FDA Submission

Name: Gabriel Ohaike

Name of your Device: ***Gab-XTech***

Algorithm Description:

> ***Gab-XTech*** is a 4-layer convolutional neural network with three dropout layers incorporated to prevent overfitting and improve generalization ability to unseen data. Trained on *NIH* chest X-ray consisting of 112,120 X-ray images with disease labels from 30,805 unique patients to predict the presence or absence of pneumonia through a 2D chest x-ray image. Practicing radiologists annotate a test set, on which we compare the performance of ***Gab-XTech*** to that of radiologists. We find that ***Gab-XTech*** was able to accurately predict the presence or absence of pneumonia using a threshold of 0.5 and F1.

1. General Information

Intended Use Statement:

- Assist radiologists to identify the presence or absence of Pneumonia in chest X-ray medical image

Indications for Use:

- Enhance the workflow of a radiologist by detecting the presence or absence of Pneumonia between for males and females between the age of 10 – 95 years.

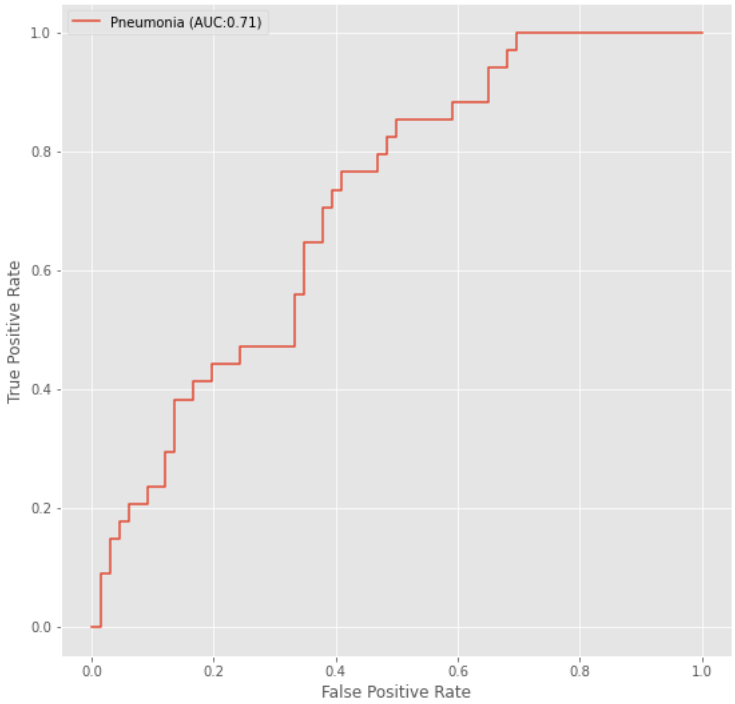
Device Limitations:

- This algorithm is not intended for use for 10 < age > 95 years of age.
- It does not have the tool or ability to identify any other disease.

Clinical Impact of Performance:

There is a clinical impact on performance that ***Gab-XTech*** put into consideration. The false-positive and false-negative. This is an indication that this algorithm is not perfect. False-positive occurs when a patient is wrongly diagnosed with Pneumonia and false-negative vice versa. Hence, predictions should be confirmed by a radiologist before making decisions. A plot is shown of AOC is shown below

.



From the plot above, the best AOC is seen at 0.71, this is the best performance that was feed to our model to minimize error.

2. Algorithm Design and Function

image processing → Data Analysis → Algorithm Design → Validation → Testing

DICOM Checking Steps:

In DICOM process, we conducted an image check to examine the body part image, modality, and patient position, thereafter examine each image, if the image does not belong to chest X-ray, the algorithm will print a message notifying the image part is not chest X-ray, this ensure for integrity and quality of the data before feeding it into the next preprocessing phase

Preprocessing Steps:

The preprocessing steps take on data from DICOM checking process. Here, the image is preprocessed to conformity to the standard, also undergoes transformation to fit into the same dimensions of the model. Resizing all the images ensures accurate predictions

Normalization Steps:

This process runs simultaneously with the preprocessing steps. We subtracted the image to the image mean and divided by the standard deviation. This ensures that all images fit perfectly before loading it into ***Gab-XTech*** for the prediction process.

3**. Algorithm training:**

The following standard was used to train our model:

Parameters:

ImageDatagGenerator was used for augmentation with a rescale of of 1. /255.0.

With these parameters adjusted for best performance as shown below.
Horizontal flip=True,
Vertical flip=False,
Height shift range=0.1,
Width shift range=0.1,
Rotation range=20,
Shear range = 0.1,
Zoom range=0.1

These parameters were used to simulate possible chest X-ray image positions. Our AI takes this into considerations for best performance

- Batch size of 9 was used in the training set. B**atch size** is a number of samples processed before model is updated. This controls the accuracy of estimate the error gradient.

- Adam(1r=1e-4) optimizer learning rate was used to optimize the efficiency of our learning rate.

- VGG16 pre-trained network from Keras for fine-tuning was incorporated into our model with the parameters below:

  Total params: 138,357,544
  Trainable params: 138,357,544
  Non-trainable params: 0

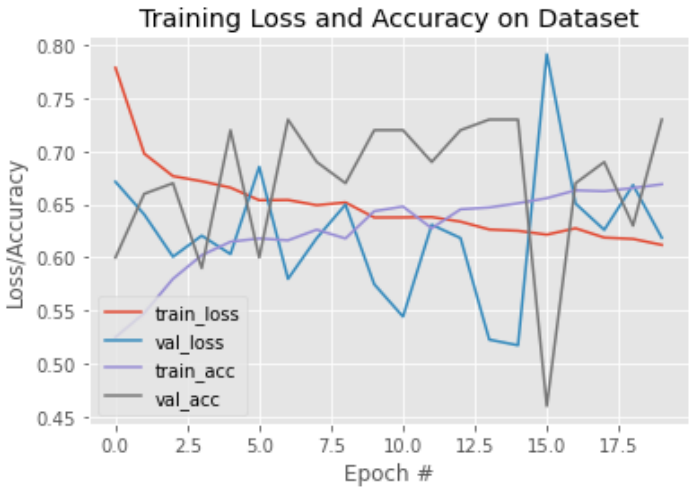
We transferred layer to block5_pool and created a VGG model.
First, choose 17 layers of VGG16 to fine-tune and freeze all but the last convolutional layer with the results below:

input_1 False
block1_conv1 False
block1_conv2 False
block1_pool False
block2_conv1 False
block2_conv2 False
block2_pool False
block3_conv1 False
block3_conv2 False
block3_conv3 False
block3_pool False

block4_conv1 False
block4_conv2 False
block4_conv3 False
block4_pool False
block5_conv1 False
block5_conv2 False
block5_conv3 True
block5_pool True

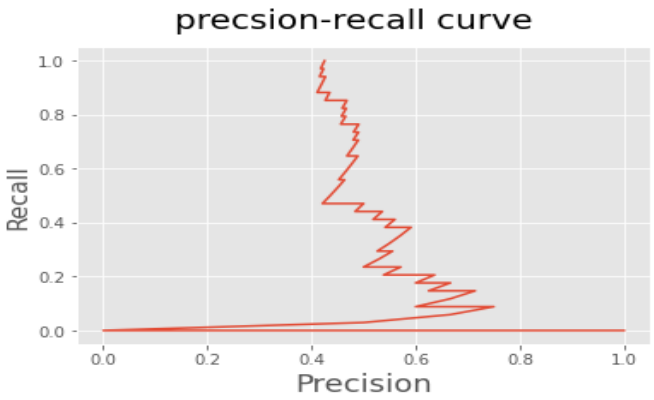
Three dropout-layer were added to the model to prevent overfitting and improve the generalization ability of unseen data. Four dense layers were added to make the rest architecture.

After training and validation, the results obtained by 20 epochs are shown in the plot below



As can be seen from the plot, the model made improvements and narrow the loss function, the smaller the loss the better our AI is able to make more accurate predictions.

Before sending for testing, we experimented with different performance matrices in order to ch oose the best threshold for best performance. Below is a precision/recall curve primarily used f or further enhancement in choosing the best parameter suitable for our model.



From our plot, the precision-recall curve is 0.425 this shows the tradeoff between precision and recall for different thresholds. A high area under the curve represents both high recall and high precision, where high precision relates to a low false-positive rate, and high recall relates to a low false-negative rate.

The image below is 100 X-ray images with classifications between 0 and 1. (1,1) confirms the pr esence or absence of pneumonia, (0,0) signifies no trace of pneumonia detected and (1,0) has a probability of either or not a patient has pneumonia.

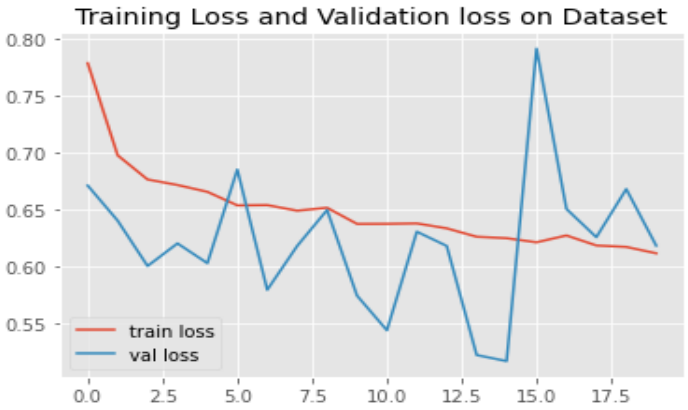
4. Databases

Description of Training Dataset:

Training set contains 89,696 X-ray images of *30,805* unique patients provided by NIH through DICOM file in NIH website. For the pneumonia detection task, we randomly split the dataset into training and validation set. Before inputting the images into the network, we performed image augmentation with random horizontal flipping and rescale to 1. /255 using an Image data generator and adjusted other parameters for best performance. We also prepared our train set to contain equal distributions of cases with pneumonia and those without pneumonia. This is done for best the algorithm performance.

Description of Validation Dataset:

The Validation set was taken from the remainder of the NIH dataset. We rescaled to 1. /255 using Image data generator. In order not to leak our validation set to our train set no other image data generator was added. The validation set was engineered to contain 20% of pneumonia cases.



The plot above is for evaluation performance of ***Gab-XTech*** training and test set. The training loss showed an improved learning rate and validation loss showed continuous improvement until about 14 epochs before declining. Overall, this is a good indication of a good learning and validation process

5. Ground Truth:

The methodology used to establish ground truth was provided by NIH. There are 112,120 X-ray images with disease labels from 30,805 unique patients in this dataset. This was labeled using Natural Language processing, this could be a major drawback that could impact their algorithm's for clinical performance. In NLP process, the pose for high probability mislabel image is always a possibility.

6. FDA Validation Plan

Patient Population Description:

GabP-Tech is built specifically for the following Patient Population detailed below:

- **Age ranges:** 9 – 85 years
- **Sex:** Male and Female
- **Type of imaging modality**: X-ray
- **Body part imaged**: chest
- **Prevalence of disease of interest:** Pneumonia

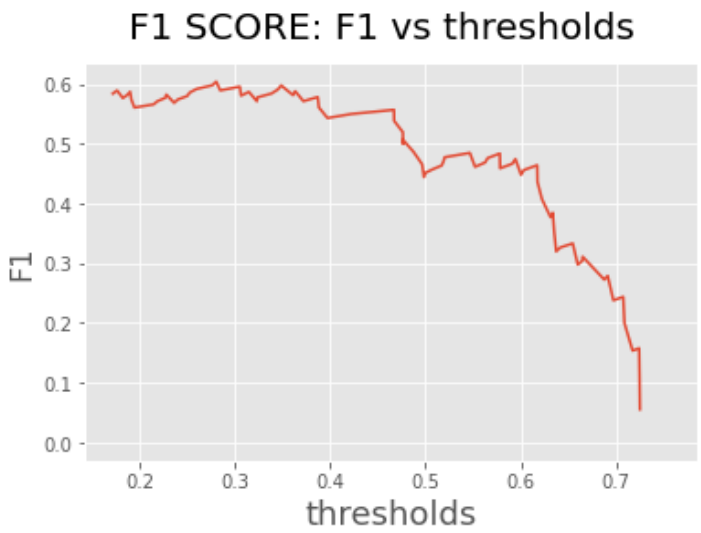
Ground Truth Acquisition Methodology:

For ground truth acquisition, I would choose a ***silver standard***. This method would require a team of radiologists with each scoring an image for the presence or absence of pneumonia based on the number of years of experience. The radiologists would be grouped into three groups, those below 5 years of experience, those with 5 – 10 years of experience, and those above 10 years of experience. This method is preferred to the gold standard because of the scoring methods considering different levels of radiologists to analyze the same image other than relying on one radiologist.

Algorithm performance:

F1 scores and thresholds would be used to measure the performance of my model. The link provides a reference to this metric.
CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning:
https://arxiv.org/pdf/1711.05225.pdf by Pranav Rajpurkar et el.

Below shows how F1 score and threshold curve. The best threshold was obtained for best performance.



GabP-Tech used data provided by the NIH to evaluate the model. A threshold of 0.54 for best performance and the following results were obtained

```
Load file test1.dcm ...
No Pneumonia
Load file test2.dcm ...
Pneumonia Present
Load file test3.dcm ...
Pneumonia Present
Load file test4.dcm ...
The model is not valid because the of image position, the image type or th
e body part are not OK
No Pneumonia
Load file test5.dcm ...
The model is not valid because the of image position, the image type or th
e body part are not OK
No Pneumonia
Load file test6.dcm ...
The model is not valid because the of image position, the image type or th
e body part are not OK
No Pneumonia
None
```

After assessing the performance of radiologists and ***Gab-XTech*** on the test set for the pneumonia detection task. ***Gab-XTech*** was able to classify the presence or absence of pneumonia using a threshold of 0.54.