FDA Submission

Your Name: Ganapathy Shankar

Name of your Device: PneumoniaX-ray

Algorithm Description

1. General Information

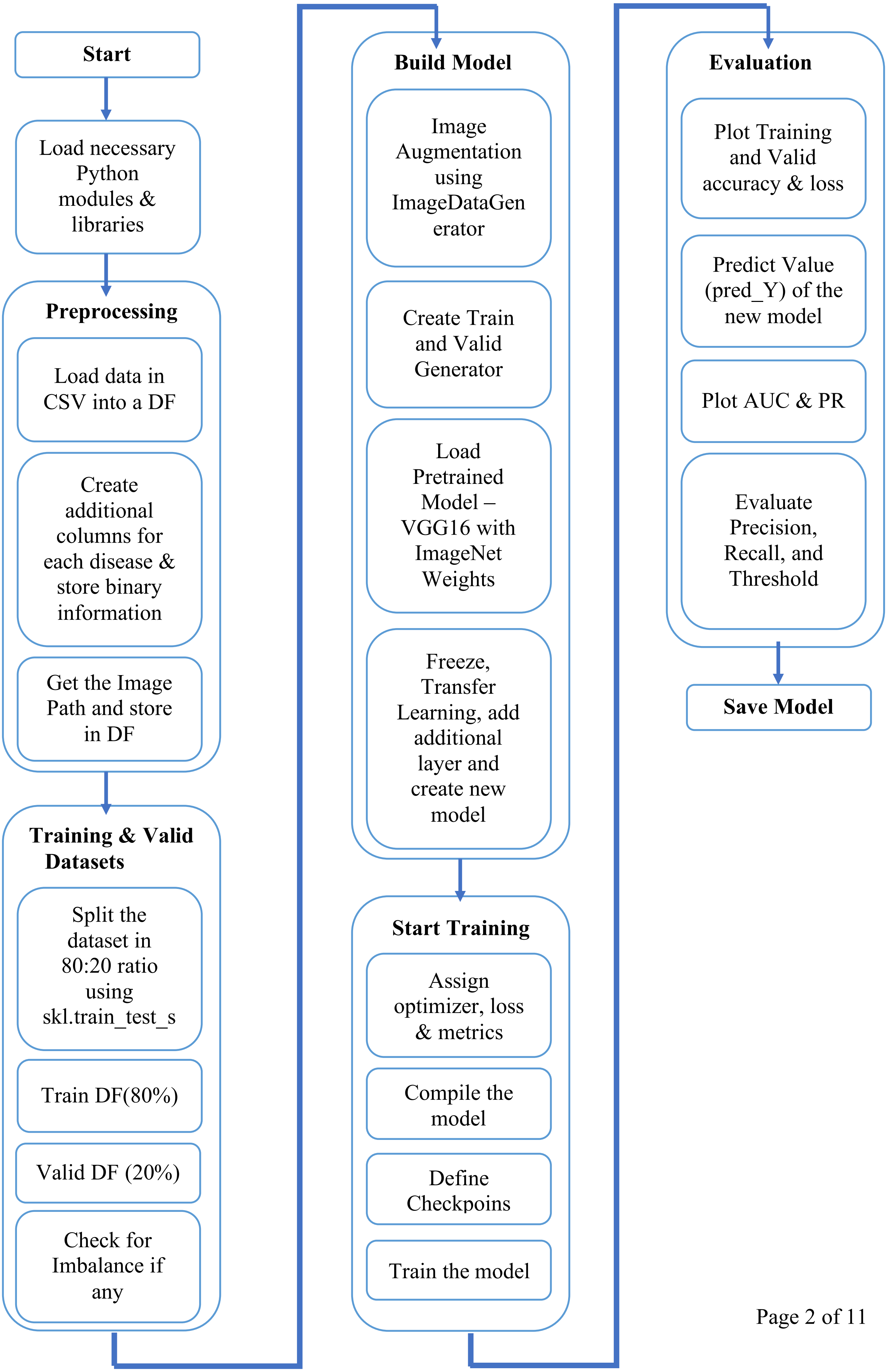
Intended Use Statement: The PneumoniaX-ray software is a device that allows review, analysis, and interpretation of chest X-ray images. It is intended for use with chest X-ray images to assist medical professionals in image analysis. It is not intended to be the primary interpretation. The software provides classification of presence of Pneumonia, numerical analysis and substance indication. The user can review, verify and correct the results of the system and generate a report of the findings.

Indications for Use: The PneumoniaX-ray is a software designed to aid the clinical assessment of new born to adults chest X-ray cases with featured suggestive of pneumonia in the medical environment. PneumoniaX-ray analyzes cases using and artificial algorithm to identify suspected findings of pneumonia. It makes case level output available in a PACS system for worklist prioritization. PneumoniaX-ray is not intended to direct attention to specific portion or anomalies of an image. Its results are not intended to be used on a standalone basis for clinical decision making nor is it intended to rule out pneumonia or otherwise preclude clinical assessment of X-ray cases.

Device Limitations: PneumoniaX-ray software is designed for analysis of X-ray images captured in 2-D format only and does not support any other form of medical images like CT or MRI scan images etc. PneumoniaX-ray performs a binary classification of prevalence of Pneumonia in a given X-ray and does not perform segmentation or localisation of Pneumonia. The software is designed by training data of a new born baby to adults up to 100 years of age of both male and female.

Clinical Impact of Performance: PneumoniaX-ray assist medical professionals in featured suggestive of pneumonia in the medical environment there by reducing manual effort and human error in analyzing a X-ray image but does not replace the activities of a physician. A prediction of false positive or false negative may affect the treatment plan for a patient, hence it is recommended to have a strong review mechanism by a qualified physician for every case.

2. Algorithm Design and Function


Start
Load necessary Python modules & libraries
Preprocessing
Load data in CSV into a DF
Create additional columns for each disease & store binary information
Get the Image Path and store in DF
Training & Valid Datasets
Split the dataset in 80:20 ratio using skl.train_test_s
Train DF(80%)
Valid DF (20%)
Check for Imbalance if any
Build Model
Image Augmentation using ImageDataGenerator
Create Train and Valid Generator
Load Pretrained Model – VGG16 with ImageNet Weights
Freeze, Transfer Learning, add additional layer and create new model
Start Training
Assign optimizer, loss & metrics
Compile the model
Define Checkpoins
Train the model
Evaluation
Plot Training and Valid accuracy & loss
Predict Value (pred_Y) of the new model
Plot AUC & PR
Evaluate Precision, Recall, and Threshold
Save Model

DICOM Checking Steps:

DICOM header provide all the attributes except for the pixel data of the image and the image file provide pixel data representing actual image. The first steps is to pre-extract key attributes from DICOM headers into a dataframe using Python package pyidcom to optimize image processing workflow and algorithm training. Since PneumoniaX-ray is specifically designed for deducting pneumonia in 2D chest X-ray, we ensure to check image position is Posterior-Anterior (PA) or Antero-Posterior(AP), image type is Digital radiography (DX) , and body part is of Chest in the given image. Any image not meeting the criteria is not processed by the software.

Pre-processing Steps:

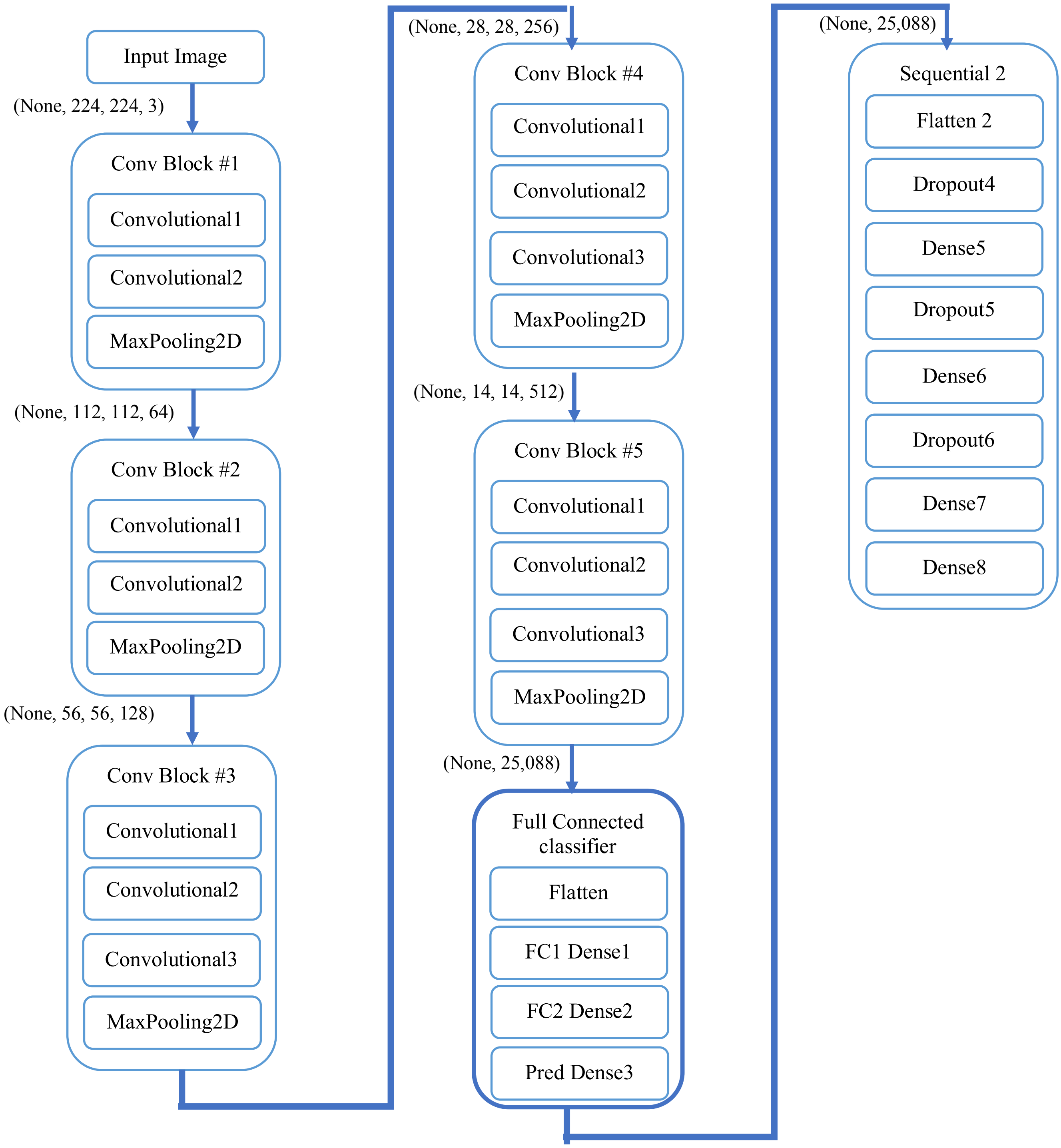
The NIH dataset is available in a csv file "Data_Entry_2017.csv". Following steps are followed for pre-processing

1) Load the csv file data in a dataframe using Python Pandas library

2) 'Finding Labels' column stores diseases findings from the X-ray image. In many cases more than one disease is prevalent in a X-ray image and each diseases is separated by '|' in Finding Labels column. As a good additional columns for each diseases (14 types in this datasets) ois created and have a binary representation for presence and no presence of pneumonia for easy data manipulation and analysis

3) X-ray image files are stored in a separate directory. Use 'os' and 'glob' modules to read the image filepath and stored in a column against each images for easy manipulation

CNN Architecture: Pre trained VGG16 architecture with weights trained on the ImageNet dataset with added dropouts and dense layer



3. Algorithm Training

Parameters:

For PneumoniaX-ray algorithm we have used a pretrained model called “VGG16” and fined tuned the model based on the dataset. Below is the parameters for “VGG16”

```
Model: "vgg16"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_1 (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
_________________________________________________________________
block5_pool (MaxPooling2D)   (None, 7, 7, 512)         0
_________________________________________________________________
flatten (Flatten)            (None, 25088)             0
_________________________________________________________________
fc1 (Dense)                  (None, 4096)              102764544
_________________________________________________________________
fc2 (Dense)                  (None, 4096)              16781312
_________________________________________________________________
predictions (Dense)          (None, 1000)              4097000
=================================================================
Total params: 138,357,544
Trainable params: 138,357,544
Non-trainable params: 0
_________________________________________________________________
```

```
Model: "sequential_2"
_________________________________________________________________
Layer (type)                 Output Shape              Param #   
=================================================================
model_1 (Model)              (None, 7, 7, 512)         14714688  
_________________________________________________________________
flatten_2 (Flatten)          (None, 25088)             0         
_________________________________________________________________
dropout_4 (Dropout)          (None, 25088)             0         
_________________________________________________________________
dense_5 (Dense)              (None, 1024)              25691136  
_________________________________________________________________
dropout_5 (Dropout)          (None, 1024)              0         
_________________________________________________________________
dense_6 (Dense)              (None, 512)               524800    
_________________________________________________________________
dropout_6 (Dropout)          (None, 512)               0         
_________________________________________________________________
dense_7 (Dense)              (None, 256)               131328    
_________________________________________________________________
dense_8 (Dense)              (None, 1)                 257       
=================================================================
Total params: 41,062,209
Trainable params: 28,707,329
Non-trainable params: 12,354,880
_________________________________________________________________
```

Types of augmentation used during training : All the images are augmented to have uniformity across all the images in the training set. Hence the images are rescaled to 1./255 - every pixel value from range [0,255] -> [0,1]. Flipping the image horizontally is set to true and vertical flip is set the false.

Batch size : Batch Size of 64 is used

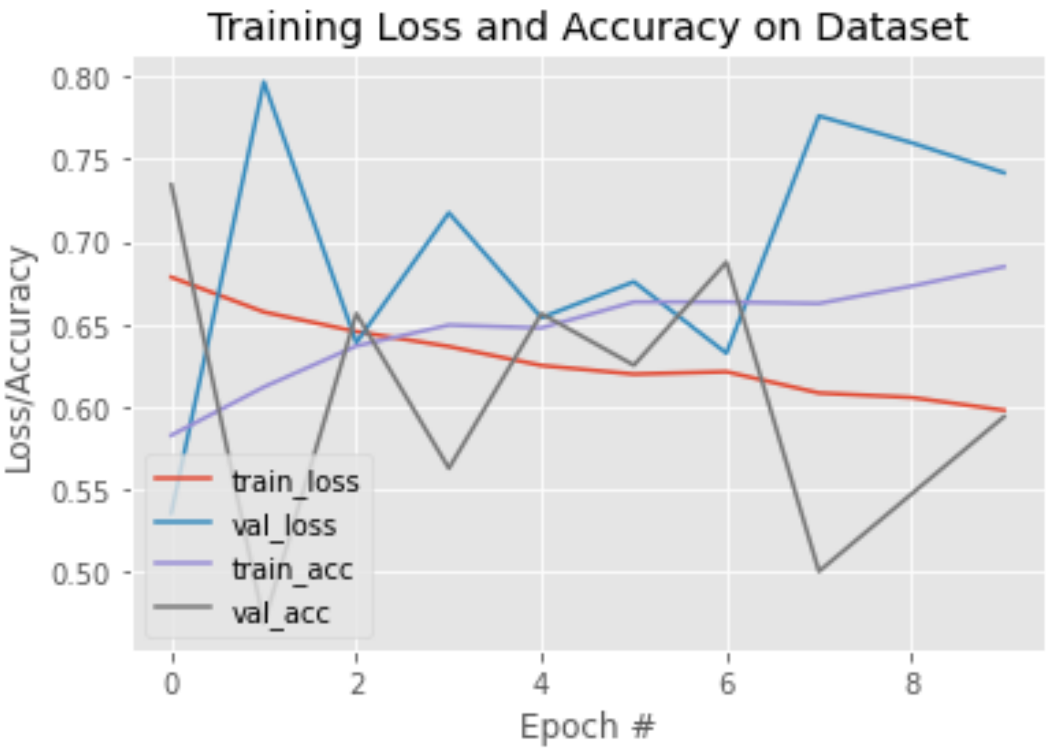
Optimizer learning rate : In PneumoniaX-ray ‘Adam’ optimizer is used. Adam optimization is a stochastic gradient descent method that is based on adaptive estimation of first-order and second-order moments. Learning rate of 1e-4(0.0001) is used

Layers of pre-existing architecture that were frozen : Please refer to the flow chart in CNN Architecture. VGG16 architecture with weights trained on the ImageNet dataset, Conv Block #1 to #5 containing is frozen

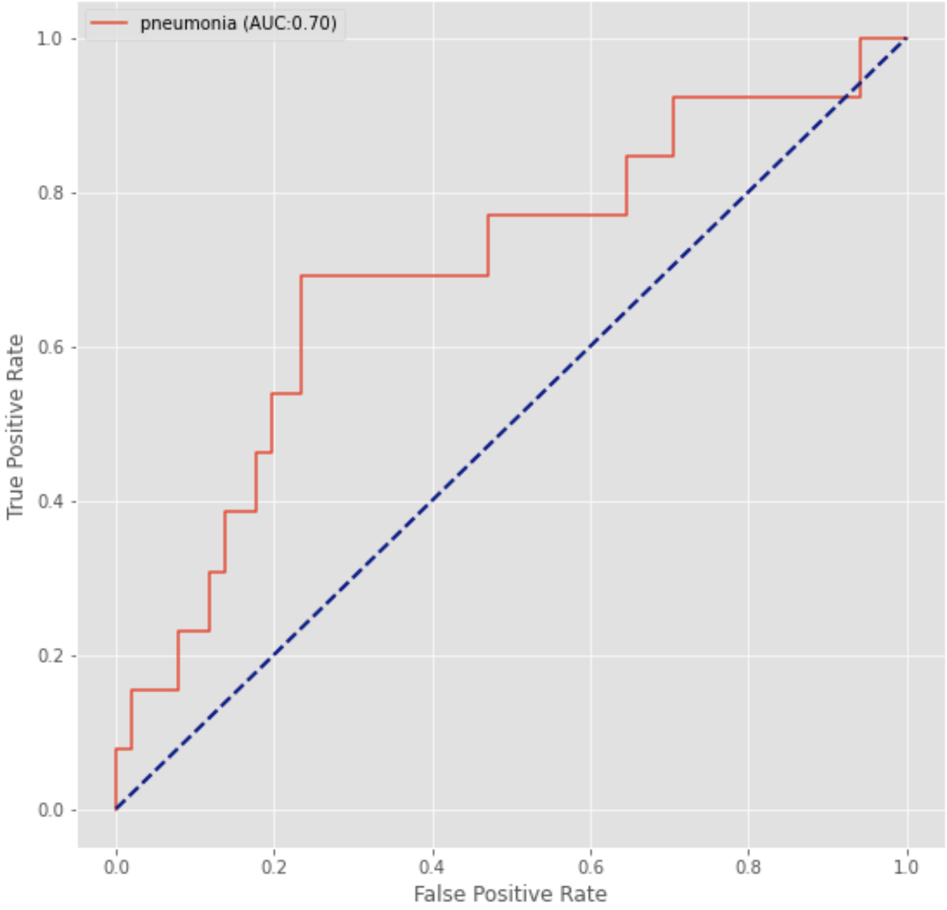
Layers of pre-existing architecture that were fine-tuned : Loaded the pretrained weights and train the complete network with a smaller learning rate of 1e-4(0.0001). This results in very good accuracy with even small datasets

Layers added to pre-existing architecture : Added a classifier on top of the convolutional base by adding a fully connected layer followed by a softmax layer with 4 outputs.

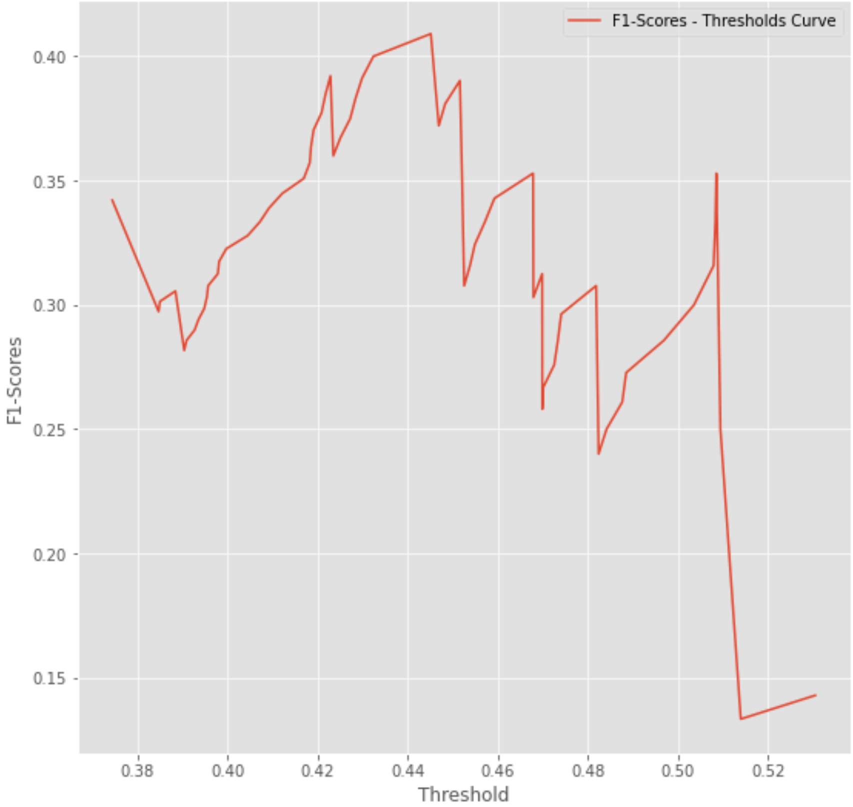
Insert algorithm training performance visualization :



Insert P-R curve



F1-Scores – Thresholds Curve



Final Threshold and Explanation: Final Threshold of 0.62786263 at 80% Precision is considered with F1-score of 0.14285. Test accuracy is at 67.18%, AUC 0.64 and AP score of 0.3. Looking at the above graph threshold of 0.445 at F1-Score of 0.42 seems to be optimum, however given the 6 test images, where in predicted threshold is 0.4775539, final threshold of 0.62786263 gives accurate result on test images.

4. Databases

The dataset was curated by NIH specifically to address the problem of a lack of large x-ray datasets with ground truth labels to be used in the creation of disease detection algorithms. There are 112,120 X-ray images with disease labels from 30,805 unique patients in this dataset. The disease labels were created using Natural Language Processing (NLP) to mine the associated radiological reports. The labels include 14 common thoracic pathologies:

- Atelectasis
- Consolidation
- Infiltration
- Pneumothorax
- Edema
- Emphysema
- Fibrosis
- Effusion
- Pneumonia
- Pleural thickening
- Cardiomegaly

- Nodule
- Mass
- Hernia

We identify three limitations of this comparison. First, only frontal radiographs were presented to the radiologists and model during diagnosis, but it has been shown that up to 15% of accurate diagnoses require the lateral view (Raoof et al., 2012); thus expect that this setup provides a conservative estimate of performance. Third, neither the model nor the radiologists were permitted to use patient history, which has been shown to decrease radiologist diagnostic performance in interpreting chest radiographs (Berbaum et al., 1985; Potchen et al., 1979); for example, given a pulmonary abnormality with a history of fever and cough, pneumonia would be appropriate rather than less specific terms such as infiltration or consolidation)

An ideal database should provide the different angles of chest X-ray along with medical history of patient for better prediction of phenomena

Description of Training Dataset: NIH X-ray image dataset which also contains additional information like patient details and annotated information like disease prevalence is split into Training Dataset and Validation dataset in the ratio of 80:20 using python library skl.train_test_split. In the training dataset it is ensured that there is balanced dataset of images with pneumonia and images without pneumonia

Description of Validation Dataset: Validation dataset represent the 20% of the overall dataset for prevalence of pneumonia with imbalanced mix of images with pneumonia and images without pneumonia. It is ensured that images used in Training dataset is not used in validation dataset

5. Ground Truth

Ground Truth is acquired from NIH prepared datasets of 112,120 X-ray images with 14 disease labels from 30,805 unique patients. The image labels were extracted using NLP so there could be some erroneous labels but the NLP labelling accuracy is estimated to be >90%.

6. FDA Validation Plan

Patient Population Description for FDA Validation Dataset: The dataset contains 30,805 unique patients in the age group of new born to 100 years of age with an approx. mix of 44% and 56% of female and male population

Ground Truth Acquisition Methodology: Ground truth acquisition methodology can be categorised into two major category

Gold standard

The gold standard for a particular type of data refers to the method that detects disease with the highest sensitivity and accuracy. Any new method that is developed can be compared to this to determine its performance. Typical sources of ground truth for Pneumonia are

- Chest X-ray to look for inflammation in lungs
- Blood test to check white blood cell count
- Sputum tests (using a microscope to look at the gunk you cough up)
- A pulse oximetry test, which measures the oxygen in your blood

Silver standard

The silver standard involves hiring several radiologists to each make their own diagnosis of an image. The final diagnosis is then determined by a voting system across all of the radiologists labels for each image. Radiologists experience levels are taken into account and votes are weighted by years of experience.

Silver standard method of ground truth acquisition is the recommended approach considering the cost and accuracy of deduction

Algorithm Performance Standard: The PneumoniaX-ray has been evaluated and verified in accordance with software specifications and applicable performance standards through Software Development and Validation & Verification Process to ensure performance according to specifications, User Requirements and Federal Regulations and Guidance documents, “Guidance for the Content of Premarket Submissions for Software Contained in Medical Devices”. The performance of the PneumoniaX-ray device has been validated in a pivotal performance study that was carried out in simulated synthetic work-flow. Below are some of the key performance metrics

```
TRAIN METRIC ----------------------
Train acc: 60.39%
TEST METRICS ----------------------
Accuracy: 67.1875%
True Negative: 35
True Positive: 8
False Negative: 5
False Positive: 16

Sensitivity: 0.6153846153846154

specificity: 0.6862745098039216

Confusion Matrix: [[35 16]
 [ 5  8]]

Threshold where Precision is 0.8--------------------
Precision is: 1.0
Recall is: 0.07692307692307693
Threshold is: 0.62786263
F1 Score is: 0.14285714285714288

Threshold where Recall is 0.8--------------------
Precision is: 0.23809523809523808
Recall is: 0.7692307692307693
Threshold is: 0.3527323
F1 Score is: 0.36363636363636365
```

Compared to some of the previous study and work by Wang et al. (2017), Yao et al. (2017), CheXNet et al. (2017), PneumoniaX-ray performance is very close and the accuracy Pneumonia deduction is as follows

Wang et al. (2017) – 0.633
Yao et al. (2017) – 0.713
CheXNet et al. (2017) – 0.7680
PneumoniaX-ray – 0.671

The best F1 Score of PneumoniaX-ray 0.357 which is close radiologist and CheXNet score as shown below

Radiologist 1 - 0.383
Radiologist 2 - 0.356
Radiologist 3 - 0.365
Radiologist 4 - 0.442
Radiologist Avg. - 0.387
CheXNet - 0.435
PneumoniaX-ray - 0.363