AWS MACHINE LEARNING ENGINEER NANODEGREE CAPSTONE REPORT

PNEUMONIA DETECTION USING DEEP LEARNING

- POOJA SINGARI

1.DEFINITION

1.1 PROJECT OVERVIEW

A large number of medical malpractice lawsuits as well as degradation of a patient's conditions attribute to misdiagnosis which can not only prove fatal to a patient's life but also to a medical practitioner's career. It is very essential to make sure that any medical condition is judged correctly by analyzing the prevalent symptoms for effective treatment of the patient.

The prevailing medical diseases call for the need of automated diagnosis which can be achieved with the help of Deep learning models based on computer vision achieved through supervised learning approach. Accurate and early detection of diseases such as Pneumonia can help prevent further fatal complications and may moreover lead to recovery at an early stage at the cost of fewer consultations and medications.

This machine learning model not only poses as a solution for achieving accurate diagnosis for the patients but can also help the doctors receive an unbiased second opinion.

1.2 PROBLEM STATEMENT

"The need for an accurate model for prediction of pneumonia amongst the rising Covid-19 Pneumonia cases"

Through it's natural approach Pneumonia is in its self a dangerous disease which when contracted affects one or more sections of the patient's lungs and may lead to several complications like lung abscesses and sepsis depending on the stage at which the medical treatment has been administered.

For this project the main focus will be on the usage of chest-xrays to diagnose the patient as it helps in checking both bacterial and viral pneumonic conditions.

With the recent spread of Covid-19 the world has also witnessed the rise of Covid-19 pneumonia cases. This is a far threatening condition wherein the

infection hijack’s the lung’s immune cells and damages certain vital organs including the kidney,brain,heart.
To help diagnose the condition an accurate machine learning model will be required to analyze as to whether or not a patient has contracted pneumonia in addition to being infected by Covid-19 virus. I propose to use a pre-trained convolutional neural network to achieve accurate results which is vital in the case proposed.

1.3 EVALUATION METRICS

As the project deals with the classification of two classes. I found accuracy to be the best suited evaluation metric. Outcomes of the model have been cross-checked with the ground truth of the image and the predicted label to validate the efficiency of the model. CrossEntropyLoss will be used as the loss function besides the Adam optimizer. Furthermore, using real-world images outside the test dataset has been an appropriate method to judge the competence of the model.

METRICS:

Accuracy: It is the fraction of predictions that the model predicted correctly

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

CrossEntropyLoss: It is a measure of the difference between two probability distributions for a given random variable or set of events.

$$H(p, q) \;=\; -\sum_i p_i \log q_i \;=\; -y \log \hat{y} - (1 - y) \log(1 - \hat{y})$$

2. ANALYSIS

2.1 DATA EXPLORATION

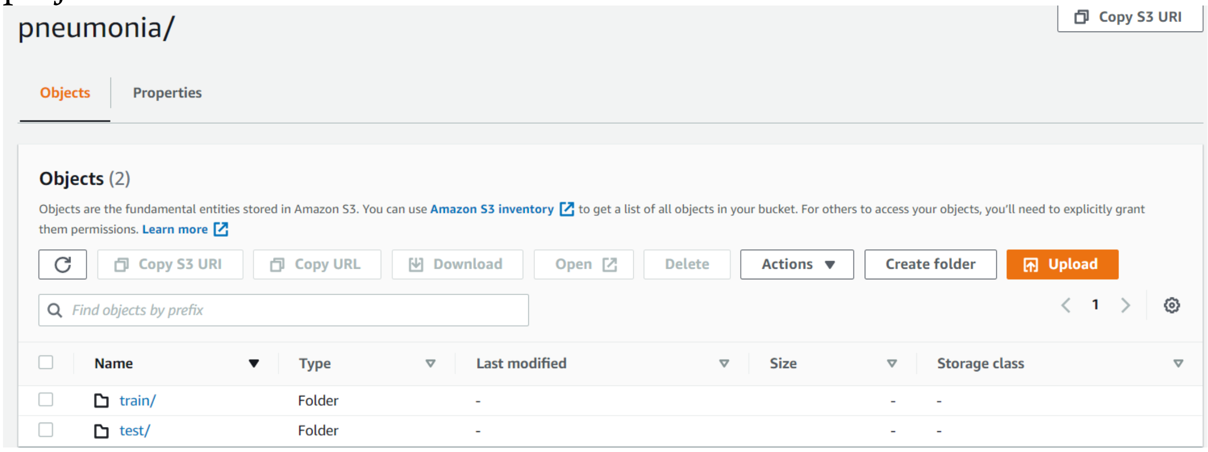
This project will focus on diagnosing Pneumonia from a Chest X-ray dataset. As the main causes of pneumonia are either bacterial, viral or mycoplasmic, for this reason Kaggle’s ‘Chest X-ray Images’ dataset will be used for this project as it’s three subsections(train, test, validation) collectively house 5,863 x-rays further labelled as normal/pneumonia.

Category	Number of Images
Train	5216
Test	624
Validation	16

Fig 1: Dataset

For the purpose of this project only train and test sections will be sufficient in modelling a classifier as the number of images in the validation section is comparatively trivial and will not have much effect on the model.

All the data had been downloaded and transferred into a s3 bucket for the project.



Objects (2)

Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

□	Name	Type	Last modified	Size	Storage class
□	train/	Folder	-	-	-
□	test/	Folder	-	-	-

Fig 2 :Dataset loaded into s3 buckets

2.2 DATA VISUALISATION

The labels were visualized via a bar graph. As depicted below the number of x-rays were more in number for Pneumonia than for normal images. Care should be taken that this factor does not lead us to building a model which is biased towards the pneumonia images.

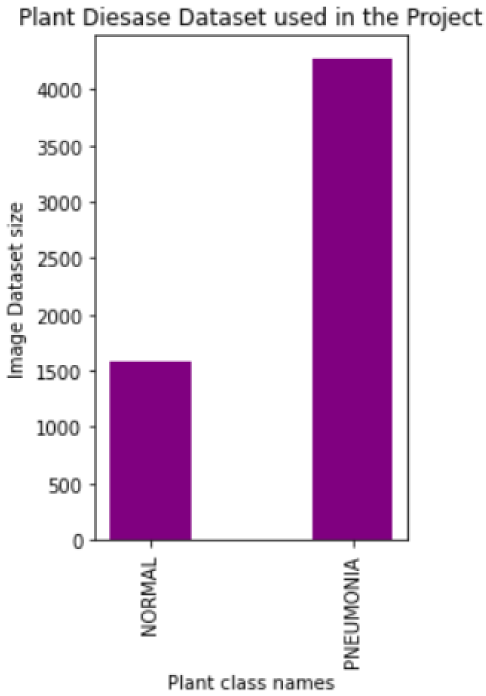


Fig 3: Bar plot of the classes

The images were further pre-processed by random rotations as augmentations are the simplest way of generalizing a dataset.

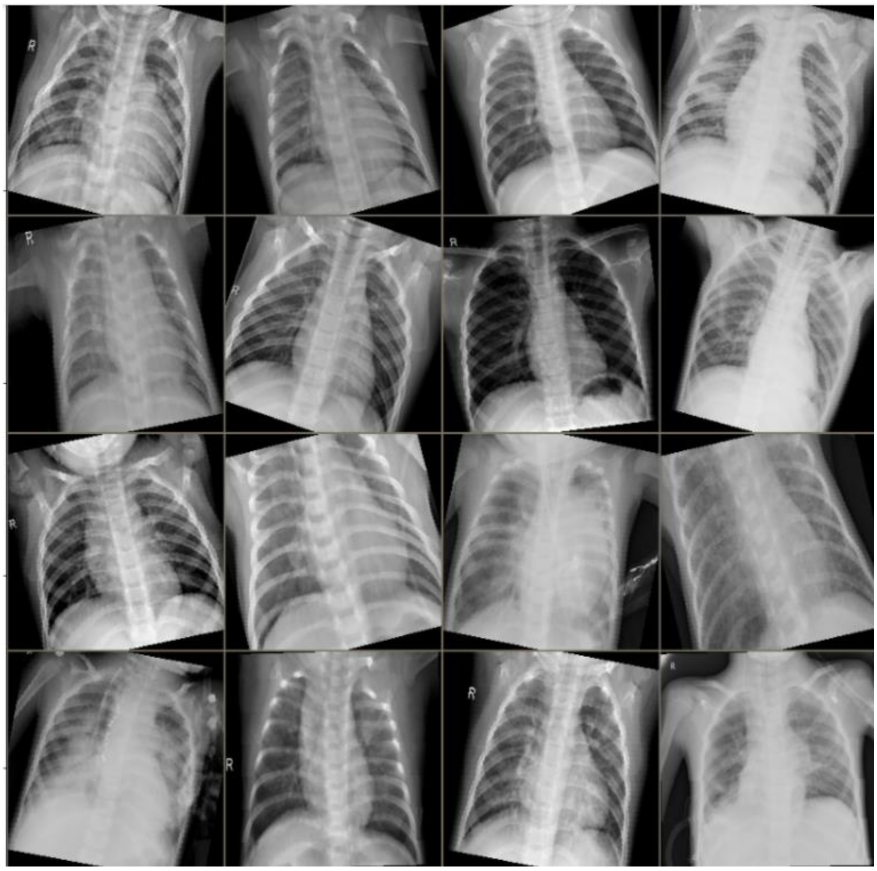


Fig 4: Augmented images

2.3 ALGORITHMS AND TECHNIQUES

The algorithm used is a pre-trained Convolutional Neural Network model for this project because it has already been trained on a larger dataset which makes the task of extracting necessary features from an image simpler as opposed to the exhaustion of computational resources while building a CNN model from scratch. The model opted for this project is EfficientNetB4 as it has already proven to give accurate results on a medical dataset which is a crucial requirement as the current project proposes to deliver an automatic pneumonia detection model.

The technique implemented is Transfer Learning wherein the features learned by an existing model are transferred to a newer model instead of training it's weights from the start. This not only saves time but also results in a more accurate model as the transferred model was trained on a large dataset and is already competent in extracting features from a smaller dataset as the one being used in this project.

2.4 BENCHMARK

Experiments have already been conducted to discover the best pre-trained model which gives the highest accuracy on medical datasets, specifically X-rays for automatic detection of Covid -19. Among the models, EfficientNetB4 achieved better performance with an accuracy of nearly 97%[1].
According to the paper[2], EfficientNet's scaling starts from a good baseline which rationale's it's success in predictive analysis and also contributes to its faster inference on the best existing ConvNet.Consequently, EfficientNetB4 will be used for this project as it is a good fit for the task.

3. METHODOLOGY

3.1 DATA PRE-PROCESSING

EfficientNet has been trained on coloured('rgb') dataset specifiacally the Imagenet dataset and expects their inputs to be tensors of pixels with values in the range of [0-255] range. For this reason all the data will be

resized into (224,224) so that essential features can be extracted for further processing.
The pre-processing includes visualizing the images in each subsection of the dataset to ensure no data leakage occurs. Techniques to be used include permuting the images and finding their mean and standard deviation after which the images will be clipped ,rotated ,resized, and squeezed into a proper normalized tensor to ensure that the model is given the right input.

The major challenge involved in using a x-rays as input to a pre-trained model is to tackle the inherent grayscale issue. X-rays are innately grayscale and have a one-channel input as opposed to coloured images which have a shape of [3,224,224] which in fact is the broadcasting shape of the model. All the images have to be transformed into 'rgb' either by using the transforms library repeating the input channel three times or by adding an initial layer to the pre-trained model to accept inputs having one channel. This project uses the former approach due to its simplistic operation and bound to succeed method.

3.2 TRANSFER LEARNING

A fully connected layer along with dropout and ReLU activation functions will be added to the pre-trained model. A simplified overview of the architecture is shown below in perspective of the design flow.

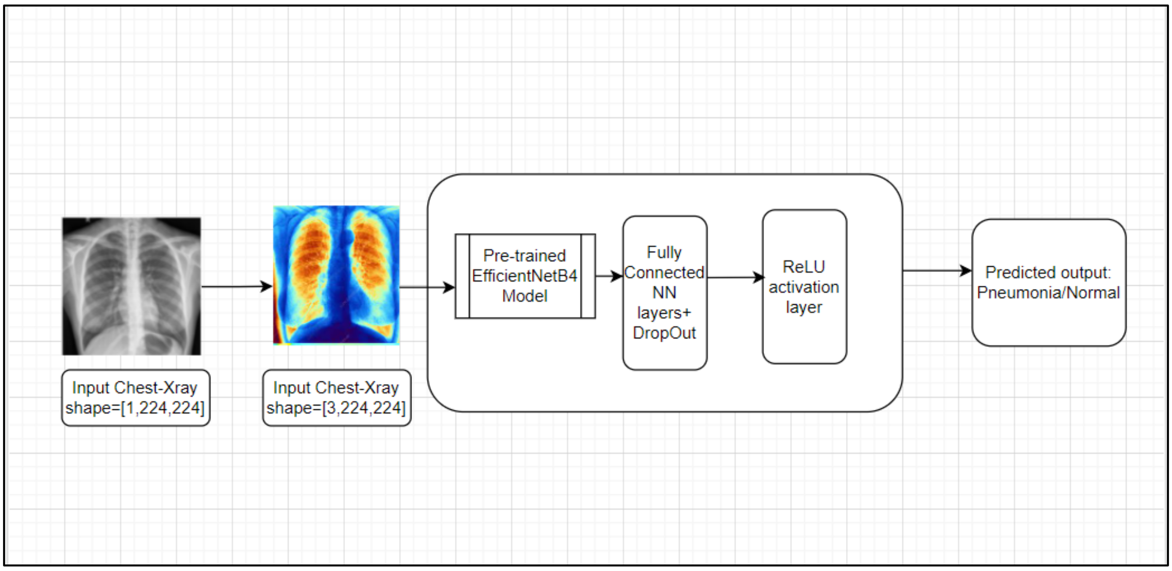
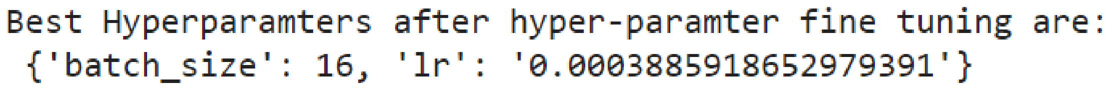

Input Chest-Xray
shape=[1,224,224]
Input Chest-Xray
shape=[3,224,224]
Pre-trained EfficientNetB4 Model
Fully Connected NN layers+ DropOut
ReLU activation layer
Predicted output: Pneumonia/Normal

Fig 5: Simplified design of fine-tuning EfficientNetB4 model

Finetuning can be achieved by trial and error of the various number or neurons to be included in each layer apart from the experimentation with the dropout and activation function. Tuner can be used to find the right learning rate and batch-size to achieve optimum results through several iterations of different combinational ranges.

3.3 FINETUNING THE PRE-TRAINED MODEL

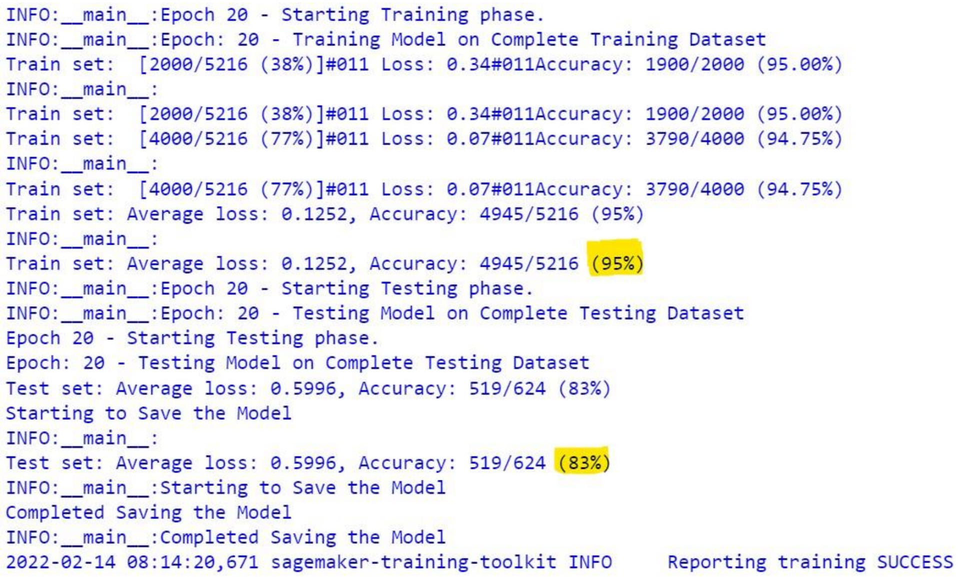
For the purpose of fine-tuning SageMaker's Tuner was utilized to find the best range for hyperparameters (namely the batch-size and learning rate) to achieve optimum results. After tuning the best configuration was found to be that of :



```
Best Hyperparamters after hyper-paramter fine tuning are:
 {'batch_size': 16, 'lr': '0.0003885918652979391'}
```

Fig 6: Snapshot of the best hyperparameters

After training the model with the above mentioned configuration for 20 epochs it was noticed that the model achieved an accuracy of **95%** in the training phase and **83%** over the training dataset



```
INFO:__main__:Epoch 20 - Starting Training phase.
INFO:__main__:Epoch: 20 - Training Model on Complete Training Dataset
Train set:  [2000/5216 (38%)]#011 Loss: 0.34#011Accuracy: 1900/2000 (95.00%)
INFO:__main__:
Train set:  [2000/5216 (38%)]#011 Loss: 0.34#011Accuracy: 1900/2000 (95.00%)
Train set:  [4000/5216 (77%)]#011 Loss: 0.07#011Accuracy: 3790/4000 (94.75%)
INFO:__main__:
Train set:  [4000/5216 (77%)]#011 Loss: 0.07#011Accuracy: 3790/4000 (94.75%)
Train set: Average loss: 0.1252, Accuracy: 4945/5216 (95%)
INFO:__main__:
Train set: Average loss: 0.1252, Accuracy: 4945/5216 (95%)
INFO:__main__:Epoch 20 - Starting Testing phase.
INFO:__main__:Epoch: 20 - Testing Model on Complete Testing Dataset
Epoch 20 - Starting Testing phase.
Epoch: 20 - Testing Model on Complete Testing Dataset
Test set: Average loss: 0.5996, Accuracy: 519/624 (83%)
Starting to Save the Model
INFO:__main__:
Test set: Average loss: 0.5996, Accuracy: 519/624 (83%)
INFO:__main__:Starting to Save the Model
Completed Saving the Model
INFO:__main__:Completed Saving the Model
2022-02-14 08:14:20,671 sagemaker-training-toolkit INFO     Reporting training SUCCESS
```

Fig 7: Snapshot of the model's accuracy

3.4 CHALLENGES FACED

- Ensuring that a proper instance type(ml.g4dn.xlarge) which is large enough to satisfy the requirements of the task which otherwise leads to an error while downloading essential libraries

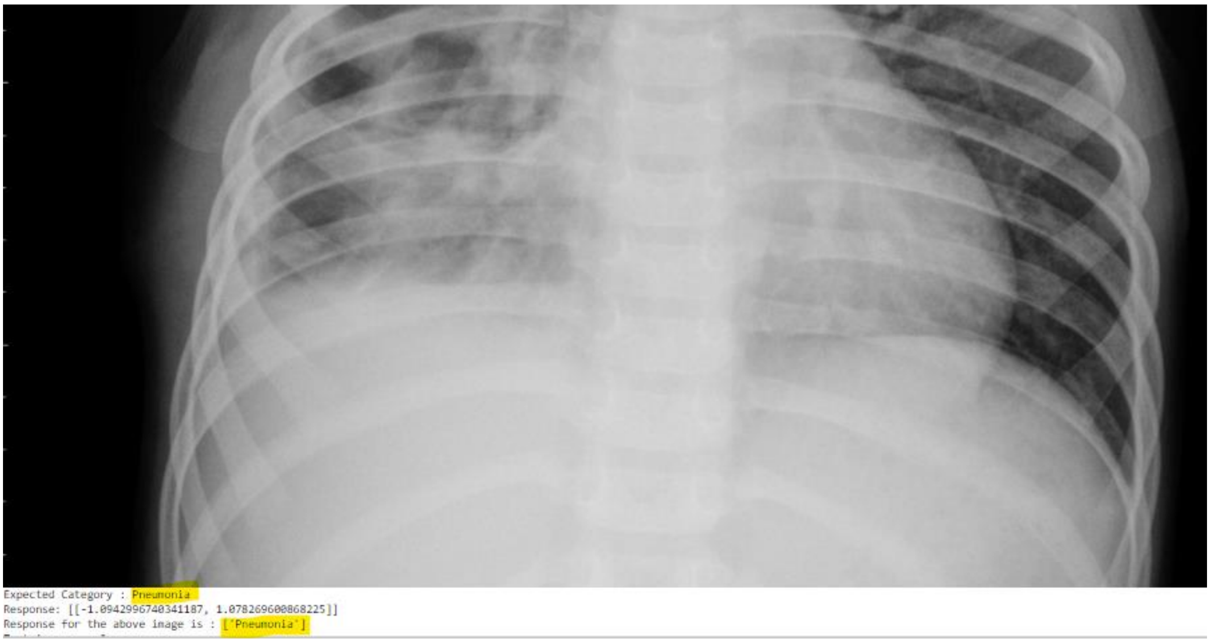
- Updating the versions of the environment variables (SageMaker,Pytorch) without which pre-trained models and training jobs will not function
- Converting grayscale images into colored images which could otherwise lead to a broadcasting error as a pre-trained CNN model requires input to have 3 channels(rgb).Hence transformation is applied on each image before serving it to the model for prediction.

4. RESULTS

4.1 INFERENCE

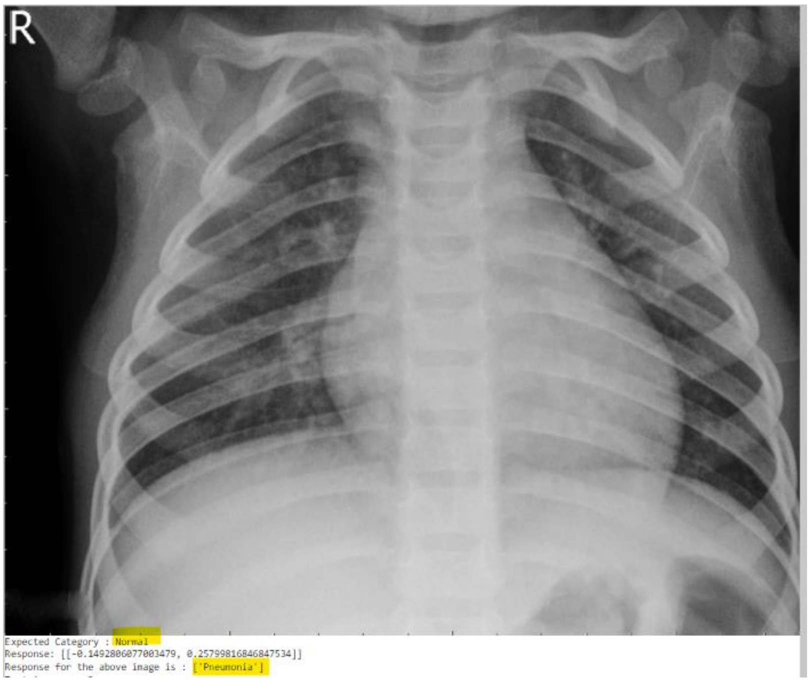
Post-training the model was presented with 3 images from the validation set(which was initially dopped out) to check the accuracy of the model.These 3 images pose to be images out of the dataset as the model has not seen them before. The model correctly identifies the classes of all the three images correctly as required.

TEST :1



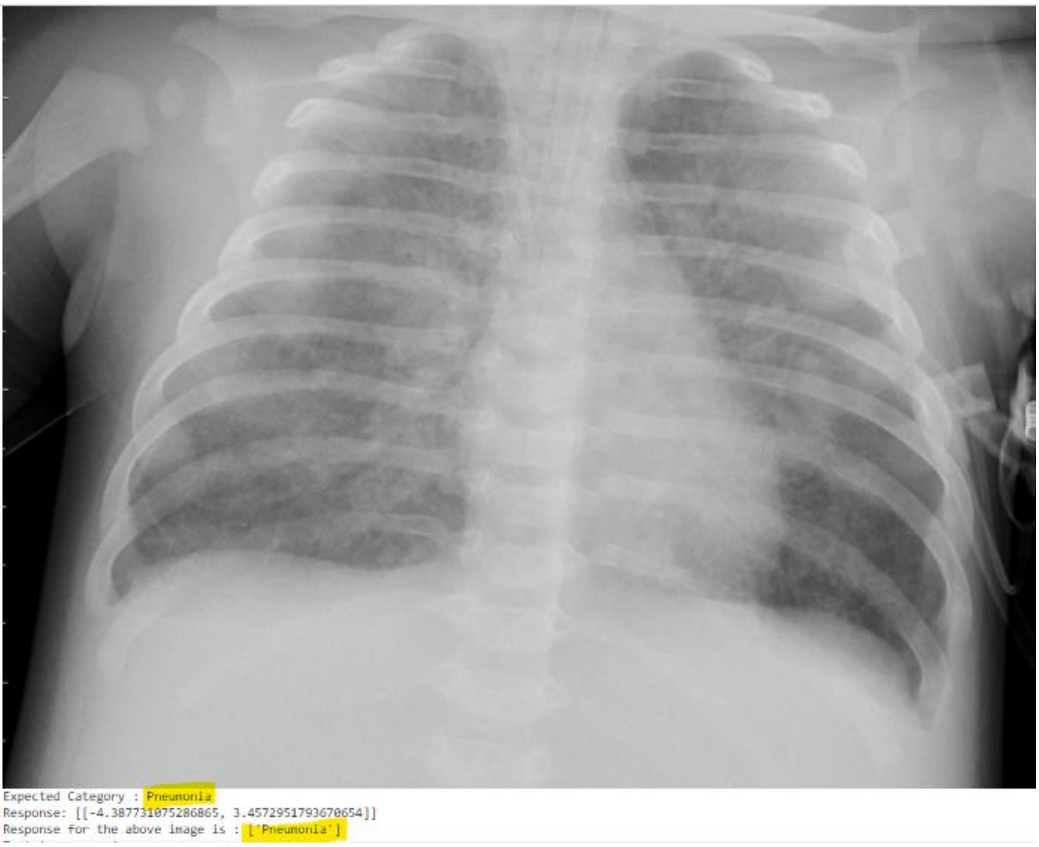
```
Expected Category : Pneumonia
Response: [[-1.0942996740341187, 1.078269600868225]]
Response for the above image is : ['Pneumonia']
```

TEST:2



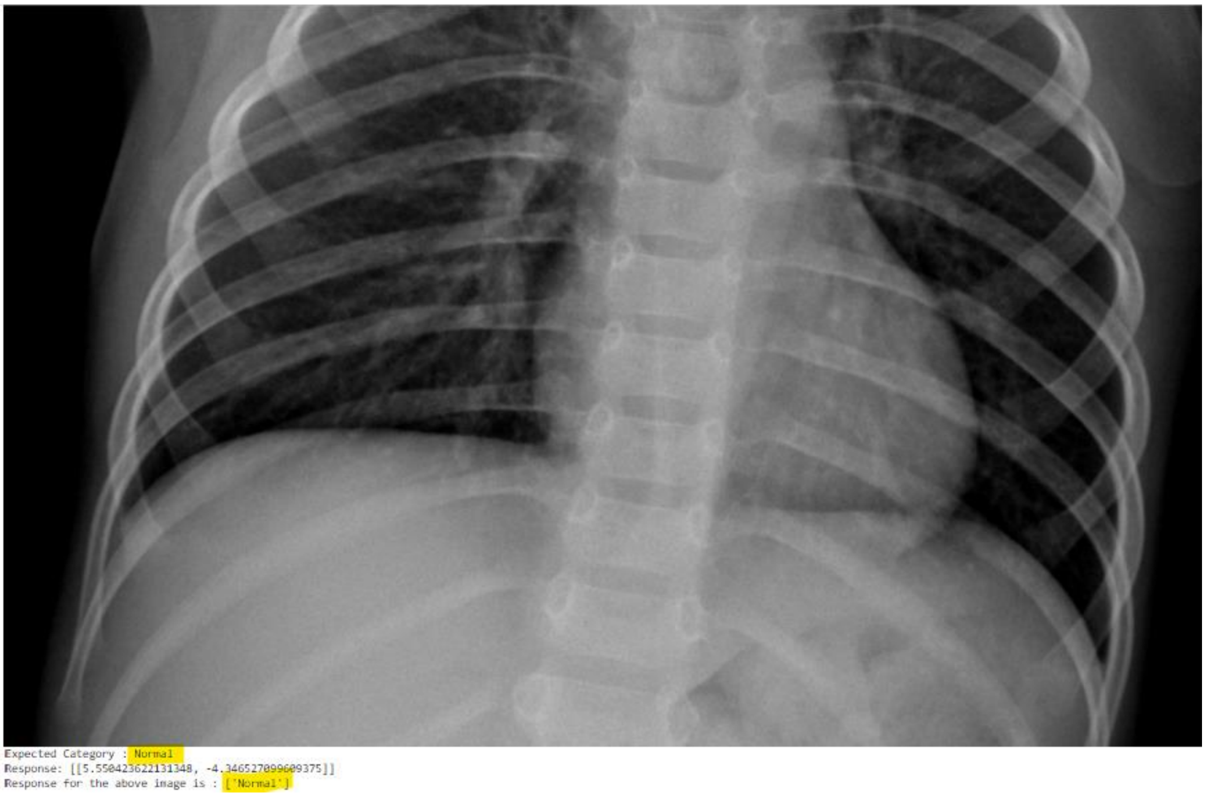
```
Expected Category : Normal
Response: [[-0.1492806077003479, 0.2579981684684753 4]]
Response for the above image is : ['Pneumonia']
```

TEST:3



```
Expected Category : Pneumonia
Response: [[-4.387731075286865, 3.4572951793670654]]
Response for the above image is : ['Pneumonia']
```

TEST:4



```
Expected Category : Normal
Response: [[5.550423622131348, -4.346527099609375]]
Response for the above image is : ['Normal']
```

4.2 JUSTIFICATION

As seen above the model was able to correctly identify both the pneumonic x-rays whereas it misclassifies a normal x-ray as Pneumonia(TEST 2). This indicates that the model has still not achieved the state-of-art accuracy which can be attributed to the imbalance in the dataset used.
Further work can be done to annotate and add more normal x-rays to counteract the effect of pneumonic x-rays on the model so that it doesn’t result in false positives.