Capstone project Pneumonia Detection

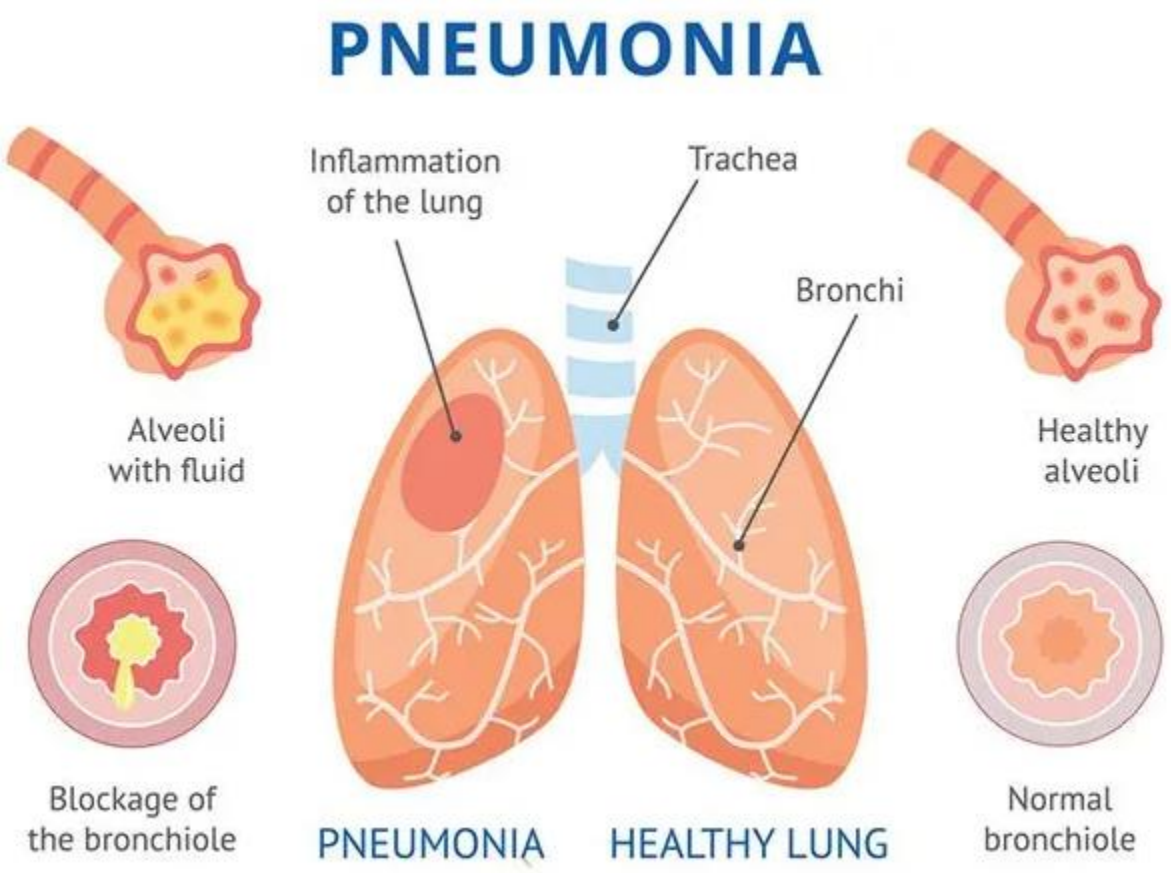
By Chandrakanth B

OCTOBER 23, 2022

Table of Contents

1. 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 your lungs, which are called alveoli. Pneumonia is an inflammatory condition of the lung affecting primarily the small air sacs known as alveoli.Symptoms typically include some combination of productive or dry cough, chest pain, fever and difficulty breathing. The severity of the condition is variable. Pneumonia is usually caused by infection with viruses or bacteria and less commonly by other microorganisms, certain medications or conditions such as autoimmune diseases.Risk factors include cystic fibrosis, chronic obstructive pulmonary disease (COPD), asthma, diabetes, heart failure, a history of smoking, a poor ability to cough such as following a stroke and a weak immune system. Diagnosis is often based on symptoms and physical examination. Chest X-ray, blood tests, and culture of the sputum may help confirm the diagnosis.The disease may be classified by where it was acquired, such as community- or hospital-acquired or healthcare-associated pneumonia.



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 time points and correlation with clinical symptoms and history are helpful in making the diagnosis.

Pneumonia Detection

Now to detection Pneumonia we need to detect Inflammation of the lungs. In this project, would challenge to build an algorithm to detect a visual signal for pneumonia in medical images. Specifically, your algorithm needs to automatically locate lung opacities on chest radiographs.

2. Literature Review

Artificial intelligence techniques can be used to diagnose various diseases such as pneumonia. Research has been done by using multiple methods of machine learning techniques for detecting medical diseases. In this section, we have illustrated the work done in the field of medical image detection. We have reviewed the finding based on strengths and limitations. Concerning medical image detection, various datasets have been used to build up an effective model.

Artificial intelligence has proven to be an effective way in the detection of many diseases. This study presents in artificial intelligence techniques used in the detection, classification and visualization of pneumonia disease in lungs using radiographs of chest. In this review, different reliable databases were searched including research gate, ELSEVIER, Applied sciences and IEEE. Pneumonia is a fatal sort of malady on the off chance that we truly couldn't care less. If we don’t diagnose it in its early stages it can be responsible for 50000 deaths every year. There are two kinds of pneumonia: viral and bacterial. Many researchers have done their research for the identification of pneumonia using machine learning and deep learning methods. This study gives you an overview of the machine and deep learning methods proposed previously for the pneumonia detection. The review is structured based on Deep learning, transfer learning and machine learning methods using chest x-rays images for the early identification of pneumonia. The main objective is to find the limitations of the previous studies and suggestions for the future work.

3. 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.

4. Business Domain Value

Automating Pneumonia screening in chest radiographs, providing affected area details through 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.

- Data info and preprocessing from the Pneumonia data set as Train dataset comprises of 30227 patients details but coordinates of bounding boxes are given, only for 9555 patients. Hence remaining is considered as null value. Please see the below table 1

Index	Data column	Non-Null Count	Data Type
0	Patient ID	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

Table 1

- It has been found that 20672 patients with target 0- don't have pneumonia (chest X rays without bounding boxes)

- 9555 patients with target 1- having pneumonia (chest X rays with bounding boxes)

	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
6	00569f44-917d-4c86-a842-81832af98c30	NaN	NaN	NaN	NaN	0
...	...	...	...	...	...	...
30216	c1cf3255-d734-4980-bfe0-967902ad7ed9	NaN	NaN	NaN	NaN	0
30217	c1e228e4-b7b4-432b-a735-36c48fdb806f	NaN	NaN	NaN	NaN	0
30218	c1e3eb82-c55a-471f-a57f-fe1a823469da	NaN	NaN	NaN	NaN	0
30223	c1edf42b-5958-47ff-a1e7-4f23d99583ba	NaN	NaN	NaN	NaN	0
30224	c1f6b555-2cb1-4231-98f6-50a963976431	NaN	NaN	NaN	NaN	0

20672 rows × 6 columns

Figure 1

- From data set have duplicate patient ids because coordinates of bounding boxes might be duplicates, it indicates that same patient has 2 bounding boxes in their DICOM images.
- Below is the distribution of `Target` and `class` column in figure 2 and values in table2.

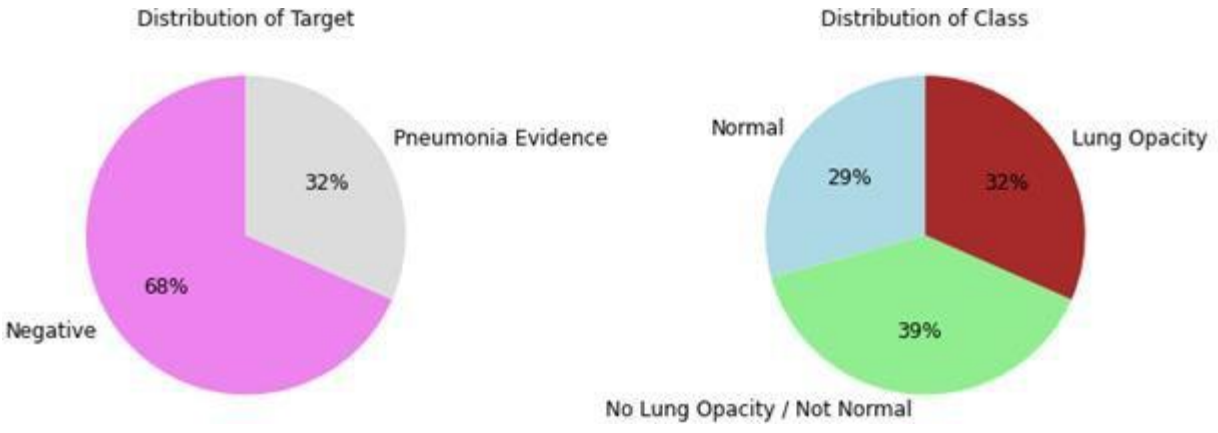


Figure 2

Class	Count
No Lung Opacity / Not Normal	11821
Lung Opacity	9555
Normal	8851

Table 2

- As per train set number of unique patient IDs in the dataset**: 6012**

5. Data Analysis

- Reading the Dicom images meta data and appending it to the training set.

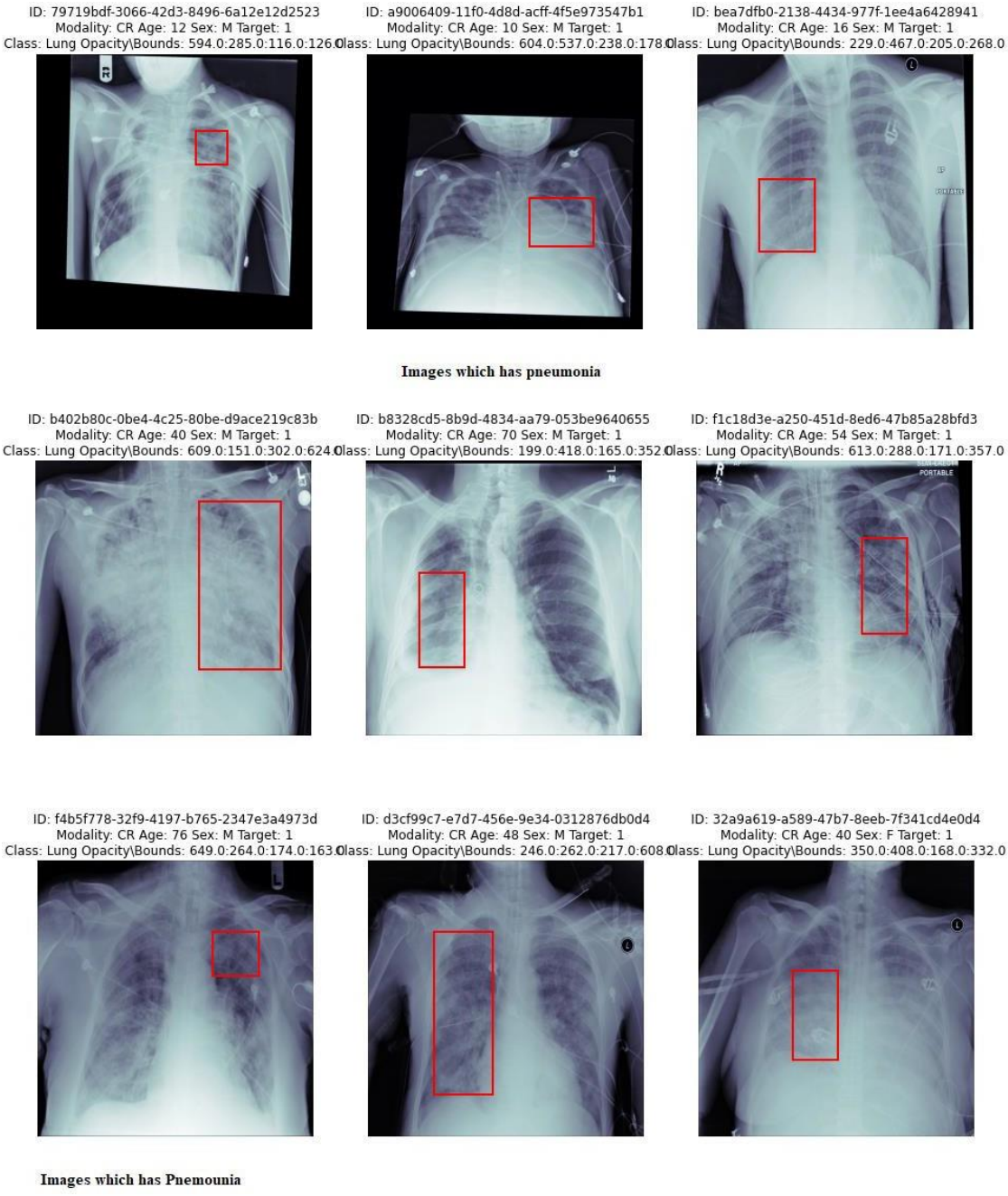
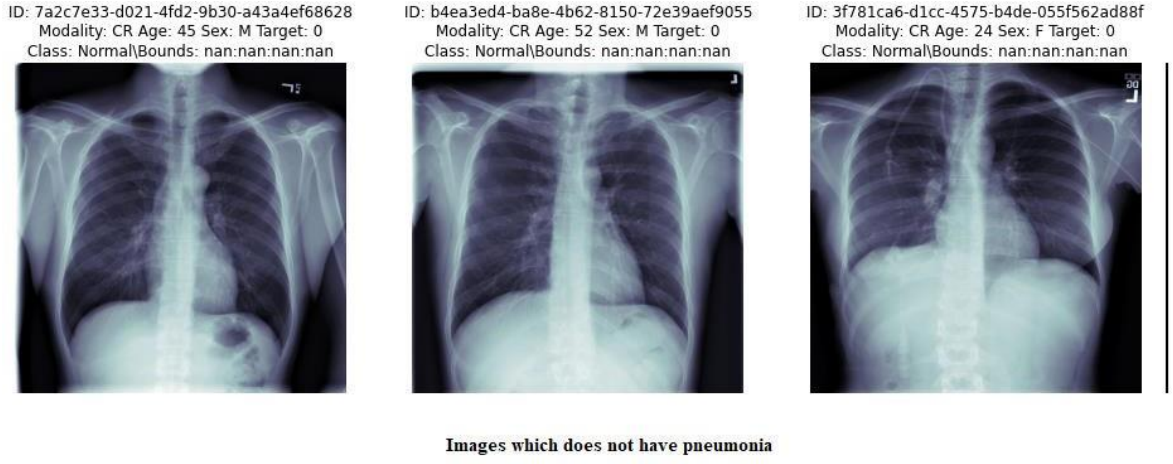


Figure 3


ID: 7a2c7e33-d021-4fd2-9b30-a43a4ef68628
Modality: CR Age: 45 Sex: M Target: 0
Class: Normal\Bounds: nan:nan:nan:nan
ID: b4ea3ed4-ba8e-4b62-8150-72e39aef9055
Modality: CR Age: 52 Sex: M Target: 0
Class: Normal\Bounds: nan:nan:nan:nan
ID: 3f781ca6-d1cc-4575-b4de-055f562ad88f
Modality: CR Age: 24 Sex: F Target: 0
Class: Normal\Bounds: nan:nan:nan:nan
Images which does not have pneumonia


ID: f25fbd2a-7adc-4701-95c4-549983ce8aa7
Modality: CR Age: 4 Sex: F Target: 0
Class: Normal\Bounds: nan:nan:nan:nan
ID: d9f7cfdf-0b34-4647-bd4c-5f64d21f717a
Modality: CR Age: 2 Sex: F Target: 0
Class: No Lung Opacity / Not Normal\Bounds: nan:nan:nan:nan
ID: a8deed82-3273-4c1c-9e72-13bdcee00803
Modality: CR Age: 32 Sex: F Target: 0
Class: Normal\Bounds: nan:nan:nan:nan

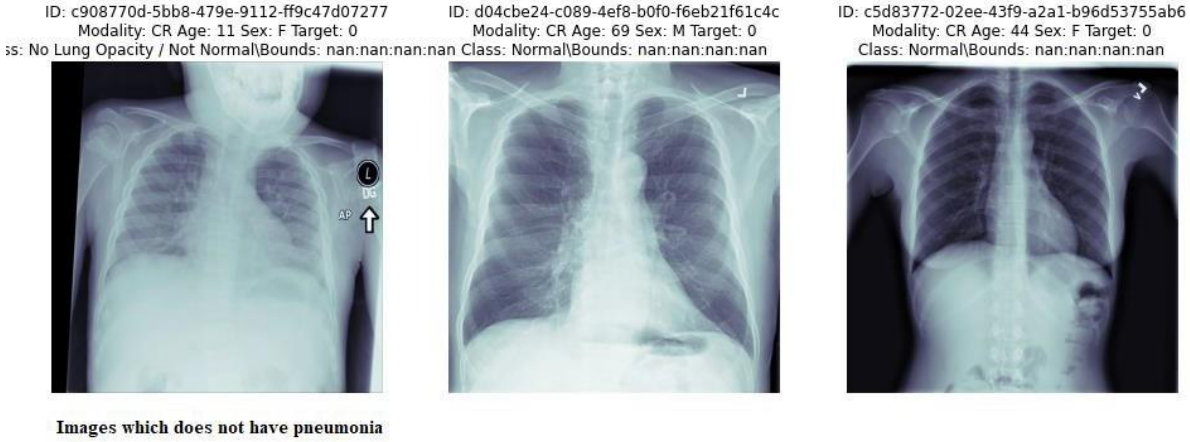

ID: c908770d-5bb8-479e-9112-ff9c47d07277
Modality: CR Age: 11 Sex: F Target: 0
ss: No Lung Opacity / Not Normal\Bounds: nan:nan:nan:nan
ID: d04cbe24-c089-4ef8-b0f0-f6eb21f61c4c
Modality: CR Age: 69 Sex: M Target: 0
Class: Normal\Bounds: nan:nan:nan:nan
ID: c5d83772-02ee-43f9-a2a1-b96d53755ab6
Modality: CR Age: 44 Sex: F Target: 0
Class: Normal\Bounds: nan:nan:nan:nan
Images which does not have pneumonia

Figure 4

Observations of age:

- The mean age is 46 years , whereas minimum age is 1 year and the maximum age is 155 which seems to be an outlier
- 50% of the patients are of around 49 ages, the std deviation is 16 which suggest that age is not normally distributed. Fig 5
- There are 31% of patients with pneumonia and the remaining are of no pneumonia
- There are 23.3% Normal and the remaining is with Lung Opacity. Fig 6

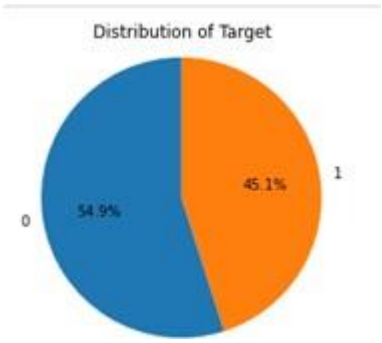


Figure 5

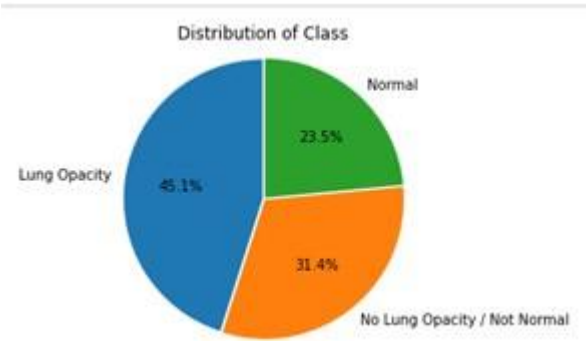


Figure 6

Distribution of target and class:

- Target 0 has only Normal or No Lung Opacity class and Target 1 has only Lung Opacity class

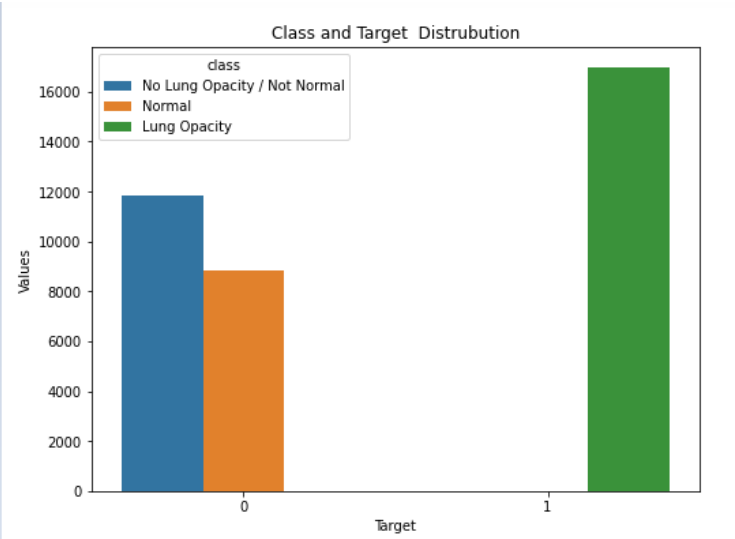


Figure 9

Distribution of Sex among the targets:

- The number of males in both categories is higher than women.

Figure 8

Distributions of Sex among the classes:

- The numbers of males in all classes are higher than women.

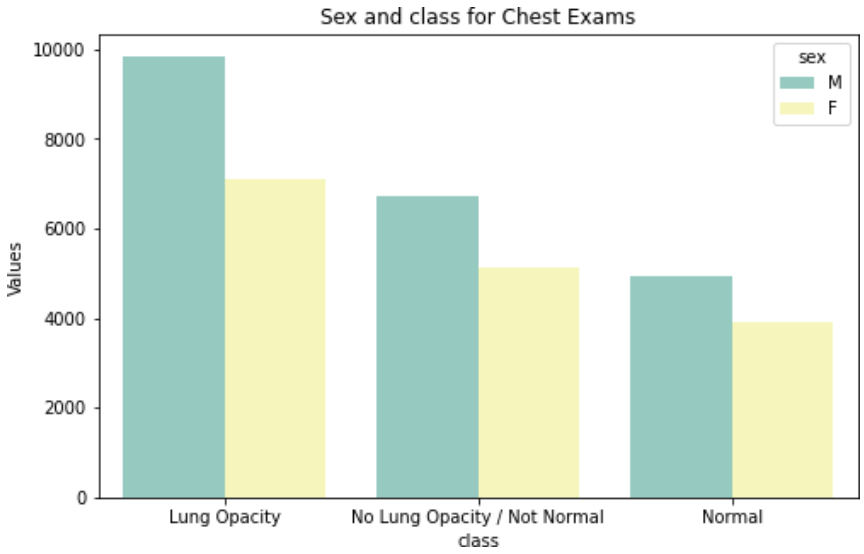


Figure 9

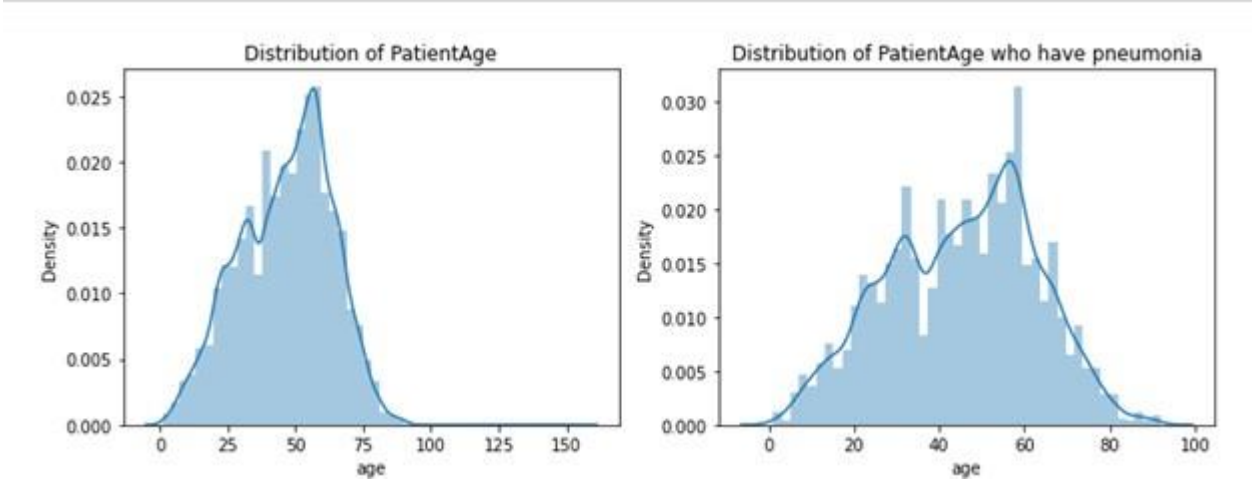


Figure 10

Distribution of Age among target:

- Patients of around 45 years age are of with target class 1(having pneumonia)

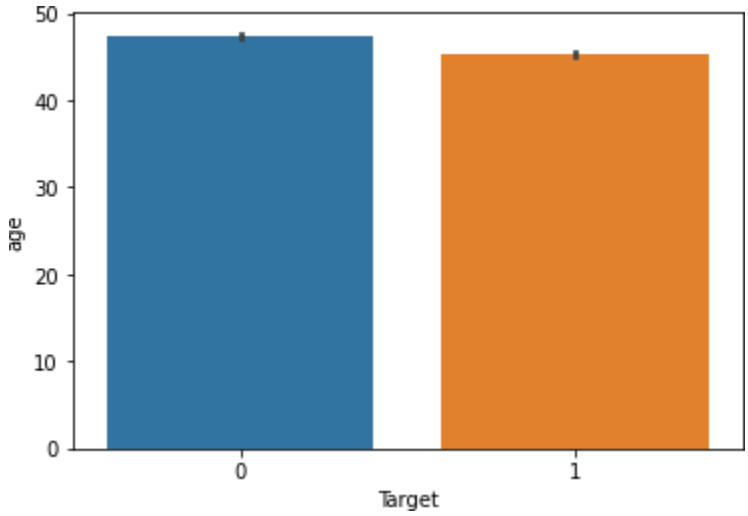


Figure 11

Distribution of Age among class:

- Patients of around 45 year's age are having lung opacity (pneumonia).

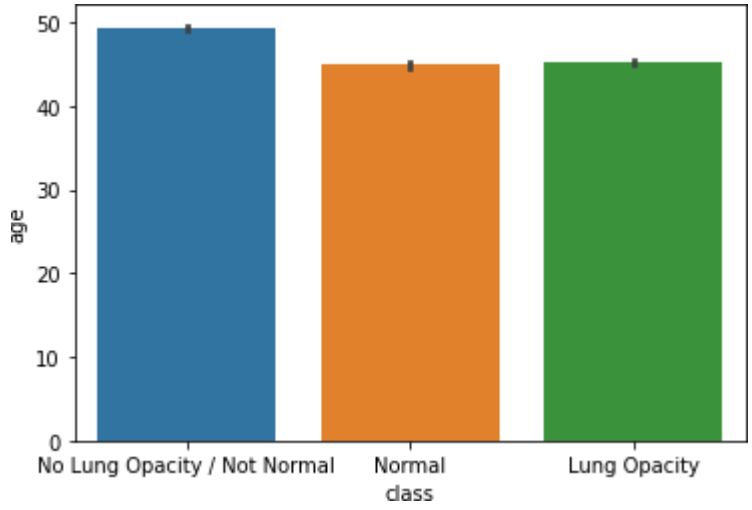


Figure 12

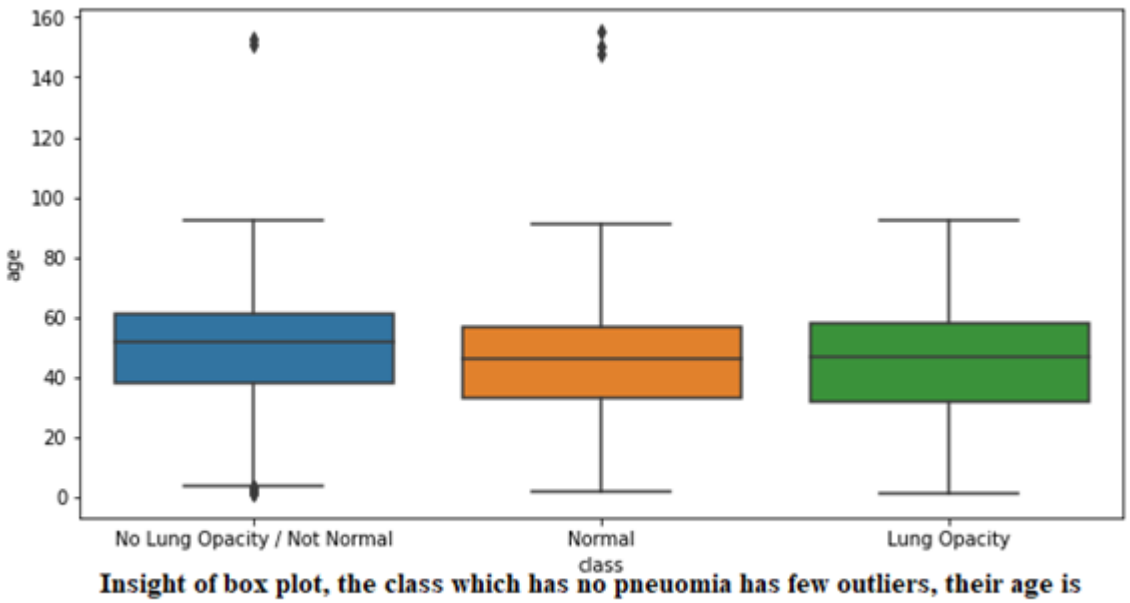


Figure 13

6. Various Modeling Techniques

Deep Learning Methods

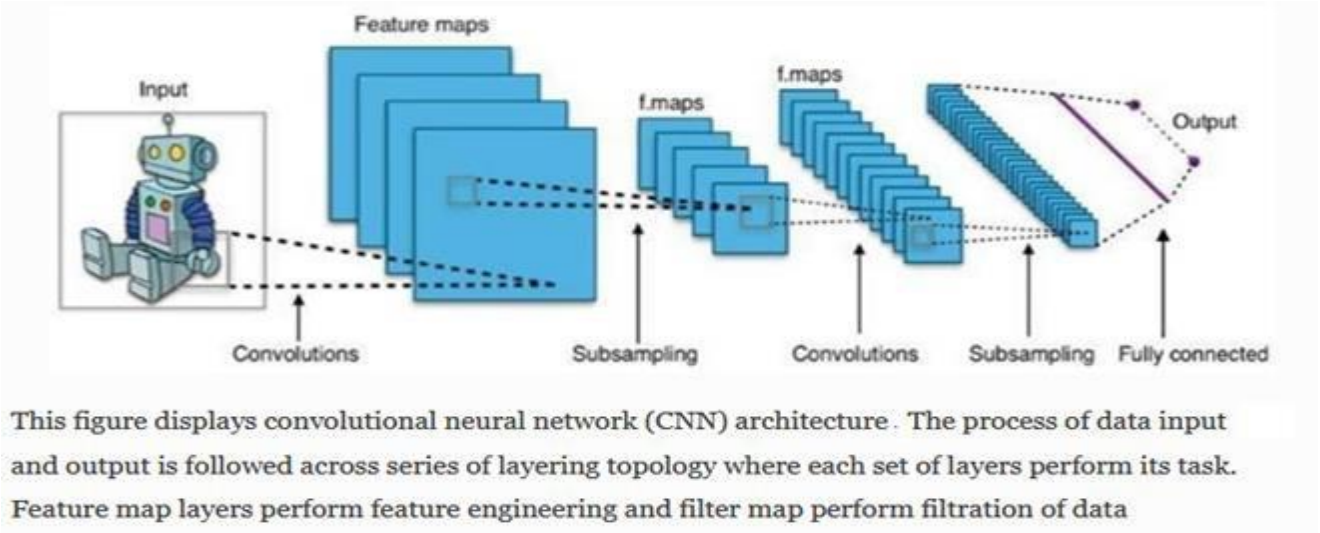
Medical image detection is a complicated task; therefore, an effective approach is needed. Deep learning is one of the techniques that can be used for the training of medical image datasets. In the study, deep learning model of RestNet-101 and RestNet50 was used for pneumonia detection. While considering these techniques, it has resulted in different results based on individual features. Therefore, to compensate this difference, an effective deep learner strategy was introduced that involves the combination of these techniques.

Artificial Neural Network

Artificial neural network (ANN) effectively detects and diagnoses various chest diseases like breast cancer, tuberculosis, and pneumonia infection. Different preprocessing techniques were used to eliminate any irrelevant data. Strategies for enhancing the imaging process were used, including Equalization of the histogram and image filtering. These techniques are crucial in reducing noises and bringing images into sharper focus, thus promoting easy detection of pneumonia. Lung segmentation is an important area of interest in diagnosing pneumonia infection.

Conventional Neural Network

Medical image classification follows complex patterns recognition; therefore, a highly effective ML model is needed. To make it possible, deep learning plays an important role, and among them, CNN is one of the effective approaches for pattern recognition due to its layering topology



This figure displays convolutional neural network (CNN) architecture. The process of data input and output is followed across series of layering topology where each set of layers perform its task. Feature map layers perform feature engineering and filter map perform filtration of data

Figure 14

Other Techniques

The usefulness of computer-aided techniques was studied in detecting lung tuberculosis. Examining various parameters like reducing patient waiting times was considered to obtain an X-ray and diagnosis lung tuberculosis. To perform diagnosis, the radiologists carried out a visual examination on textual features of thoracic X-ray images. They also used the principal component analysis method in measuring the outcomes of the study. It was identified, classified, and differentiated between TB and non-TB objects centered on various arithmetical features from experiment. The challenge of considering the PCA includes the lower interpretability issues, and also data organisation is an essential requirement for PCA to work effectively. PCA finds linear correlation among the variable, which is not ideal in many cases.

7. CNN Basic Modeling Technique

- In the Target section, there are 20672 records with no pneumonia.9555 with pneumonia.
- In class section there are 8851 normal cases. 9555 - Person with lung opacity
- 11821 - Person with No Lung Opacity / Not Normal (May be not yet recovered during/after treatment)
- CNN architecture is used to train the model.

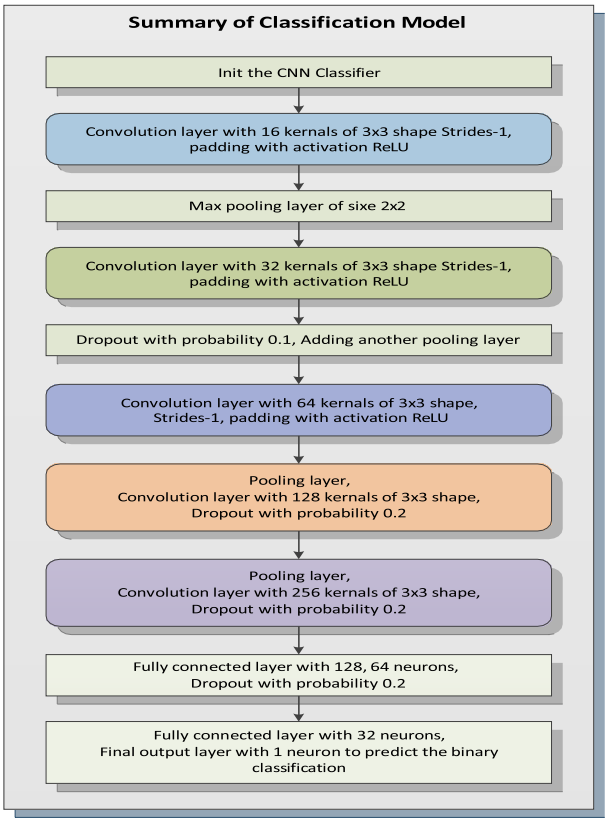


Figure 15

Model-1 with 2 classes

- Evaluating the accuracy, there has been dip in the training loss but validation loss has no much significant dip in loss.
- Validation accuracy approximately went up to 81.14%.
- While training accuracy was 83.45% and test accuracy was 81.34%.

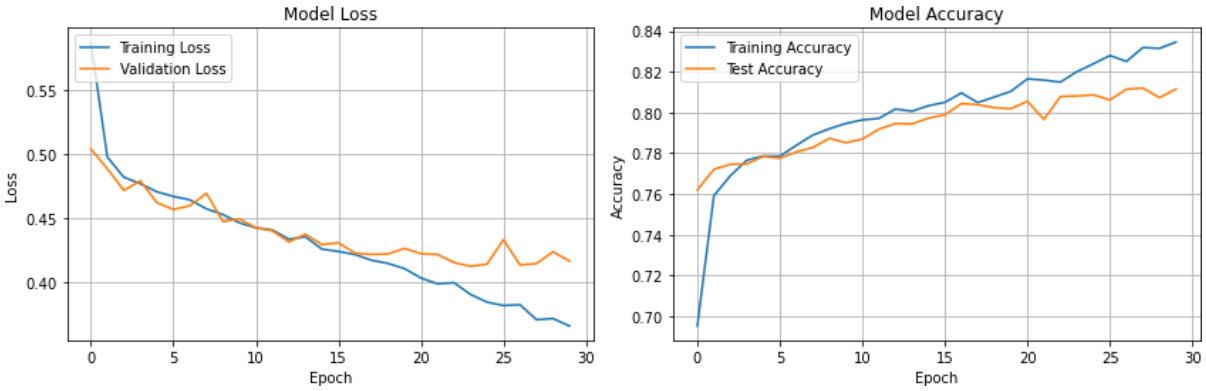
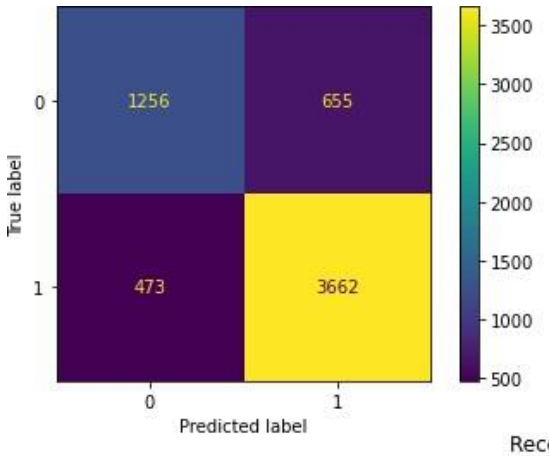


Figure 16

- Evaluating the Model, We should be fairly satisfied with the base model's performance. Below produces an evaluation report borrowing some of the better evaluation pieces of Scikit-Learn.



----------Classification Report----------

	precision	recall	f1-score	support
0	0.73	0.66	0.69	1911
1	0.85	0.89	0.87	4135
accuracy			0.81	6046
macro avg	0.79	0.77	0.78	6046
weighted avg	0.81	0.81	0.81	6046

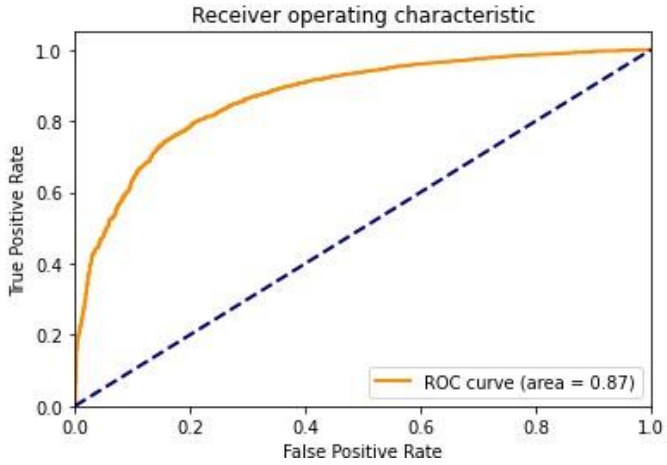


Figure 17

Model-2 with 3 classes

- There has been dip in the training loss but validation loss increased at the later epochs without any much significant difference.
- Validation accuracy went up to 64.89%.
- While training accuracy was 70.06% and testing accuracy was 64.80%.
- The model has predicted mostly wrong in this case to the Target 0. Type 2 error.

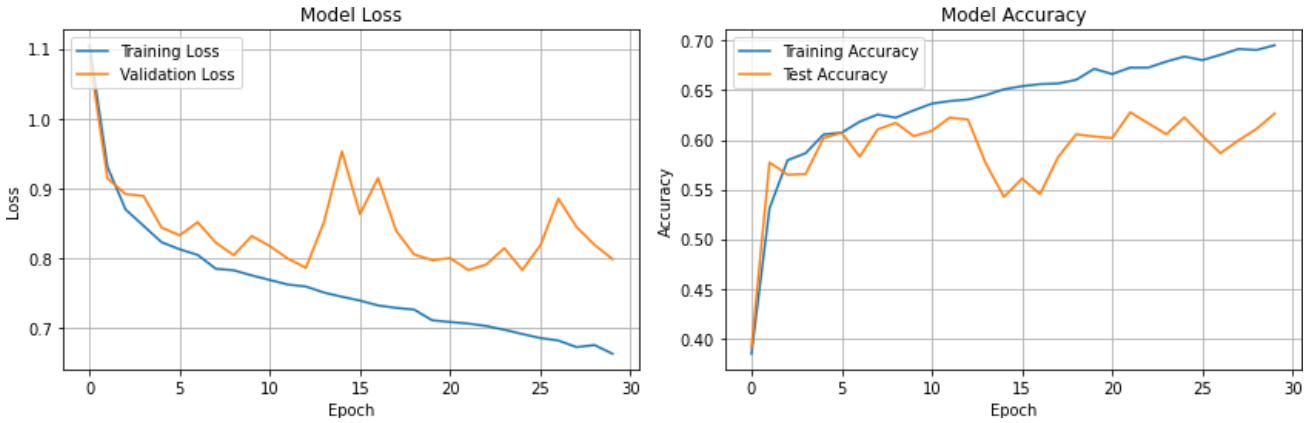


Figure 18

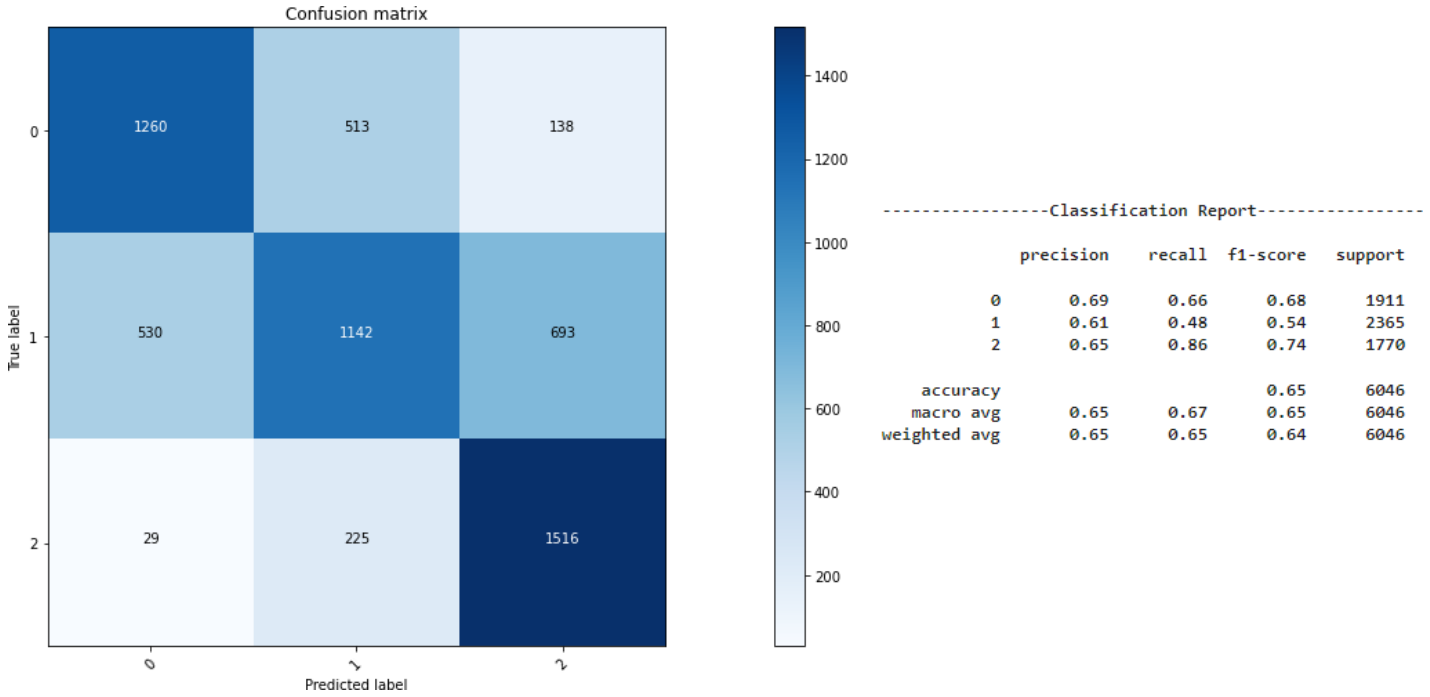


Figure 19

8. Data Augmentation

- CNN architecture to train the model with data augmentation
- Data augmentation involves perturbing the training data to increase the size of the data and reduce overfitting.
- Normalize the pixel values by dividing by 255.

- Used horizontal flips, rotations of up to 40 degrees, and other distortions.

- Importantly, the validation and test sets only rescale the pixel intensity.
- Plotting the augmented images gives a sense of how the images are manipulated.

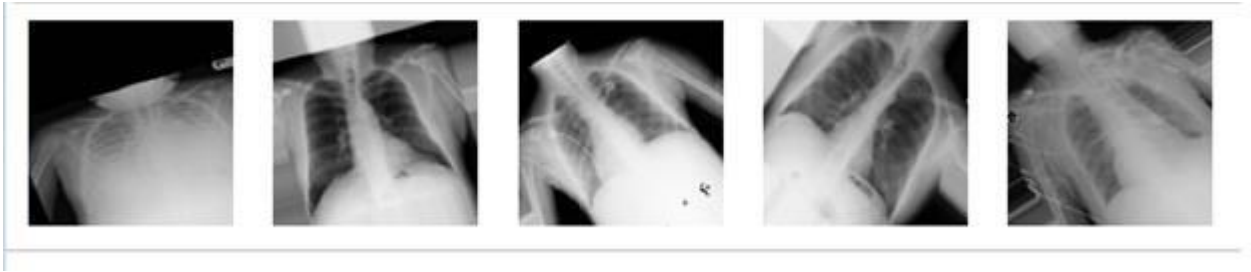


Figure 20

Model-1 with Data Augmentation (2 classes)

- Evaluating the accuracy, we have got training accuracy of 79.72% and test accuracy of 80.28%.
- If you train this model-1 with more images will get good accuracy score. Plotting the accuracy vs loss graph.

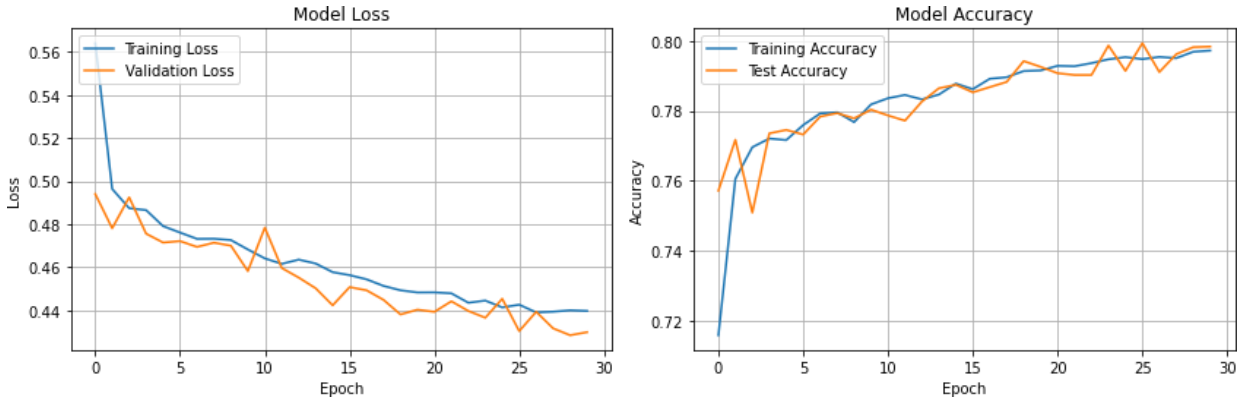


Figure 21

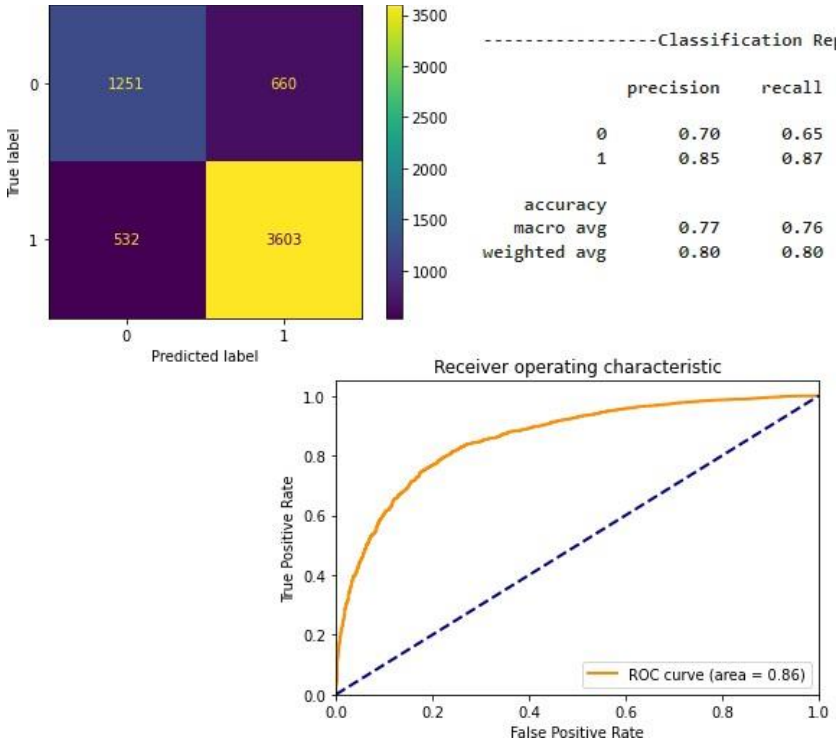


Figure 22

9. Transfer Learning

- Transfer learning is a machine learning method where a model developed for a task is reused as the starting point for a model on a second task.
- It is a popular approach in deep learning where pre-trained models are used as the starting point on computer vision and natural language processing tasks given the vast compute and time resources required to develop neural network models on these problems and from the huge jumps in skill that they provide on related problems.

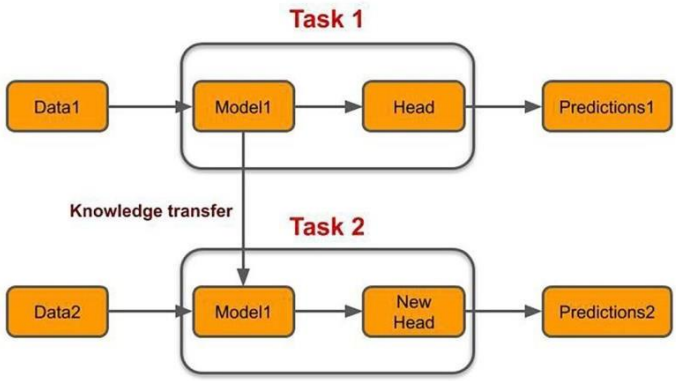


Figure 23

VGG16

- **VGG-16** is a transfer learning algorithm which means it's an algorithm with 16 convolutional layers that focuses on storing knowledge that can be applied to different but related problems.
- **VGGNet** is a well-documented and globally used architecture for convolutional neural network.
- **Include_top=False** to remove the classification layer that was trained on the ImageNet dataset and set the model as not trainable.

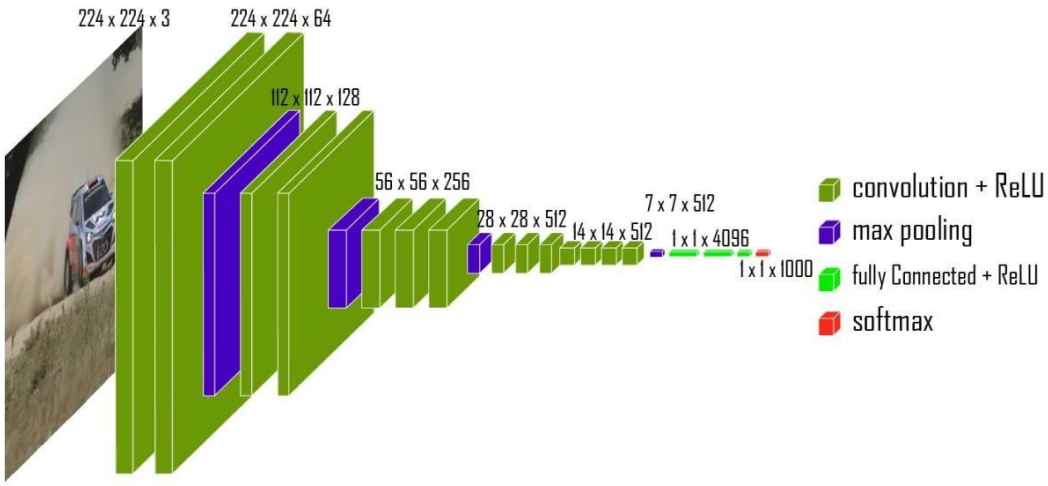


Figure 24

Modeling - Using VGG16 with trainable = False

- There has been dip in the training loss but validation loss increased at the later epochs without any much significant difference.
- Validation accuracy went up to 80.23%.
- While training accuracy was 81.97% and testing accuracy was 79.85%.

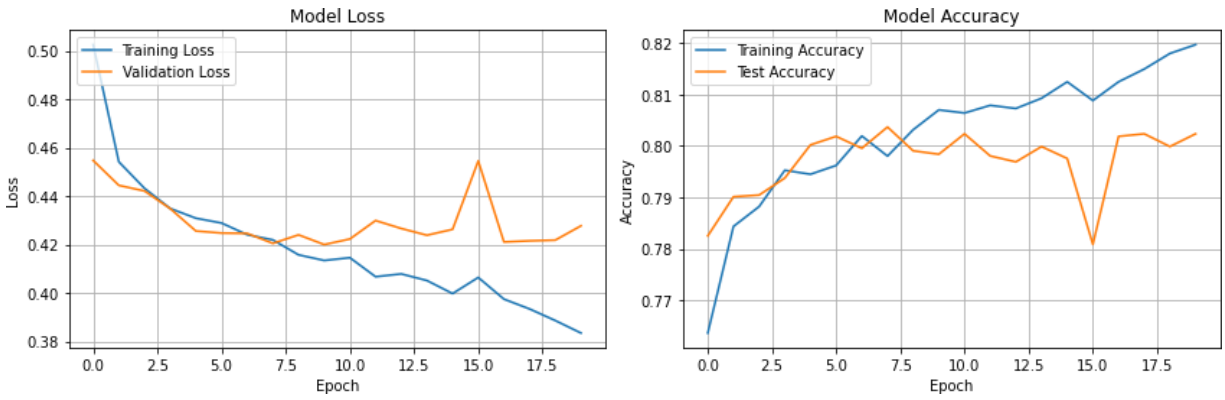


Figure 25

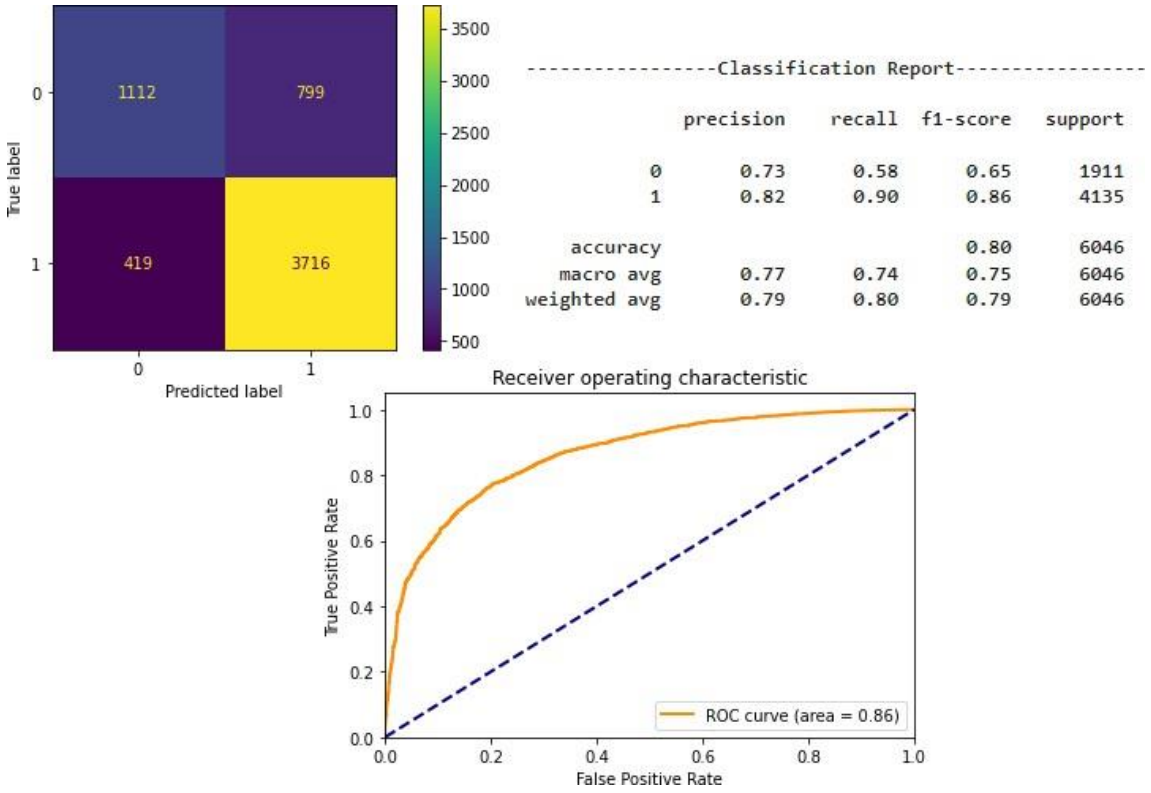


Figure 26

Modeling - Using VGG16 with trainable = True

- There has been dip in the training loss but validation loss increased at the later epochs with much significant difference.
- Validation accuracy went up to 68.39%.
- While training accuracy was 68.39% and testing accuracy was 68.39%.

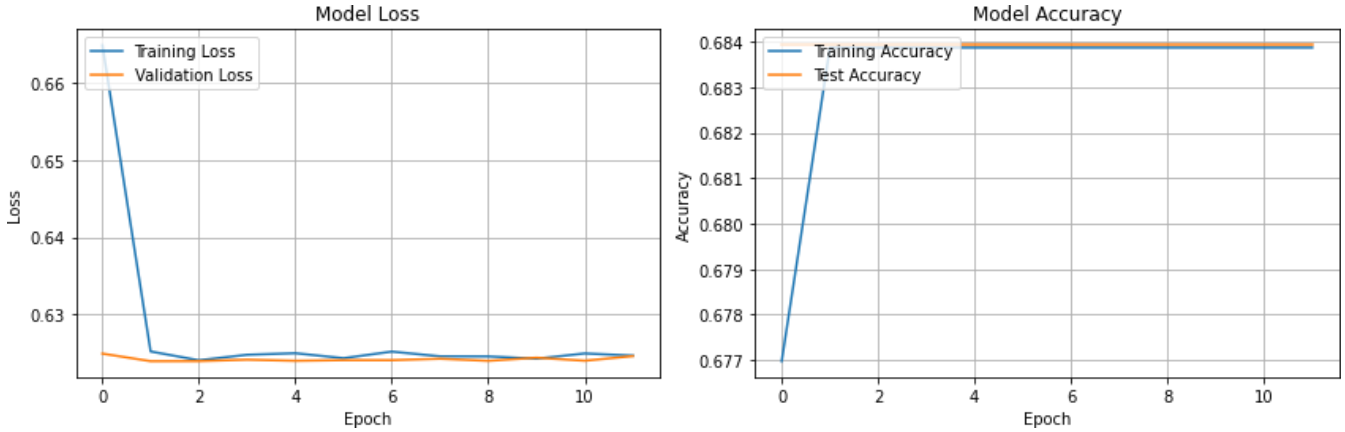


Figure 27

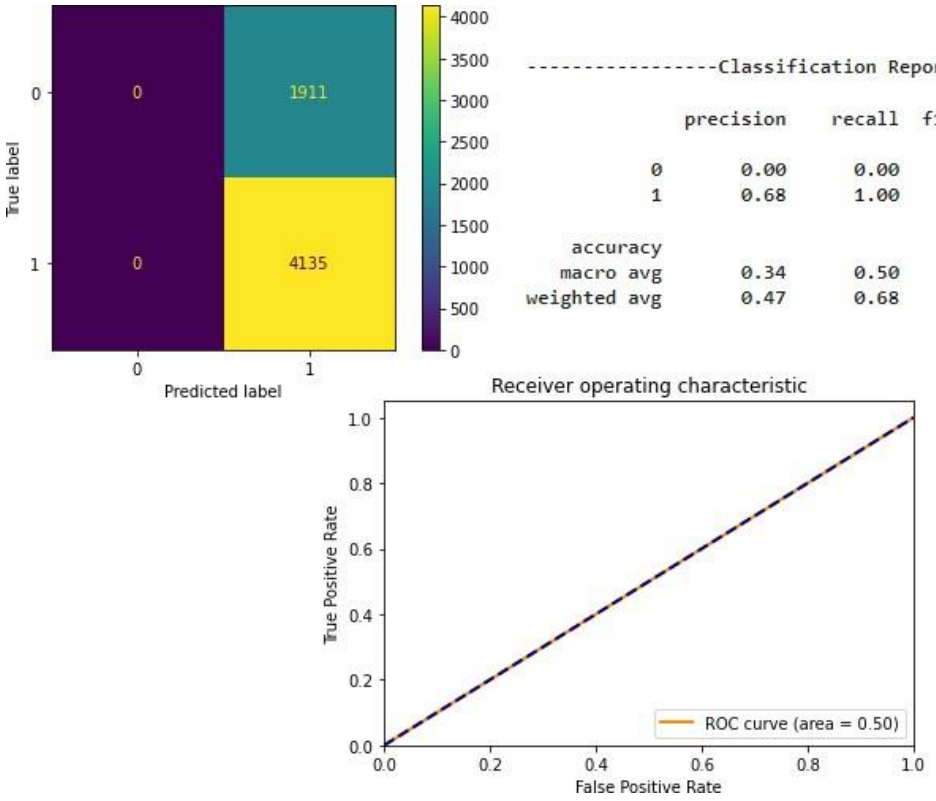


Figure 28

InceptionNet

- An inception network is a deep neural network with an architectural design that consists of repeating components referred to as Inception modules.
- It is also known as **GoogleNet**, this architecture presents sub-networks called inception modules, which allows fast training computing, complex patterns detection, and optimal use of parameters.

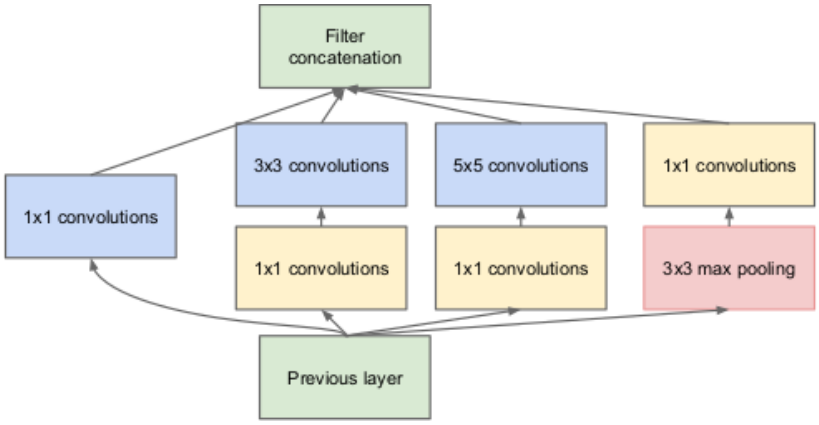


Figure 27

- There has been dip in the training loss but validation loss increased and decreased at the later epochs without any much significant difference.
- Validation accuracy went up to 82.15%.
- While training accuracy was 88.92% and testing accuracy was 82.04%.

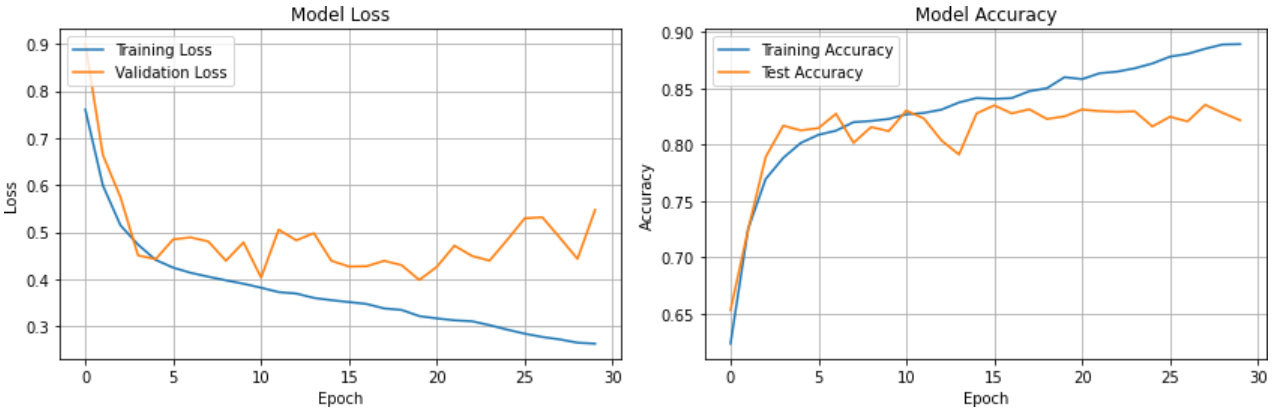


Figure 30

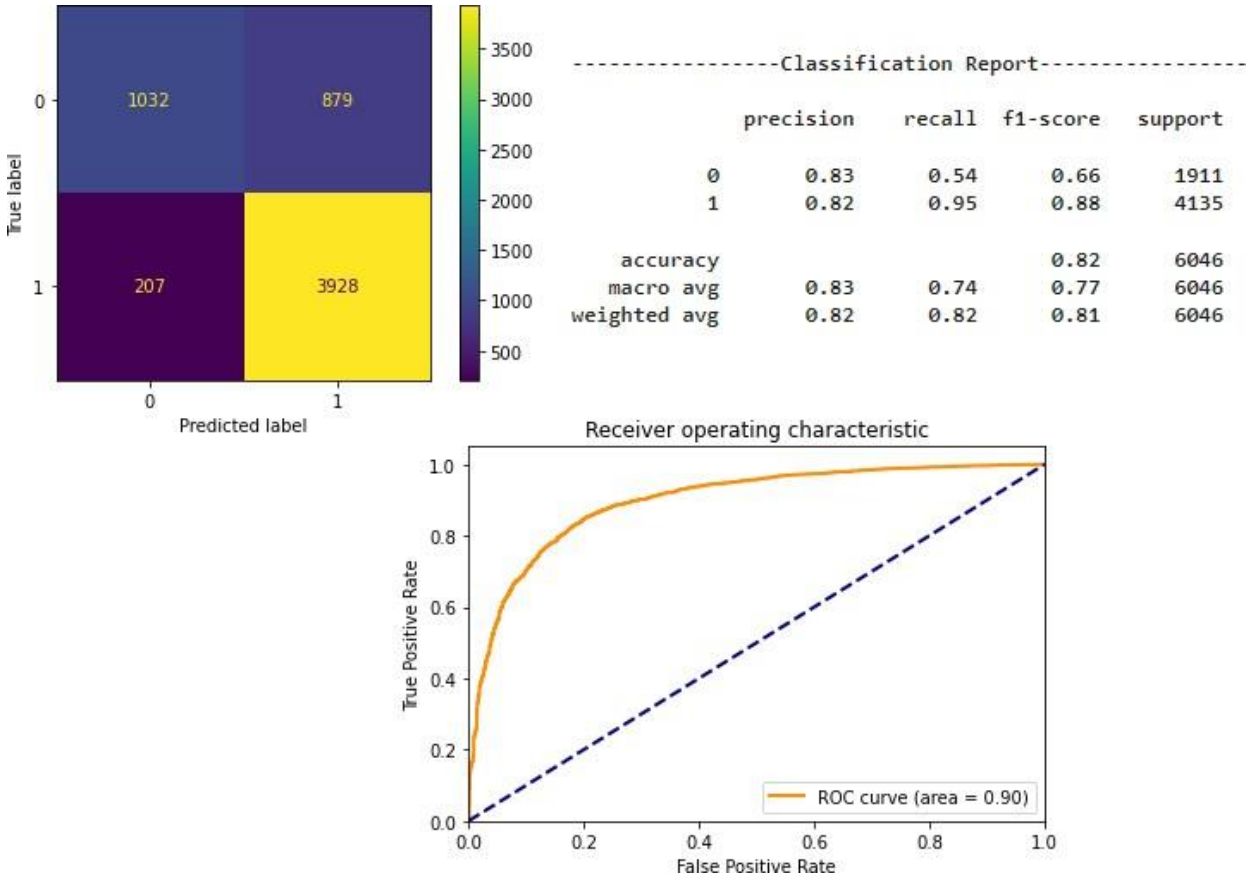


Figure 31

DenseNet

- DenseNet is one of the new discoveries in neural networks for visual object recognition.
- DenseNet is quite similar to ResNet with some fundamental differences.
- ResNet uses an additive method (+) that merges the previous layer (identity) with the future layer, whereas DenseNet concatenates (.) the output of the previous layer with the future layer.

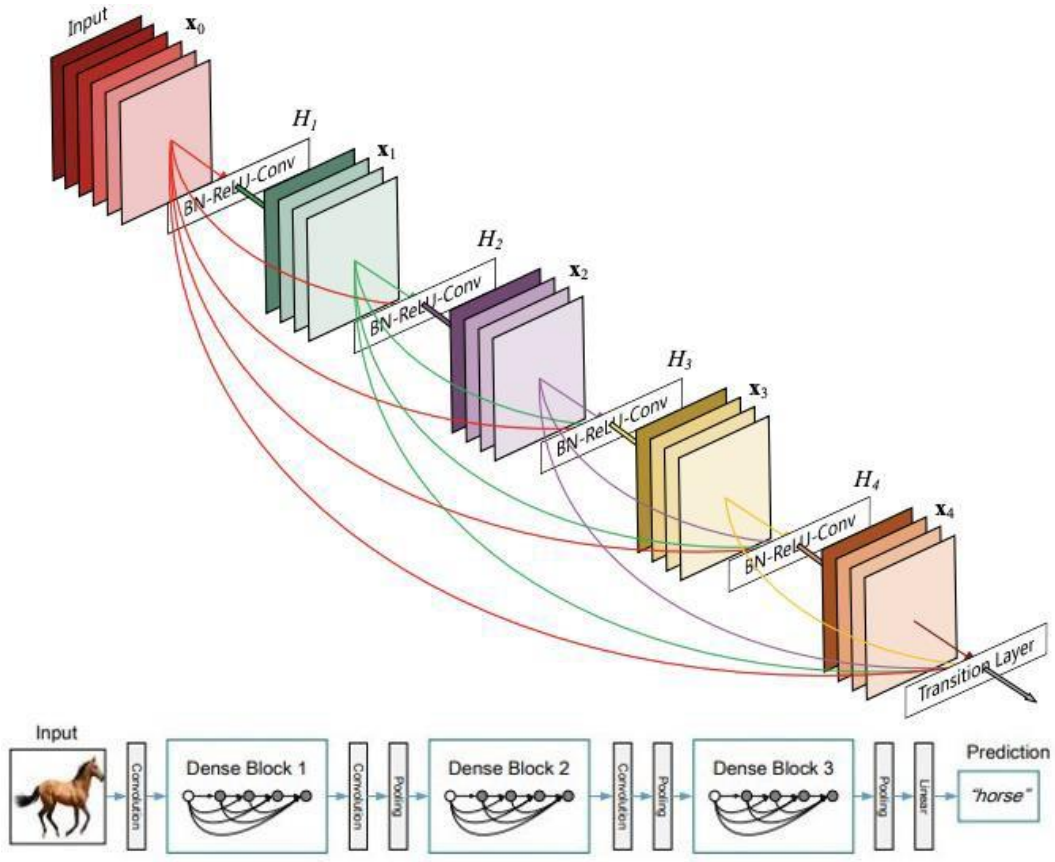


Figure 32

- There has been dip in the training loss but validation loss increased at the later epochs without any much significant difference.
- Validation accuracy went up to 81.41%.
- While training accuracy was 98.81% and testing accuracy was 81.79%.

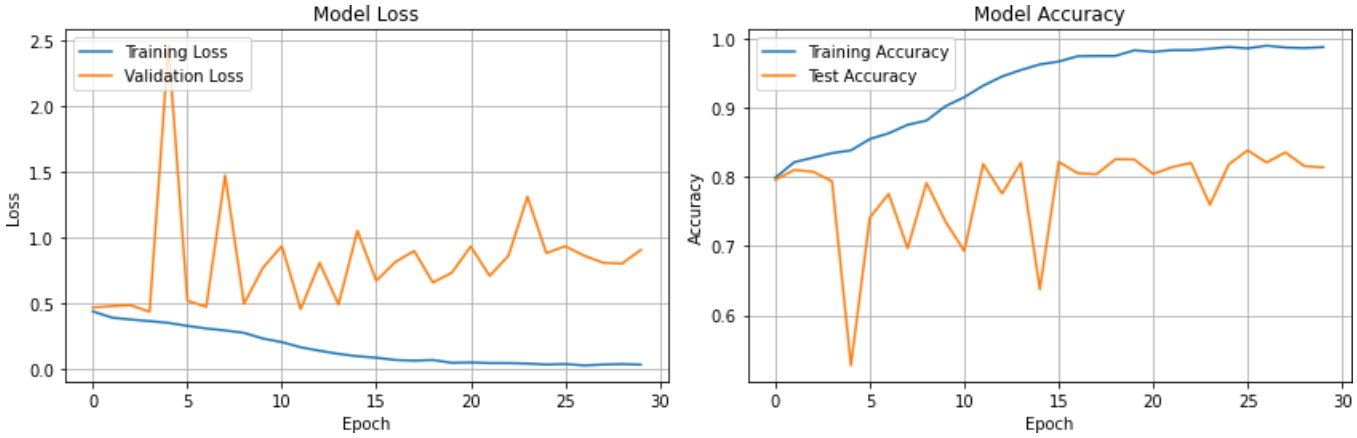
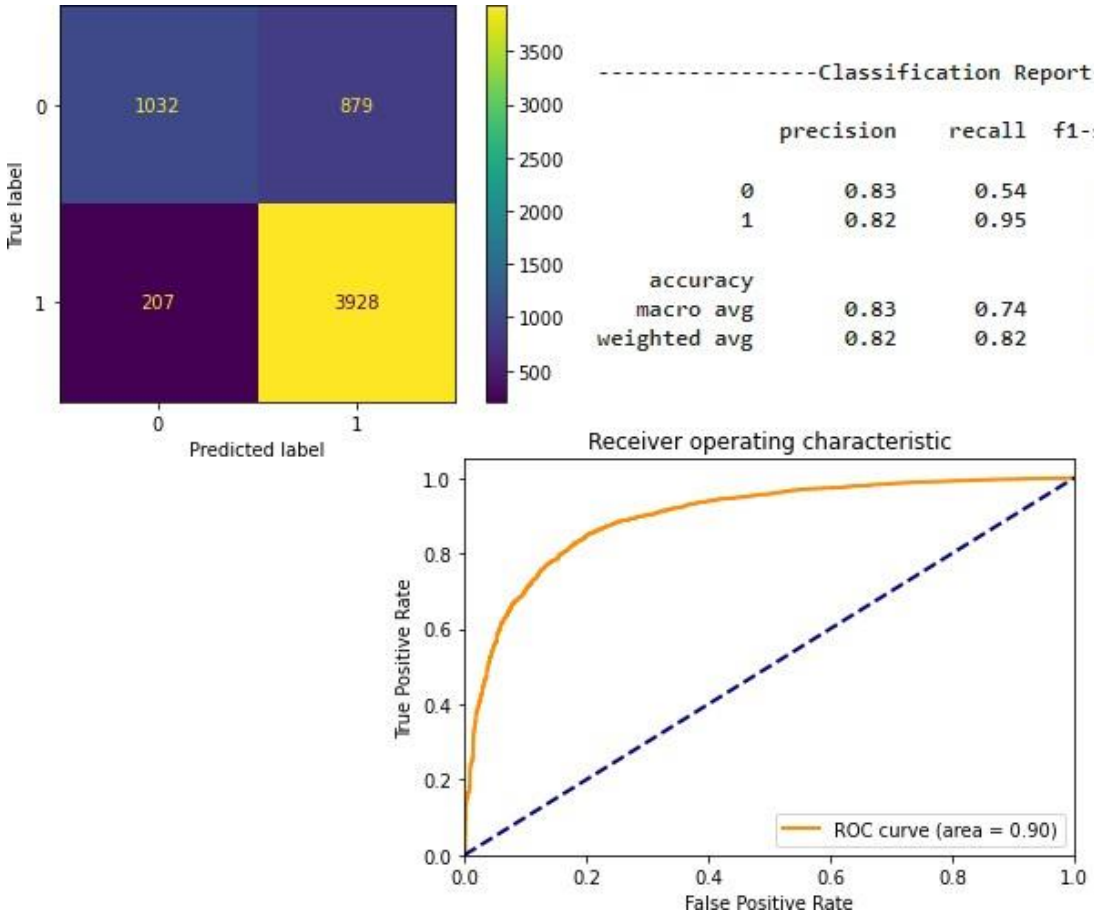


Figure 33



----------Classification Report----------

	precision	recall	f1-score	support
0	0.83	0.54	0.66	1911
1	0.82	0.95	0.88	4135
accuracy			0.82	6046
macro avg	0.83	0.74	0.77	6046
weighted avg	0.82	0.82	0.81	6046

Figure 34

10. Conclusion

Insights of transfer learning and final conclusion:

- Among the CNN model with 2 classes (clubbed the classes of no lung opacity and normal) and CNN model with 3 classes using stratify sampling (in order to make sure of the same ratio of classes for modelling), it has been found that accuracy and recall of pneumonia class (class 1) are higher for model with 2 classes, so we have used 2 classes only for further analysis/modelling.
- Then we have tried the CNN model (with 2 classes) after augmenting the images and this model has shown slightly increase in recall percentage with relatively no change in test accuracy in spite of slight decrease in train accuracy.
- Later on we have tried with transfer learning models of VGG 16. Among the trainable- false and true, it has been observed that VGG 16 trainable -False has shown relatively higher accuracy when compared to VGG trainable- True.
- Further we have tried with inception net (Google Net) and dense net. Dense net has given highest raining accuracy among all the techniques but quite less recall percentage.
- Finally, it has been found that inception net is the best among all transfer learning models being tried and has given highest recall percentage which is the main moto of this project since this belongs to health care domain and need to focus on decreasing the false negative rate.

11. References

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

https://link.springer.com/chapter/10.1007/978-981-15-6321-8_12

https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/