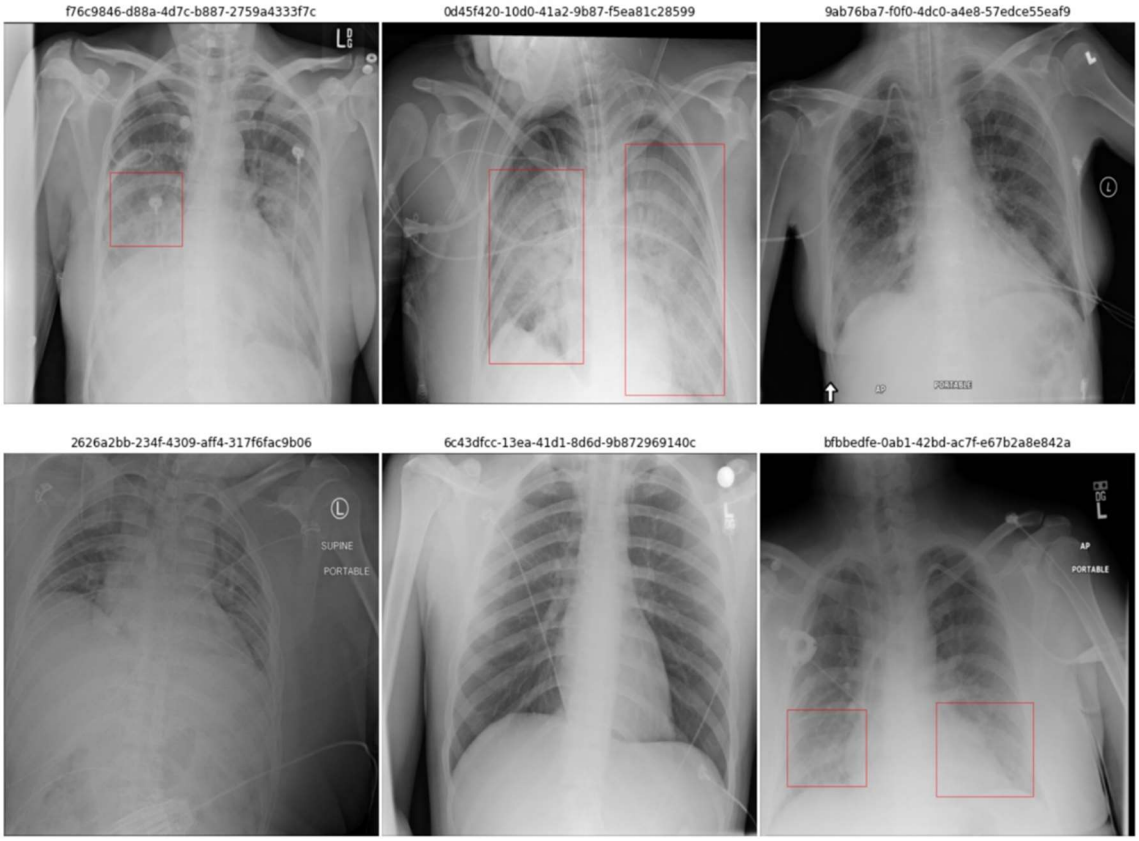


Samples of various X rays



Sample images for X rays with Bounding Boxes





7. Implications

The size, shape, and position of pneumonia may vary. This leads to serious difficulty with accuracy and levels of confidence for the detection. Business Domain Values are as follows:

- Automating Pneumonia screening in chest radiographs, providing affected area details through a bounding box.

- Assist physicians to make better clinical decisions or even replace human judgement in certain functional areas of healthcare (e.g., radiology).
- 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.

8. Limitations

There are still problems with the backbone network of the current detection algorithms. For example, Resnet generally has two problems: a large network depth leading to long training time and massive down-sampling that leads to the target position and semantic information being lost.

The dataset contains three categories which are normal, pneumonia, and abnormal (cancer or other diseases) but only provides the bounding box for pneumonia images. However, the features of pneumonia and abnormal (cancer or other diseases) are similar, which caused the failure to distinguish pneumonia and abnormal images for Resnet50. This results in predicting bounding box for abnormal images.

In general, it's not necessary to explicitly include "negative images". However, in this case since a big portion of the data set contains negative examples, TensorFlow fails to train properly. The loss function does not penalize for detections on negative images. As a result, the model has many false positives where the not normal lung, but no Pneumonia case is annotated as Pneumonia.

Further limitation was due to imbalanced distribution of train and test datasets, most likely due to different labelling approach. Lastly, the limitations of computation resources. Processing Dicom images on Google Collab environment (even with Google Pro) takes longer time to process and execute leading to inefficiency due to timeouts. This constantly affects proper hyperparameter tuning which improves the Model performance.

In real world implementation, the Model requires an API for user inputs before deployment. Flask API is still under development. Furthermore, YOLO v3 can be replaced with faster versions such as YOLO v5 or Detectron that also show the confidence level of the various bounding boxes.

9. Closing Reflections

CNN techniques demonstrate how humans perceive real time objects and offer useful applications to classify images and videos. These procedures can also be applied to detect phishing attacks, in face recognition, scene labelling, image classification, action recognition, human pose estimation, document analysis, speech recognition and text classification. Results observed in comparison to other traditional methods suggest that CNNs provide better accuracy and boosts the performance of the system due to unique

features like shared weights and pretrained models. CNNs are better than other deep learning methods in applications pertaining to computer vision and natural language processing because they mitigate most of the traditional problems.

In future, we will include metrics such as MAP (Mean Average Precision) at different thresholds of IoU to evaluate the performance of object detectors. There is need to train the models for longer epochs and on varying datasets such as NIH to achieve better results. Further image augmentation and hyperparameter tuning and utilization of unlabelled images may result in better generalisation of the Model.

Acknowledgements

Sincere gratitude to all the Great Learning staff for the assistance and guidance throughout the PGP in AIML. Many thanks the Mentor Mr. Girijesh Prasad. His immense experience has equipped us with vast knowledge in this domain. His industry expertise provided a lot of insights to deal with various challenges.

Project Code

https://github.com/girijeshcse/pneumonia_detection

References

[1] https://www.kaggle.com/c/rsna-pneumonia-detection-challenge/overview
[2] https://keras.io/api/applications/vgg/
[3] https://www.kaggle.com/keras/resnet50
[4] https://keras.io/api/applications/inceptionv3/
[5] YOLOv3: An Incremental Improvement, Joseph Redmon, Ali Farhadi, University of Washington.