FDA Submission

Padma Chandramouli
Chest X Ray Analyzer for detecting Pneumonia

Algorithm Description

1. General Information

Intended to assist Radiologist in identifying if x ray imaging indicates Pneumonia condition.It is not intended for use in supporting life or sustaining life.

Indications for Use: Assist Radiologist in screening Pneumonia between the ages 35-65 year old patients

Device Limitations: Difficult to detect Pneumonia when the patient has multiple conditions.

Clinical Impact of Performance:

From a clinical perspective,we need to identify as many Positive cases of Pneumonia as possible. If we misclassify certain negative cases as positives(False Positives are a bit higher) it should be acceptable because we can perform further tests to confirm Pneumonia. However we should not miss Positive cases(least False Negatives) Hence Recall plays a critical role and becomes most important measure of performance in clinical scenario for identifying Pneumonia.

2. Algorithm Design and Function

Algorithm Flowchart: Architecture: Architecture uses transfer learning to use the VGG16 model with the last Maxpool layer replaced with 4 Fully connected layers. Each of these 4 layers have a Relu activation function and are followed by a Dropout layer to reduce overfitting. The 5th layer has an activation Sigmoid to determine probability of Pneumonia.

Architecture of VGG model:

Model: "vgg16"

Layer (type)	Output Shape	Param #
input_4 (InputLayer)	(None, 224, 224, 3)	0
block1_conv1 (Conv2D)	(None, 224, 224, 64)	1792
block1_conv2 (Conv2D)	(None, 224, 224, 64)	36928
block1_pool (MaxPooling2D)	(None, 112, 112, 64)	0

block2_conv1 (Conv2D)	(None, 112, 112, 128)	73856
block2_conv2 (Conv2D)	(None, 112, 112, 128)	147584
block2_pool (MaxPooling2D)	(None, 56, 56, 128)	0
block3_conv1 (Conv2D)	(None, 56, 56, 256)	295168
block3_conv2 (Conv2D)	(None, 56, 56, 256)	590080
block3_conv3 (Conv2D)	(None, 56, 56, 256)	590080
block3_pool (MaxPooling2D)	(None, 28, 28, 256)	0
block4_conv1 (Conv2D)	(None, 28, 28, 512)	1180160
block4_conv2 (Conv2D)	(None, 28, 28, 512)	2359808
block4_conv3 (Conv2D)	(None, 28, 28, 512)	2359808
block4_pool (MaxPooling2D)	(None, 14, 14, 512)	0
block5_conv1 (Conv2D)	(None, 14, 14, 512)	2359808
block5_conv2 (Conv2D)	(None, 14, 14, 512)	2359808
block5_conv3 (Conv2D)	(None, 14, 14, 512)	2359808

Architecture of Classifier:

Layer (type)	Output Shape	Param #
model (Functional)	(None, 7, 7, 512)	14714688
flatten (Flatten)	(None, 25088)	0
dropout (Dropout)	(None, 25088)	0
dense (Dense)	(None, 2048)	51382272
dropout_1 (Dropout)	(None, 2048)	0

dense_1 (Dense)	(None, 1024)	2098176
dropout_2 (Dropout)	(None, 1024)	0
dense_2 (Dense)	(None, 512)	524800
dropout_3 (Dropout)	(None, 512)	0
dense_3 (Dense)	(None, 256)	131328
dropout_4 (Dropout)	(None, 256)	0
dense_4 (Dense)	(None, 1)	257

Total params: 68,851,521
Trainable params: 54,136,833
Non-trainable params: 14,714,688

DICOM Checking Steps: Run inference on only valid Dicom images for Pneumonia i.e. we check if xray images are of chest, the position the image was taken i.e.'AP'or 'PA' and modality being DX

Exploratory Data Analysis:
1) Data Distribution based on Patient Demography - Age,Gender: We have analysed that data based on Gender and Age these are the only details that have been provided to us. As seen in graphs above, the majority of Pneumonia cases we found are between 20-75 years. There are more Male count (830) than Female count(550) but the difference is not stark. We do not need to split the training data based on gender or Age inorder to have a balanced training data. However we should mention that the intended use of the algorithm is restricted to patients of age between 20-75 Yrs.

2) The x-ray views taken - AP,PA - We analysed that AP has higher counts(800) then PA(600) however again we do not need to split the training data to create a balanced data based on views.

3) The number of cases of Pneumonia: There is a stark difference between Pneumonia and non Pneumonia cases.There are 1431 Pneumonia Cases. Whereas there are 110689 Non Pneumonia Cases. So we do need to split the training data in order to have Non Pneumonia cases close to Pneumonia cases in order to have a balanced data set for training.

4) Distribution of other diseases that are comorbid with pneumonia : The most common disease that co-exists with Pneumonia is Infiltration.We can assume that this will be a challenge if we

have to identify Infiltration condition from Pneumonia cases, however that's not our current use case.

EDA Conclusion:
Number of Pneumonia vs Non-Pneumonia needs to be taken into account while splitting training data. The need to be equal in order to create a balanced training dataset.
'Infiltration','Edema','Effusion' and 'Atelectasis' are pretty close in intensity level to pneumonia.We can include these diseases that are comorbid as well.

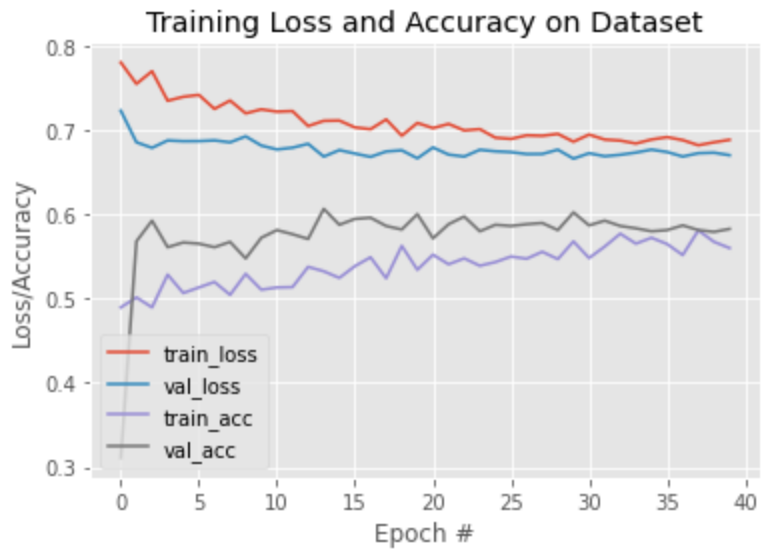
Preprocessing Steps for training set:
All images where Resized to (224,224) so it could be accepted as input to the VGG model and normalized

Preprocessing Steps for validation set: No Image augmentation was done in validation set, the images were however normalized and resized to(224,224).

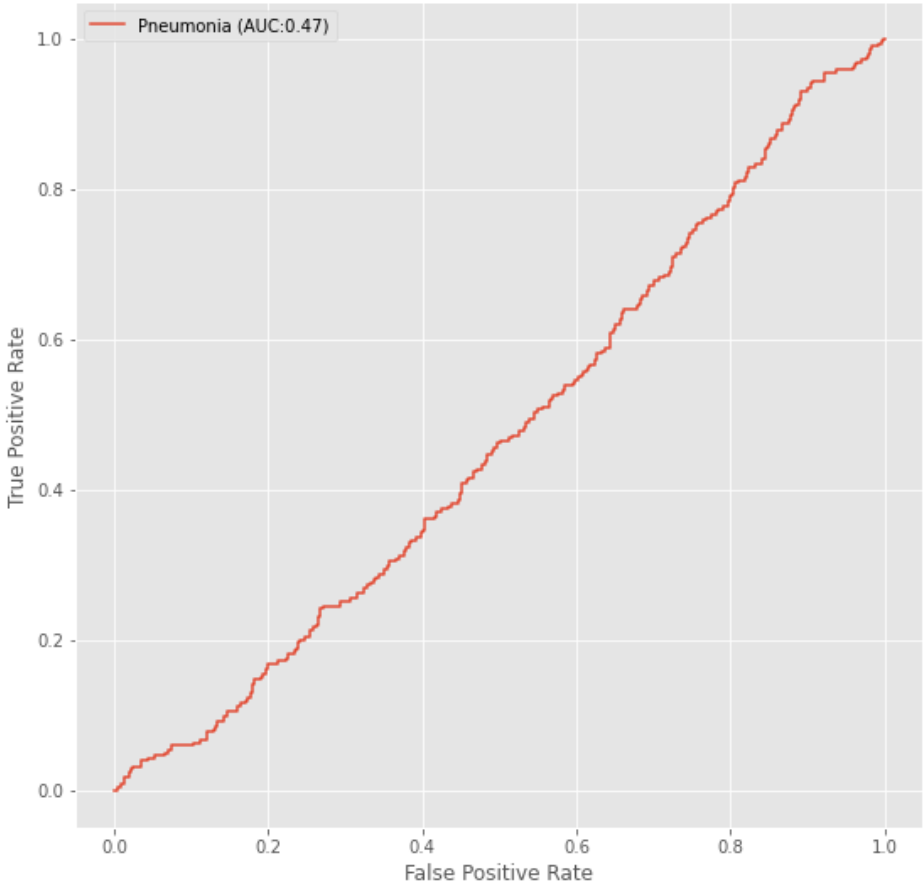
3. Algorithm Training

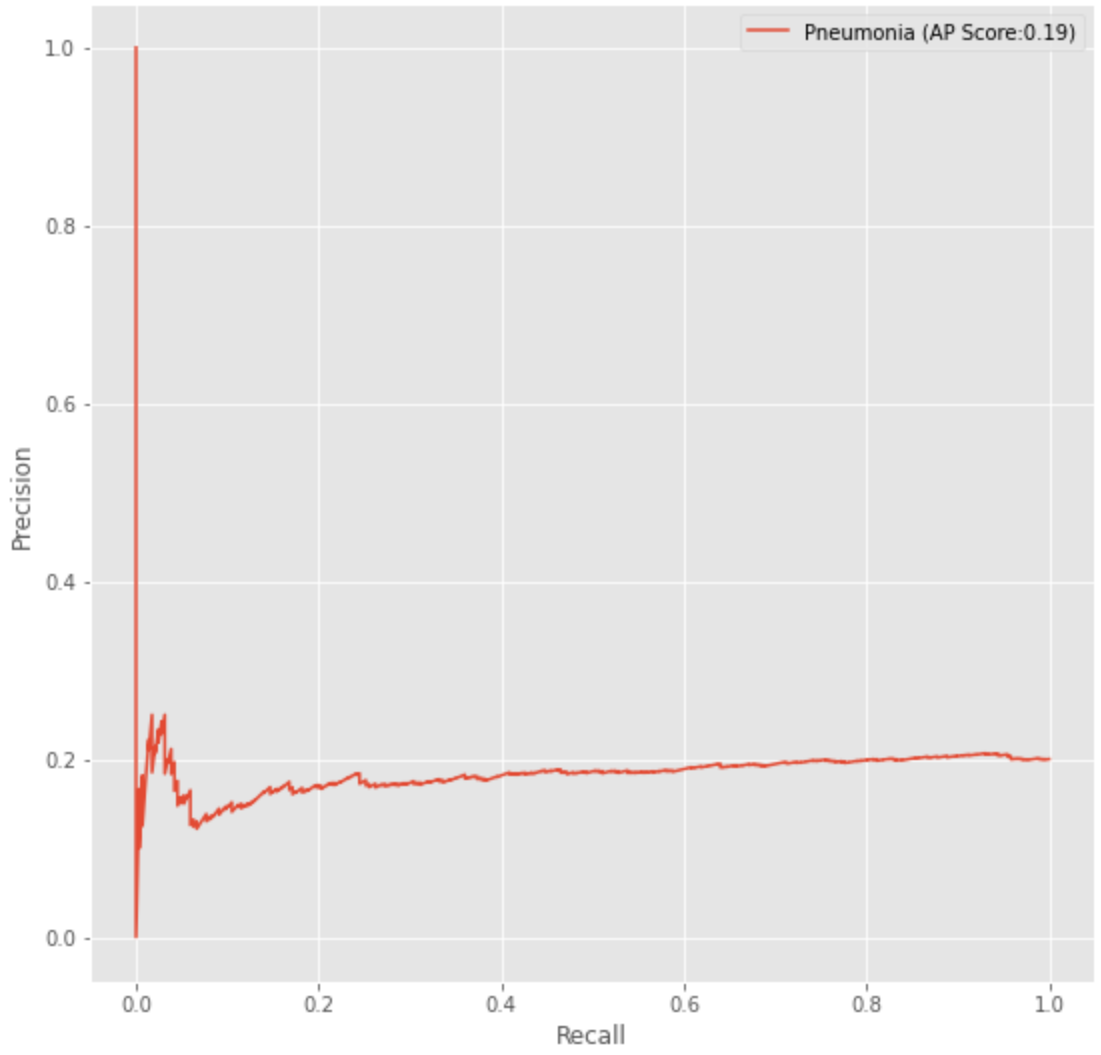
Parameters:
Types of image augmentation used for training dataset : Image Augmentation was done on training data set as follows:Normalize the image,random image data sets where horizontally flipped,used a bit of shearing, shifted height and width a bit by 0.01 and rotated the image by an angle of 1.0 . Also zooming by a factor of 0.01.

Batch size =32
Optimizer learning rate : 1e-6. the low learning rate provided smoother validation and training loss curves. Batch size of 32 was used for training and a batch size of 100 for validation dataset.
Layers of pre-existing architecture that were frozen: block1_conv1 to block5_conv3
Layers of pre-existing architecture that were fine-tuned:block5_pool
Layers added to pre-existing architecture: 5 fully connected layers.



plot_roc_curve() and plot_precision_recall_curve() from Build and Train Model notebook





Final Threshold and Explanation:
the average precision is 0.2 ,high False Positives,this means a huge number of non-Pneumonia pateints are being diagnized as having Pneumonia.The recall being 0.64 could be improved if this needs to be used in clinical scenario.With Recall being 0.64 we know that good number of positive cases are being identified. I will conclude that my algorithm cannot be used in clinical scenario.We need to work on improving Recall to atleast 90%+ and precision to 80% . If patients with no pneumonia are being misdiagnized with Pneumonia because we can still run additional tests to confirm patient's condition.Hence I am not concerned about improving Precision a lot.

4. Databases
(For the below, include visualizations as they are useful and relevant)

Description of Training Dataset:Split data to 20% validation set and 80% Training set. We also create a balanced training set containing the same number of Pneumonia and non Pneumonia cases.

Description of Validation Dataset: Validation dataset should not be augmented. However they need to be resized to match the input dimensions and normalize them. There is no need to create Balanced data set.However we used Non Neumonia cases as 4 times the count of Pneumonia since there is a huge number of dataset for non Pneumonia cases.

5. Ground Truth

"pneumonia_class" can be used as ground truth for the algorithm. However in reality we need to rely on existing device/algorithm to compare performance with or use silver standard.

6. FDA Validation Plan

Patient Population Description for FDA Validation Dataset:
In reality we need equal Dicom records of various patients in the age group of 0-12, 13-30,31-50, 51-70,71 and above.

We need equal distribution of images based on positions the xray was taken.
We need all the xray images to of chest and modality DX.
We need equal xray images distribution based on gender and race.

Ground Truth Acquisition Methodology: In reality we need to collaborate with hospitals/clinics to get data and make sure that we have permission to use patient personal information or rather not use data that would identify patient and violate HIPAA . We cannot get a GOLD standard however we can get a silver standard by running the images through 2 or more radiologists and taking their weighted feedback.

Algorithm Performance Standard: The ideal performance would be 80% Precision and 99.99% Recall. Recall is the most important metric because we need to identify all the positive cases i.e. no False Negatives. Lower precision is fine even if a patient is misdiagnosed with Pneumonia further tests can be conducted to ascertain the condition.