Capstone Project-Interim Report

Dhinesh Kumar Ganeshan

10-Feb-2020

Executive Summary

This project represents a culmination of the Ten modules of the AI and ML Specialization offered by Great Lakes Executive Learning and University of Texas at Austin via Great Learning.

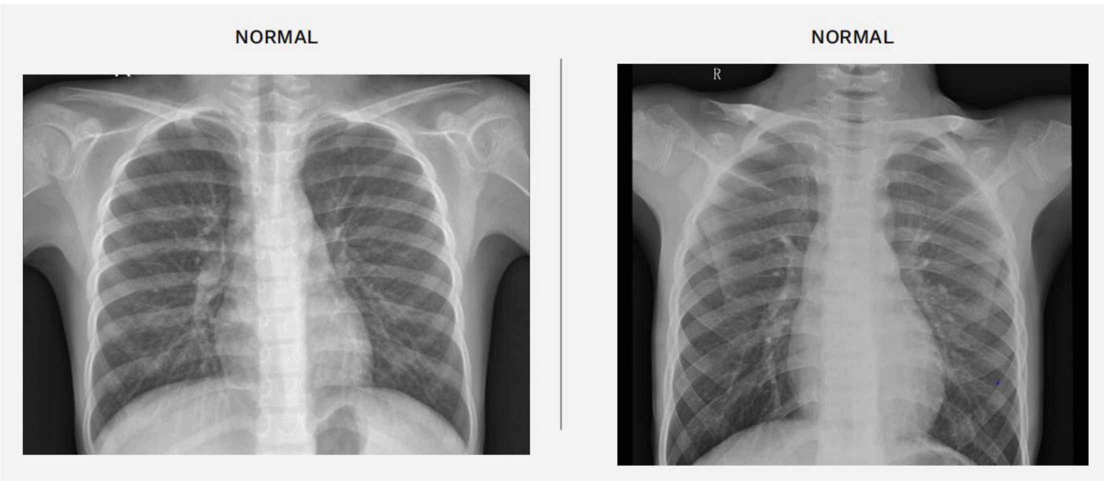
The Pneumonia Detection prediction model is built based on the basics of Computer Vision Technique techinques learned throughout the specialization.

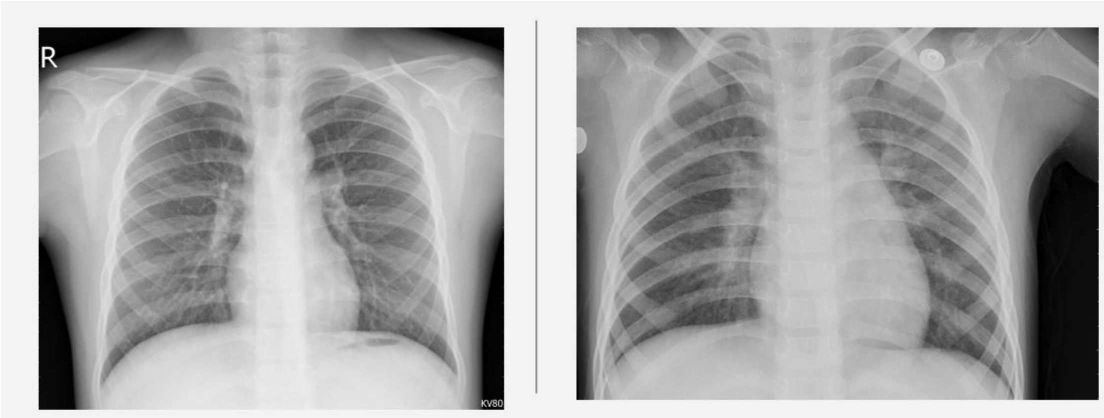
Sampling the Data

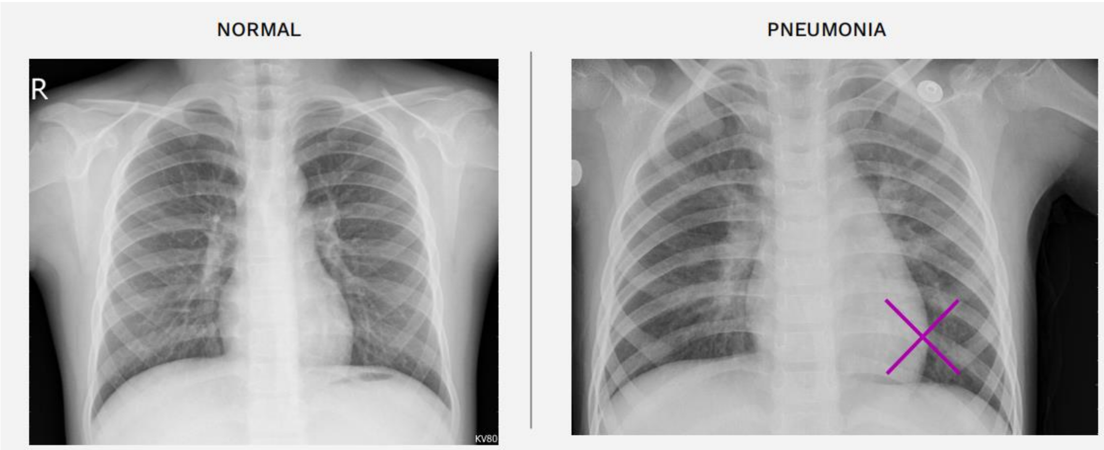
The corpora given comprises X-RAY of Lung images of very large datasets from Kaggle competition - of more than1000 and above DICOM images with file size of over 4 GB.

Kaggle dataset - quick look

- 5’863 X-Ray images
- pediatric patients 1-5 years old
- labeled by several specialists







In order to be accomodated within my system limitations (and in keeping with the approach recommended) a sample of the corpus was selected for study in order to build and train the prediction model.

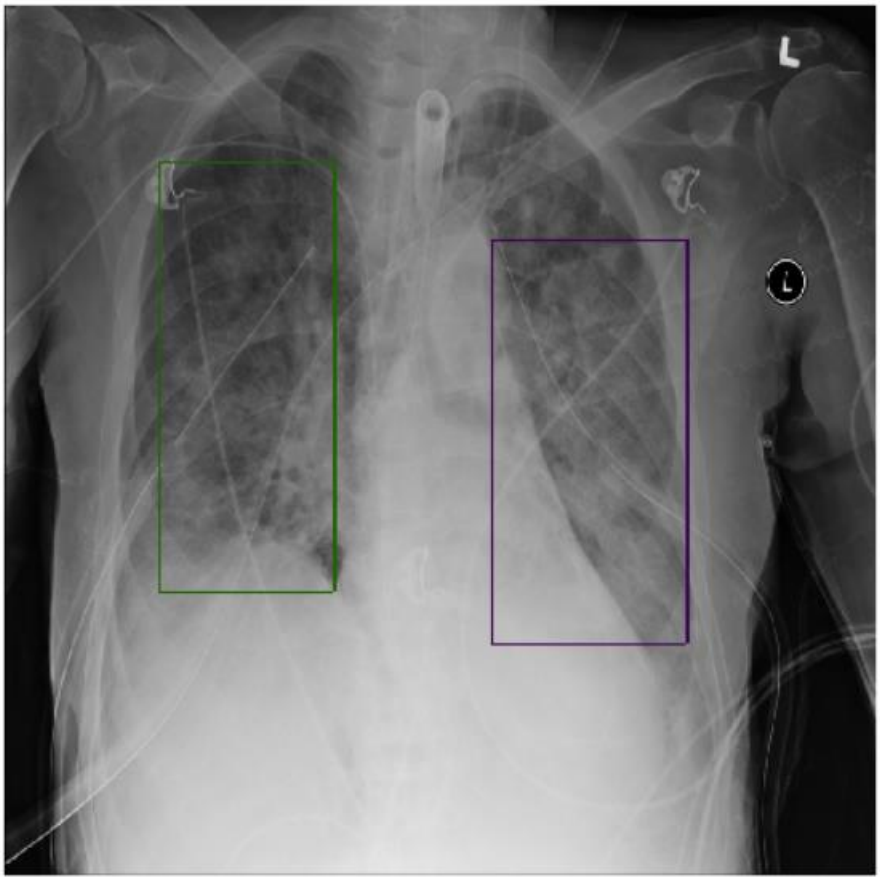
It is very important to understand the data in DICOM files before we work on Prediction Models.

```
import pydicom
dcm_file = '../content/RSNAdata/data/stage_2_train_images/%s.dcm' % pat_choose

dcm_data = pydicom.read_file(dcm_file)

print(dcm_data)
```

```
(0008, 0005) Specific Character Set              CS: 'ISO_IR 100'
(0008, 0016) SOP Class UID                       UI: Secondary Capture Image Storage
(0008, 0018) SOP Instance UID                    UI: 1.2.276.0.7230010.3.1.4.8323329.1556.1517874291.545552
(0008, 0020) Study Date                          DA: '19010101'
(0008, 0030) Study Time                          TM: '000000.00'
(0008, 0050) Accession Number                    SH: ''
(0008, 0060) Modality                            CS: 'CR'
(0008, 0064) Conversion Type                     CS: 'WSD'
(0008, 0090) Referring Physician's Name          PN: ''
(0008, 103e) Series Description                  LO: 'view: AP'
(0010, 0010) Patient's Name                      PN: '00f08de1-517e-4652-a04f-d1dc9ee48593'
(0010, 0020) Patient ID                          LO: '00f08de1-517e-4652-a04f-d1dc9ee48593'
(0010, 0030) Patient's Birth Date                DA: ''
(0010, 0040) Patient's Sex                       CS: 'M'
(0010, 1010) Patient's Age                       AS: '58'
(0018, 0015) Body Part Examined                  CS: 'CHEST'
(0018, 5101) View Position                       CS: 'AP'
(0020, 000d) Study Instance UID                  UI: 1.2.276.0.7230010.3.1.2.8323329.1556.1517874291.545551
(0020, 000e) Series Instance UID                 UI: 1.2.276.0.7230010.3.1.3.8323329.1556.1517874291.545550
(0020, 0010) Study ID                            SH: ''
(0020, 0011) Series Number                       IS: "1"
(0020, 0013) Instance Number                     IS: "1"
(0020, 0020) Patient Orientation                 CS: ''
(0028, 0002) Samples per Pixel                   US: 1
(0028, 0004) Photometric Interpretation          CS: 'MONOCHROME2'
(0028, 0010) Rows                                US: 1024
(0028, 0011) Columns                             US: 1024
(0028, 0030) Pixel Spacing                       DS: [0.139, 0.139]
(0028, 0100) Bits Allocated                      US: 8
(0028, 0101) Bits Stored                         US: 8
(0028, 0102) High Bit                            US: 7
(0028, 0103) Pixel Representation                US: 0
(0028, 2110) Lossy Image Compression             CS: '01'
(0028, 2114) Lossy Image Compression Method      CS: 'ISO_10918_1'
(7fe0, 0010) Pixel Data                          OB: Array of 143458 elements
```

Understanding the data from the DICOM files is imperative to being able to ensure one's conceptualization of bounding boxes on the arrays from those files. We need to visualize those boxes in order to augment the knowledge regarding the visual aspects of pneumonia:

```
Patient ID: 00f08de1-517e-4652-a04f-d1dc9ee48593
```




```
patientId    00f08de1-517e-4652-a04f-d1dc9ee48593
x                                             181
y                                             184
width                                         206
height                                        506
Target                                          1
class                                Lung Opacity
Name: 19, dtype: object
```

Now that we have visualized the bounding boxes and the XRAYs, we should take a look at the demographics of our features

Next Perform EDA of the dataset Setup a dataframe - Gender, Viewing Position , Age etc.

	patientId	x	y	width	height	Target	class
0	0004cfab-14fd-4e49-80ba-63a80b6bddd6	NaN	NaN	NaN	NaN	0	No Lung Opacity / Not Normal
1	00313ee0-9eaa-42f4-b0ab-c148ed3241cd	NaN	NaN	NaN	NaN	0	No Lung Opacity / Not Normal
2	00322d4d-1c29-4943-afc9-b6754be640eb	NaN	NaN	NaN	NaN	0	No Lung Opacity / Not Normal
3	003d8fa0-6bf1-40ed-b54c-ac657f8495c5	NaN	NaN	NaN	NaN	0	Normal
4	00436515-870c-4b36-a041-de91049b9ab4	264.0	152.0	213.0	379.0	1	Lung Opacity

After dropping 29 features

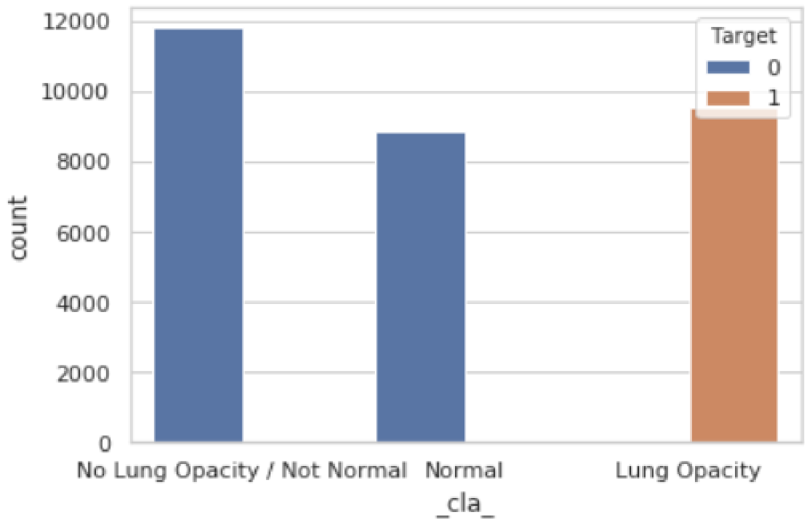
```
Dropped 29 useless features
```

```
df_meta.head()
```

	patientId	x	y	width	height	Target	class	PatientAge	PatientSex	SeriesDescription	ViewPosition
0	0004cfab-14fd-4e49-80ba-63a80b6bddd6	NaN	NaN	NaN	NaN	0	No Lung Opacity / Not Normal	51	F	view: PA	PA
1	00313ee0-9eaa-42f4-b0ab-c148ed3241cd	NaN	NaN	NaN	NaN	0	No Lung Opacity / Not Normal	48	F	view: PA	PA
2	00322d4d-1c29-4943-afc9-b6754be640eb	NaN	NaN	NaN	NaN	0	No Lung Opacity / Not Normal	19	M	view: AP	AP
3	003d8fa0-6bf1-40ed-b54c-ac657f8495c5	NaN	NaN	NaN	NaN	0	Normal	28	M	view: PA	PA
4	00436515-870c-4b36-a041-de91049b9ab4	264.0	152.0	213.0	379.0	1	Lung Opacity	32	F	view: AP	AP

Frequency Chart of our detailed class (Colored by Binary Class)

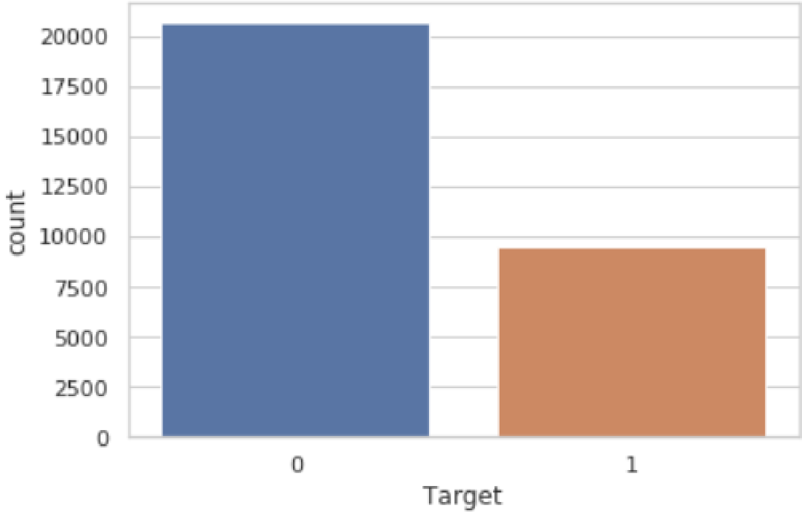
```
<Figure size 10368x864 with 0 Axes>
```




```
<Figure size 10368x864 with 0 Axes>
```

Frequency Chart of Binary Class (Colored by Binary Class)

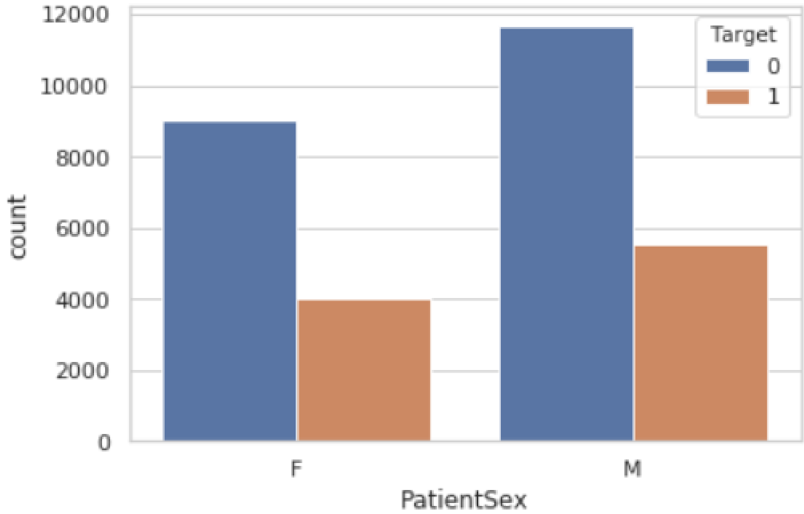
<Figure size 10368x864 with 0 Axes>



<Figure size 10368x864 with 0 Axes>

Frequency Chart of Sex (Colored by Binary Class)

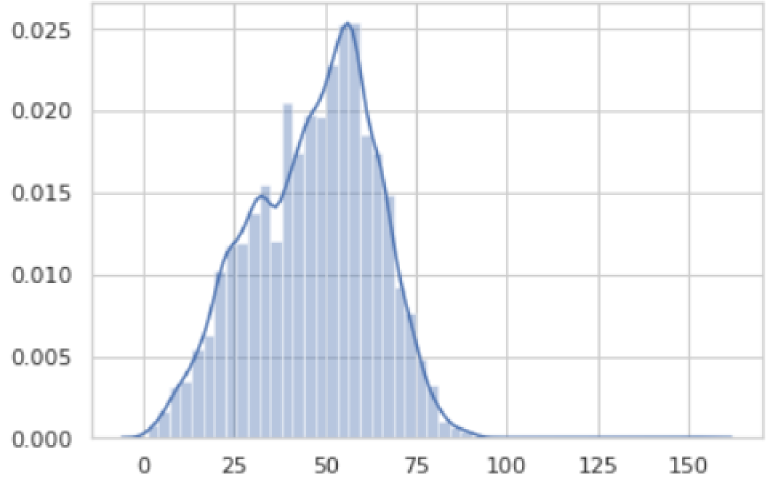
<Figure size 1800x360 with 0 Axes>



<Figure size 1800x360 with 0 Axes>

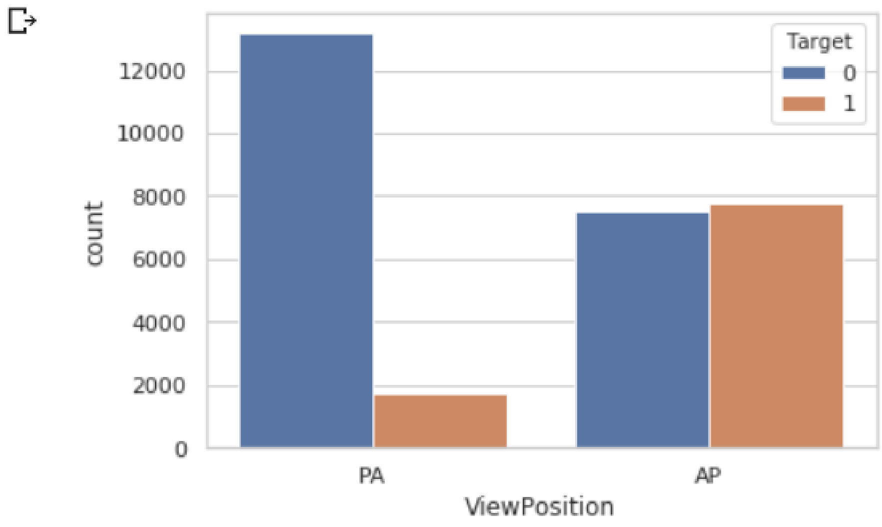
Distribution Plot of Patient Age

<Figure size 1800x360 with 0 Axes>

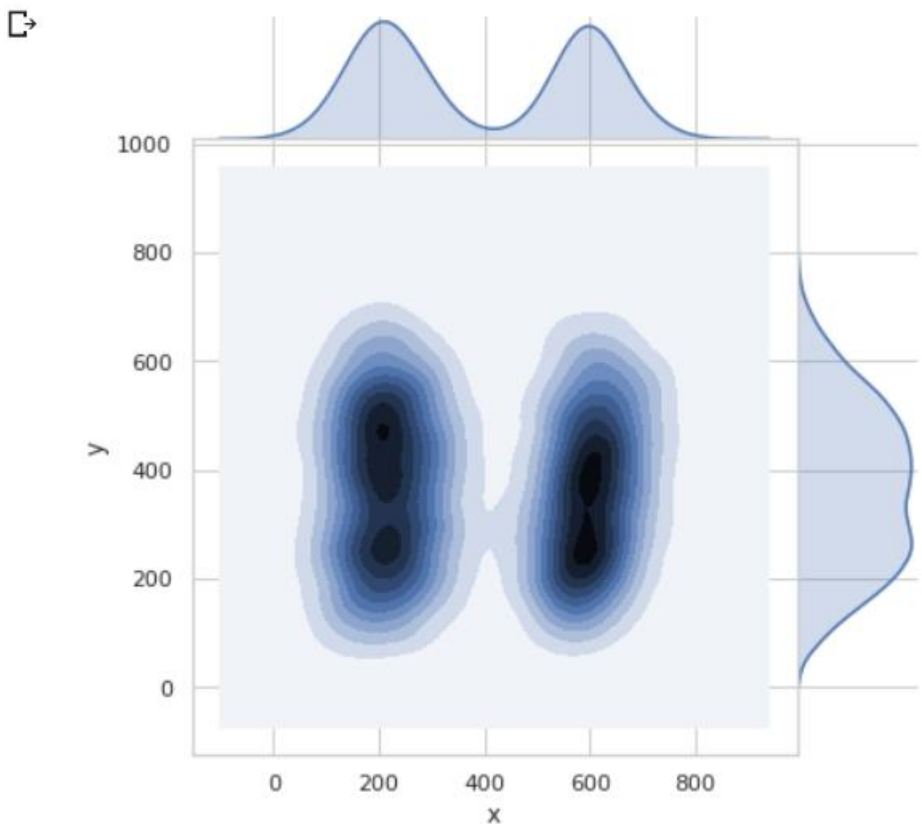


<Figure size 1800x360 with 0 Axes>

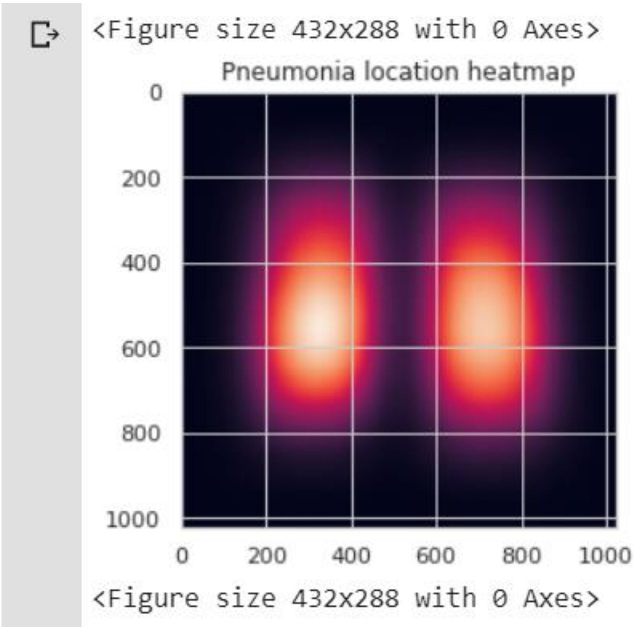
Clustered Column Chart based on viewing position



Heat map for x & y corners of each bounding box. But the below heatplot is imperfect.



So below will display Heat map of Pneumonia Presence in the sample image.



Run Models within GCP

Setting up Google Virtual Machine

Deploy the Models

The Approach Going Forward

Build prediction model based on following :

- Building a pneumonia detection model starting from basic CNN and then improving upon it.
- Train the model

- To deal with large training time, save the weights so that you can use them when training the model for the second time without starting from scratch.

- Test the model and report as per evaluation metrics

- Try different models (SSD, YoloV3,

- Set different hyper parameters, by trying different optimizers, loss functions, epochs, learning rate, batch size, check pointing, early stopping etc..for these models to finetune them

- Report evaluation metrics for these models along with your observation on how changing different hyper parameters leads to change in the final evaluation metric.