FDA Submission

Your Name: Tasnim Abusharkh

Name of your Device: Detecting Pneumonia in 2D chest x-ray images software

Algorithm Description

1. General Information

Intended Use Statement: This Algorithm detects the pneumonia in chast x-rays. It could work to help the radiologist to detect if there is or not pneumonia is x-ray image. If the algorithm missed diagnosing the pneumonia, the patient could die if he was not treated correctly and on time.

Indications for Use: The model was trained on data for people with age range 1 to 95, the race is not determined in this dataset, and patient position is either 'PA' or 'AP', the dataset contains x-rays for males and females, the percent of males to females is 3 to 2.

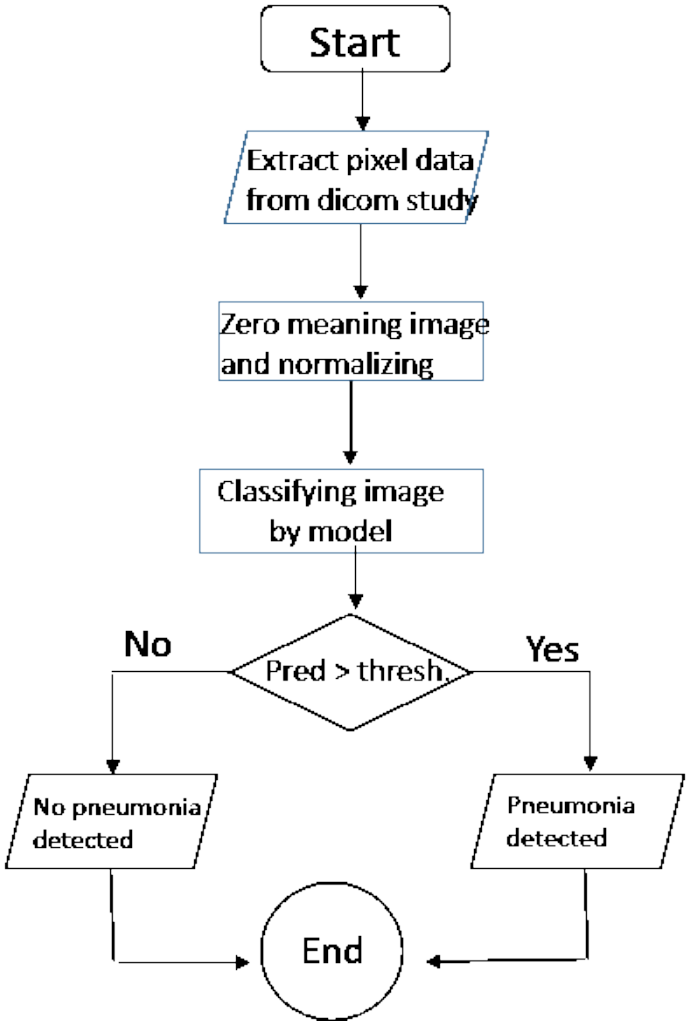
Device Limitations: There was no much impact of other diseases with penumonia on detecting it by the machine, the model failed in 15% of the cases.

Clinical Impact of Performance: There will be no impact of FP on patient by this algorithm, because the x-ray image will be reviewed by the radiologist, so he will be able decide that there is no pneumonia in the taken image. But maybe there will be an impact of the patient was misdiagnosed FN, which happened 15% of the cases while training this model.

2. Algorithm Design and Function

<< Insert Algorithm Flowchart >>

Following is the model flowchart:



DICOM Checking Steps: We extract the pixel data from dicom file and resize the image, and check the following: 1- the Modality of the image: we accept only the 'DX' modality, 2- we check the patient position it should be 'AP' or 'PA', 3- we check the part of the body examined it should be 'Chest', we dismiss the image if one of those conditions was not applied.

Preprocessing Steps: After extracting the image we resize it and reshape it in order to be suitable for model input, it should be in size (1,224,224,3), and we normalize the image by subtracting the mean of the images was used to train the model and then we dived image pixels by the strandard diviation was used to standarize the training images.

**CNN Architecture: To build the model we find-tuned VGG16 by adding: 1- dense layer with 1024 input, and 'relu' activation method 2- dense layer with 512 input, and 'relu' activation method, 3- dense layer with 1 input, wtih 'sigmoid' activation method. **

3. Algorithm Training

Parameters:

* Types of augmentation used during training: the augmentation method used the following parameters:

horizontal_flip=True, vertical_flip = False, height_shift_range=0.1, width_shift_range=0.1, rotation_range=20, shear_range=0.15, zoom_range=0.20.

* Batch size : for the training data it was: 50, and for validation data it was: 100

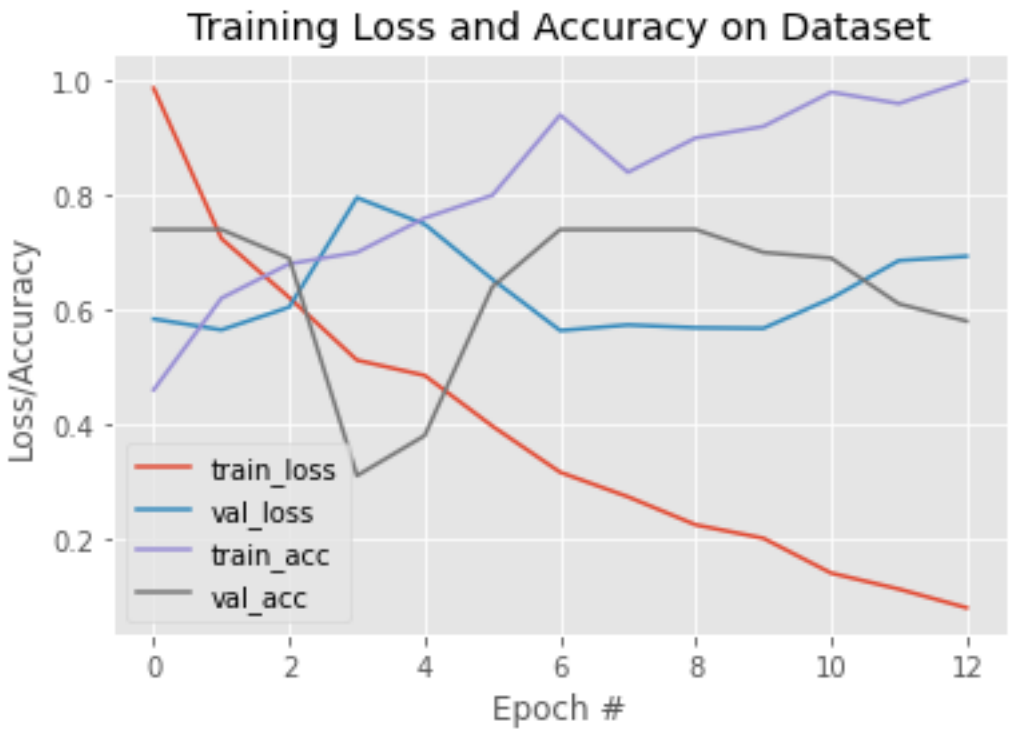
* Optimizer learning rate: 0.01

* Layers of pre-existing architecture that were frozen: the first 17 layers.

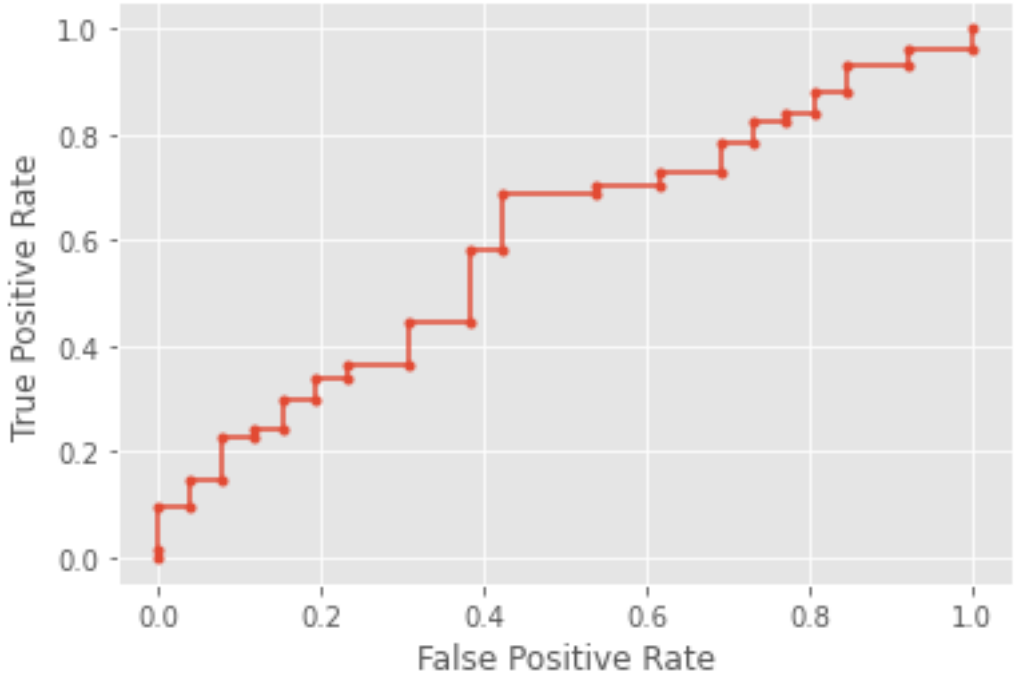
* Layers of pre-existing architecture that were fine-tuned: 1 layer

* Layers added to pre-existing architecture: 4 dense layers

<< Insert algorithm training performance visualization >>

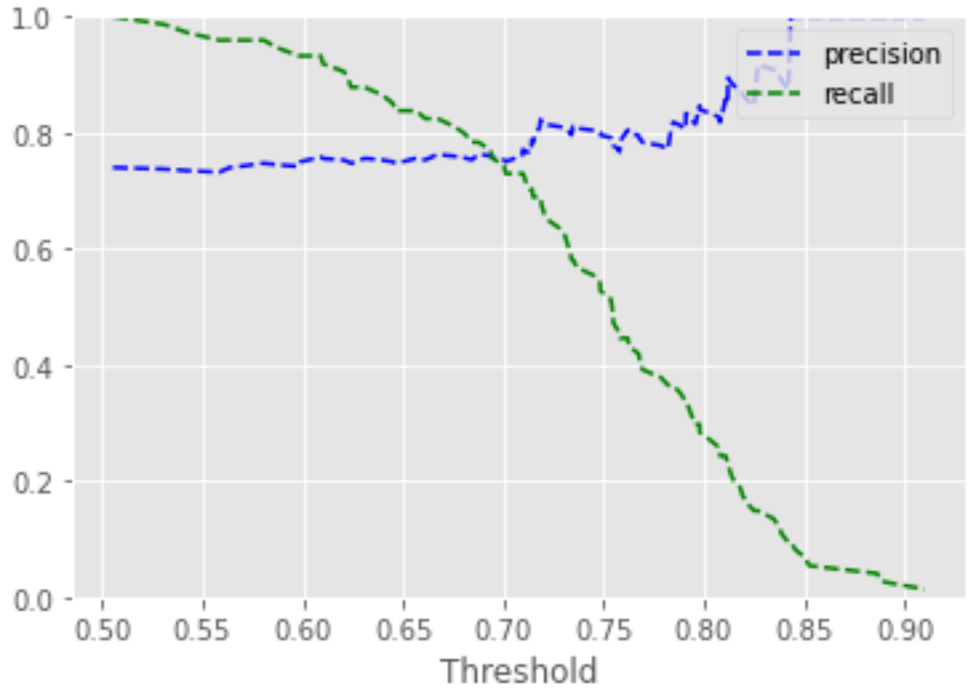


<< Insert P-R curve >>



Final Threshold and Explanation:

Following is the threshold curve:



And from the threshold curve we find that threshold more than 0.55 will increase the FN cases, this is not correct, so, we chose 0.55 threshold and after choosing it we can see the following performance statistics:

```
AUC score=0.8186616780781717
F1 score=0.8505747126436781
```

As we aim to maximized the precision over recall, because in our case if the patient was misdiagnosed positively (FP) will be less risky than if he was negatively misdiagnosed. So, we chose 0.55 threshold as a higher threshold a worse precision percentage.

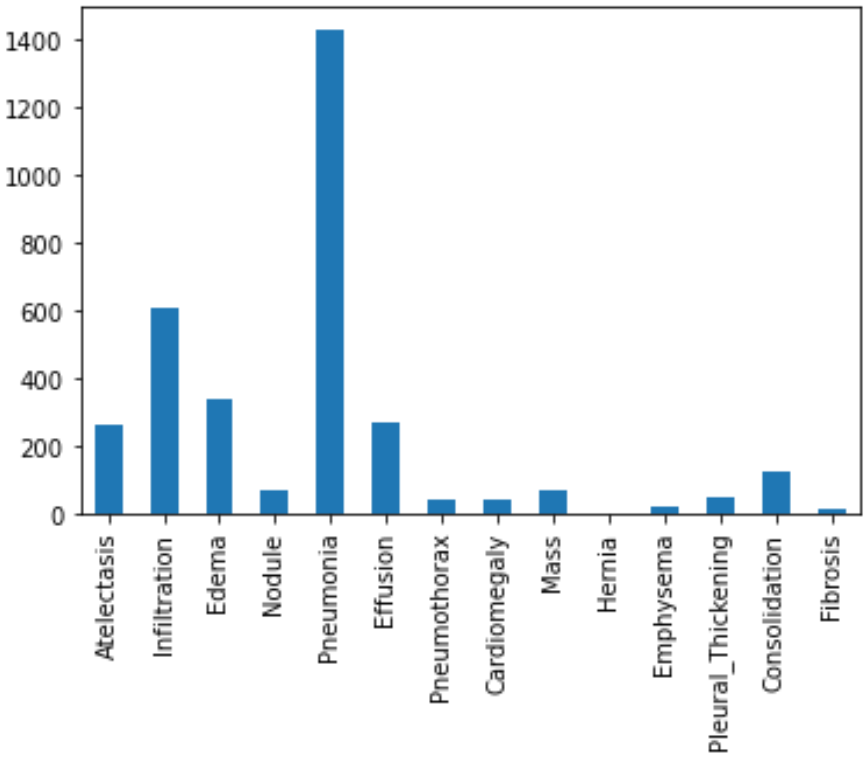
4. Databases

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

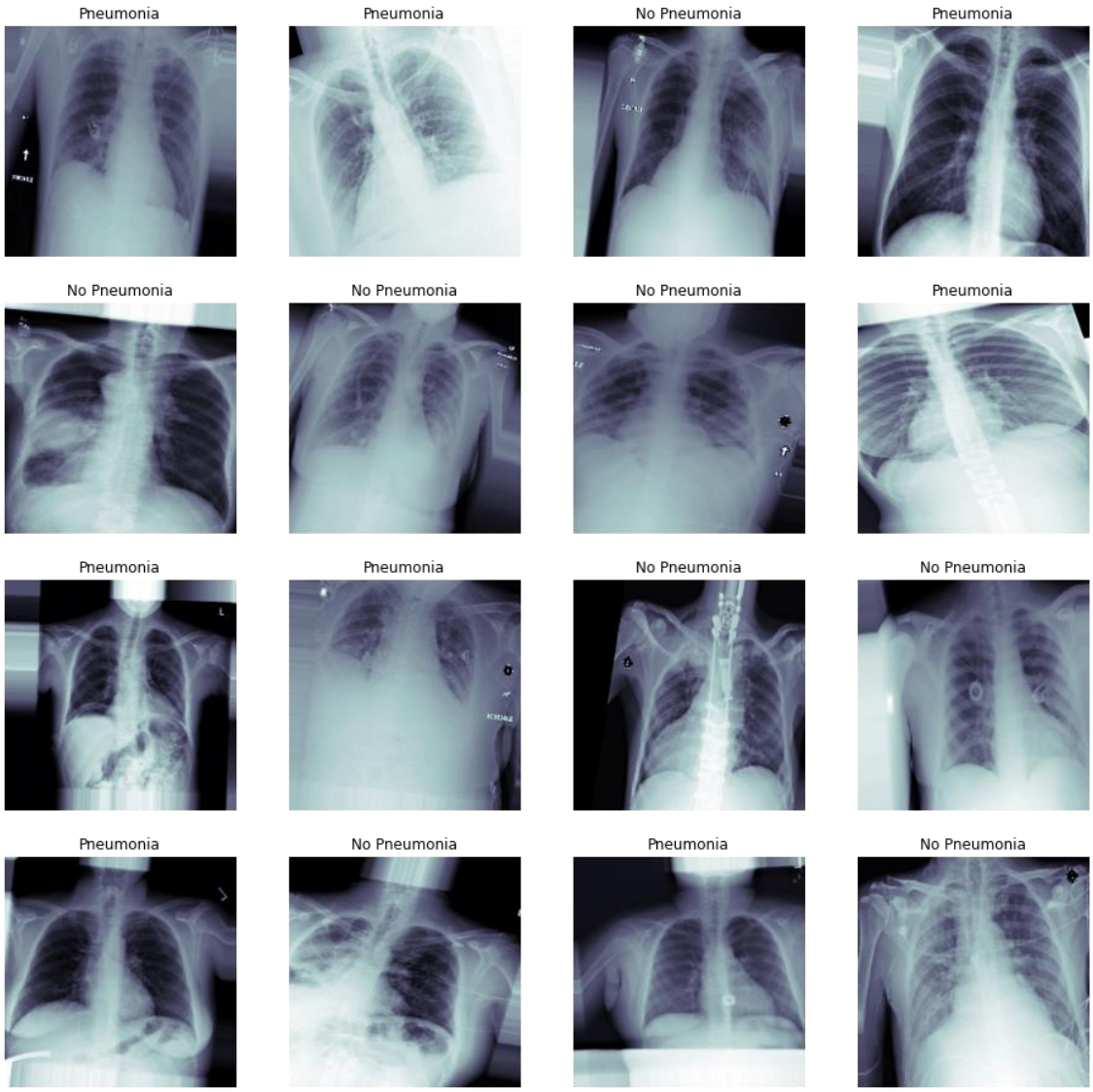
Description of Training Dataset:

The training set was taken from the NIH (National Institute of Health). It is a data contains x-ray images for the chest part form the human body, the data contains images taken for both males and females in ages from 1 to 95. The view position of the patients was either ‘AP’ or ‘PA’. And there. The pneumonia cases in the training data was 50%, but it was 0.25 in the validation data.

There was also an existence for another 13 issue comorbid to pneumonia in the dataset we worked on. We can see the following plot for the cases and their counts.



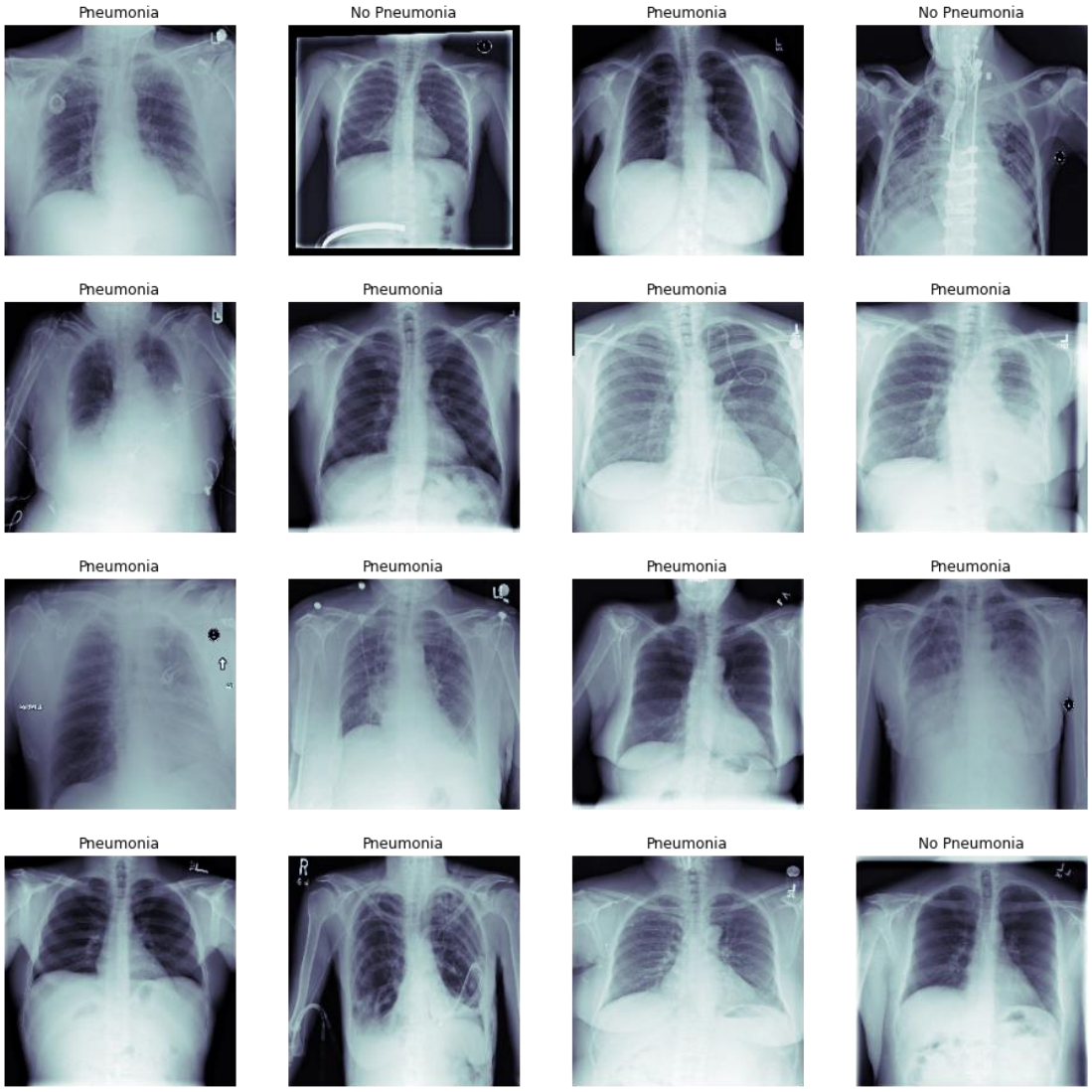
Following are sample of images from the training dataset:



Description of Validation Dataset:

The validation set was taken from the same data as the training dataset. It has the same distribution for the gender, ages, view position, and other diseases as the training dataset.

Following are sample of images in validation dataset:



5. Ground Truth

The final labels for each image were assigned via adjudicated review by three radiologists. Each image was first reviewed independently by 3 radiologists. For the test set, radiologists were selected at random for each image from a cohort of 11 American Board of Radiology certified radiologists. For the validation set, the 3 radiologists were selected from a cohort of 13 individuals, including board certified radiologists and radiology residents.

If all readers were in agreement after the initial review, then that label became final. For images with label disagreements, images were returned for additional review. Anonymous labels and any notes from the previous rounds were also available during each iterative review. Adjudication proceeded until consensus, or up to a maximum of 5

rounds. For the small number of images for which consensus was not reached, the majority vote label was used.[1].

In our model the NLP-derived labels are sub-optimal because they will be extracted by a machine which make it not the best solution as the machine will not perform like the human in this field. So, we better not use such technique in labeling data or ground truth acquisition.

6. FDA Validation Plan

Patient Population Description for FDA Validation Dataset: For validation the model needs data with high resolution images, in order for the intensities to be clear and differences to be sharp between images with finding and images without. The images should be taken for the human's chest by x-ray, the dataset should contain images for both males and females and the distribution should reflect the distribution of the disease in both genders in the real life. The age of the patients should be distribution also on all ages. If other diseases in normal situations comes comorbid to pneumonia, it also should be considered when we preparing data for training and validation.

**Ground Truth Acquisition Methodology: To create the optimal data for training model like this, we need data labeled after making polymerase chain reaction (PCR) test, this will be better in order to ensure training the model on images with correct data in 100%.

Algorithm Performance Standard: The algorithm performance should be measured by F1 score[2] and AUC method.

References:

[1] NIH Chest X-ray dataset at NIH Chest X-ray dataset. Reviewed 26 January 2021.

[2] P. Rajpurkar et al., "CheXNet: radiologist level pneumonia detection on chest x-rays with deep learning,", [Online]. Available: http://arxiv.org/abs/1711.05225.