FDA Submission

Name: Wyckliffe Aluga

Devive Name: PD

Algorithm Description

1. General Information

Intended Use Statement

Fast-tracking pneumonia diagnostics.

Indications for Use

Used to help radiologists fast track pneumonia patients to receive immediate medical care for pneumonia conditions while awaiting biopsy results. This algorithm can be used for both male and female patients of august 20 to 80. The images to be used must be:

- Image of the Chest
- Taken by a Digital X-ray
- The image should be taken in the either posteroanterior or anterior-posterior positions

Device Limitations

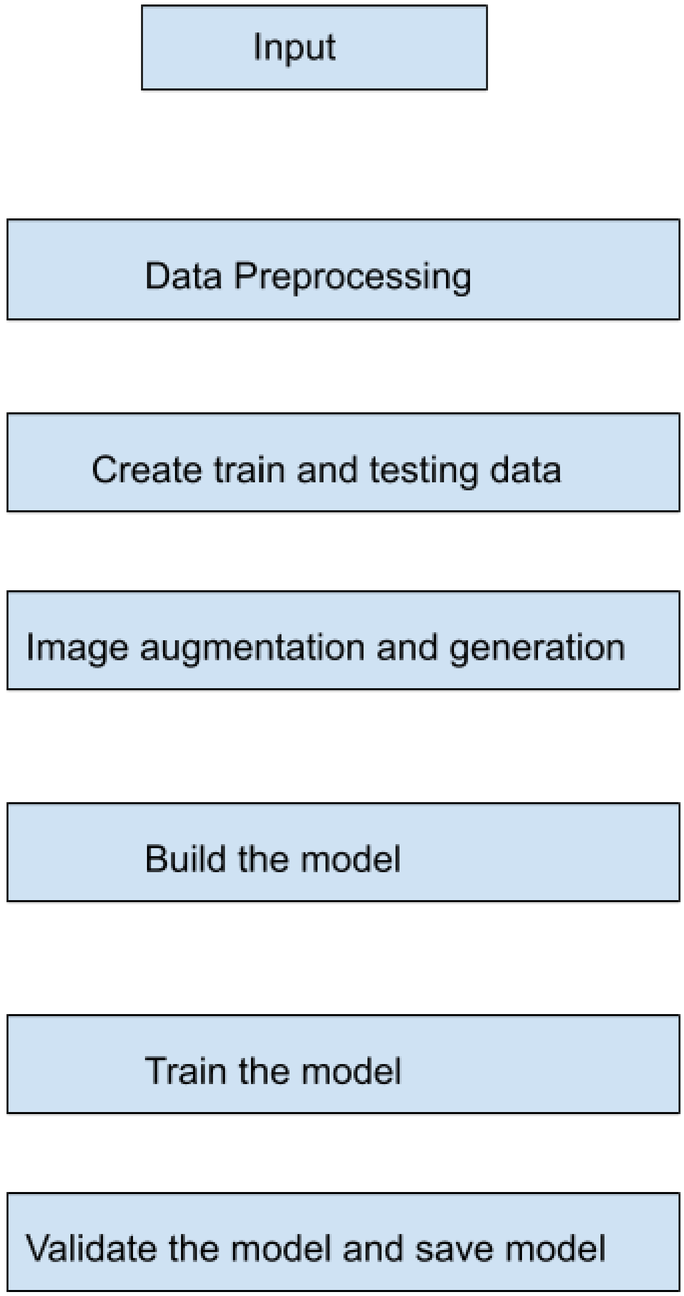
It cannot be used to confirm pneumonia without radiologists and/or biopsy results. When sampled with 14 common diseases that could be misdiagnosed as Pneumonia, this algorithm,

- In the presence of Infiltration disease, it performed poorly because it is not able to differentiate between Pneumonia and infiltration
- In addition, in the presence of edema diseases, the algorithm could not definitively differentiate it with Pneumonia.

Clinical Impact of Performance

- It is important to track the FP and FN of the model.
- The ideal situation would be to minimize both FP and FN but for this case, since Pneumonia could be fatal if not quickly diagnoses and treated the most important metric to reduce if the FP as could lead to catastrophic effects if Pneumonia is not detected early.

2. Algorithm Design and Function



DICOM Checking Steps

- The images are checked if they are valid to be fed into the model these checks include:
    - Checking if the body part examined is the chest

- Confirming that indeed the image was taken by a digital x-ray
- Confirming the image position is anterior-posterior or posteroanterior

Preprocessing Steps

- Images we normalized by subtracting the mean intensities and dividing by the standard deviation of the intensities.

CNN Architecture

- The CNN architecture used was a transfer learning from Keras VGG model.
- After flattening, 4 dense layers and 4 dropout layers were added onto the VGG model

3. Algorithm Training

Parameters

Types of augmentation used during training

- rescalling
- Horizontal shift range
- width shift range
- rotation range
- shear range
- zoom range
- horizontal flip

Batch size

- batch_size = 32

Optimizer learning rate

- learning_rate=1e-4

Layers of pre-existing architecture that were frozen

- conv2d_1
- conv2d_2
- max_pooling2d_1
- conv2d_3
- conv2d_5

Layers of pre-existing architecture that were fine-tuned

- Block 5

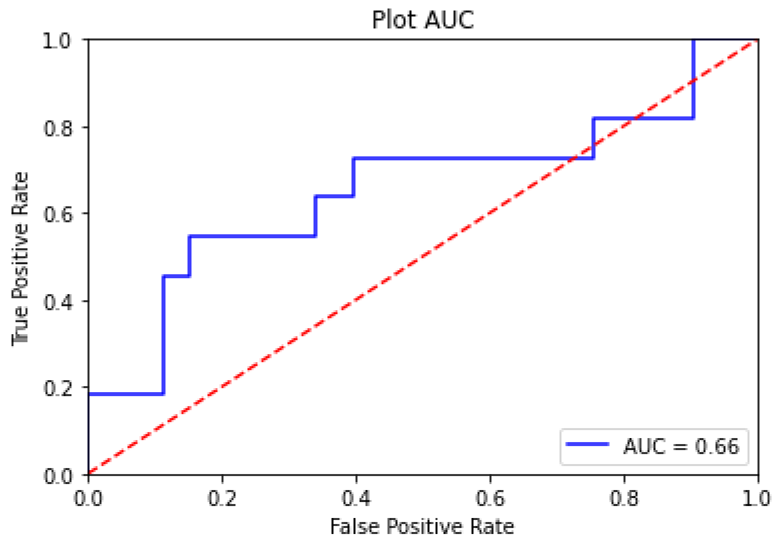
Layers added to pre-existing architecture

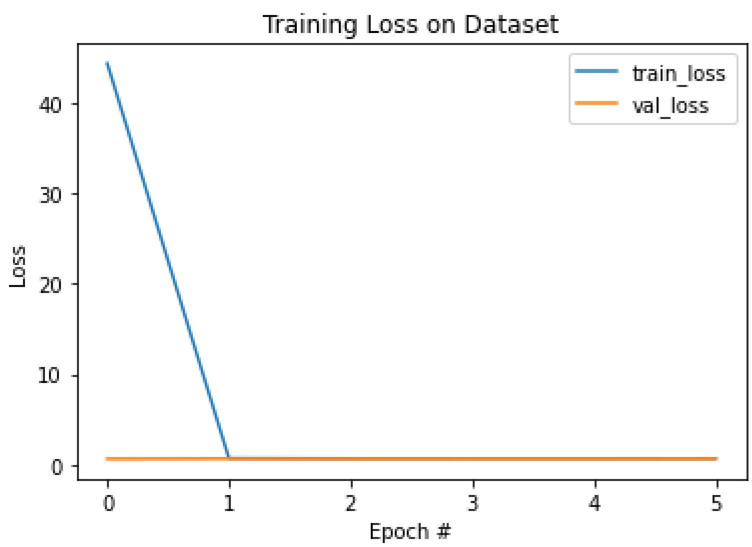
- Flatten
- Dropout_1
- Dense_1
- Dropout_2
- Dense_2
- Droupout_3
- Dense_3
- Droupout_4
- Dense_4

Insert algorithm training performance visualization

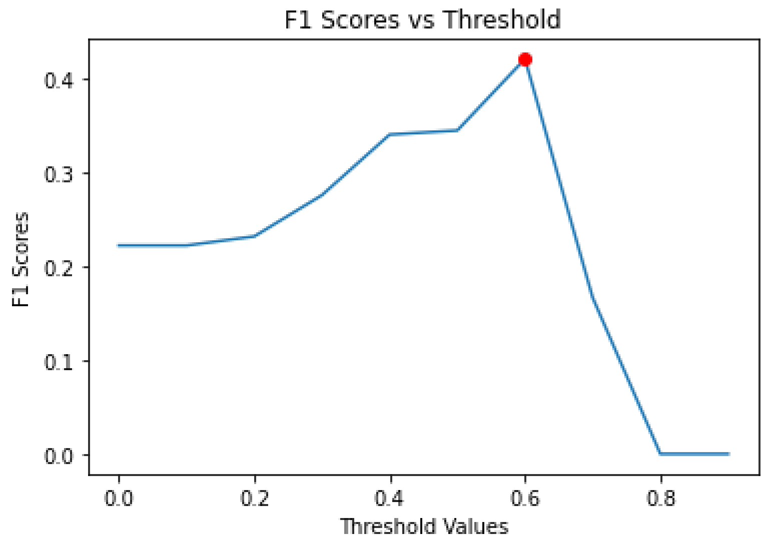
```
Epoch 1/20
36/36 [==============================] - 61s 2s/step - loss: 44.0859 - binary_accuracy: 0.5052 - val_loss: 0.6548 - val_bina
ry_accuracy: 0.8750

Epoch 00001: val_loss did not improve from 0.58290
Epoch 2/20
36/36 [==============================] - 61s 2s/step - loss: 0.7211 - binary_accuracy: 0.4952 - val_loss: 0.6722 - val_binar
y_accuracy: 0.8750

Epoch 00002: val_loss did not improve from 0.58290
Epoch 3/20
36/36 [==============================] - 60s 2s/step - loss: 0.6941 - binary_accuracy: 0.5017 - val_loss: 0.6784 - val_binar
y_accuracy: 0.8750

Epoch 00003: val_loss did not improve from 0.58290
Epoch 4/20
36/36 [==============================] - 60s 2s/step - loss: 0.6944 - binary_accuracy: 0.5052 - val_loss: 0.6869 - val_binar
y_accuracy: 0.8750

Epoch 00004: val_loss did not improve from 0.58290
Epoch 5/20
36/36 [==============================] - 61s 2s/step - loss: 0.6951 - binary_accuracy: 0.4939 - val_loss: 0.6872 - val_binar
y_accuracy: 0.8750

Epoch 00005: val_loss did not improve from 0.58290
Epoch 6/20
36/36 [==============================] - 60s 2s/step - loss: 0.6933 - binary_accuracy: 0.4878 - val_loss: 0.6922 - val_binar
y_accuracy: 0.8750

Epoch 00006: val_loss did not improve from 0.58290
```

Insert P-R curve





Final Threshold and Explanation



-
- The threshold of 0.60 gives best f1 at 0.4211

4. Databases

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

Description of Training Dataset

- The training set was created by splitting the datasets 80% training and 20% percent stratifying with the pneumonia class. The equal proportions of pneumonia in both training and validation datasets were created before augmentation of the images.

Description of Validation Dataset

- The validation dataset was a 20% split of the entire dataset stratified with pneumonia class. The set was balanced and rescaled as part of the augmentation

5. Ground Truth

- The dataset provided to you for this project was curated by the NIH

- There are 112,120 X-ray images with disease labels from 30,805 unique patients. The disease labels were created using Natural Language Processing (NLP) to mine the associated radiological reports.
- The biggest limitation of this dataset is that image labels were NLP-extracted so there could be some erroneous labels but the labeling accuracy is estimated to be >90%.

6. FDA Validation Plan

Patient Population Description for FDA Validation Dataset

- Both male and female patients ages 20-80 with at least 20% diagnosed with Pneumonia
- For an ideal dataset, these patients the following diseases should not be present
    - Atelectasis
    - Consolidation
    - Infiltration
    - Pneumothorax
    - Edema
    - Emphysema
    - Fibrosis
    - Effusion
    - Pleural thickening
    - Cardiomegaly
    - Nodule
    - Mass
    - Hernia
- Imaging modality should be DX
- Body part must be Chest and the patient position of AP and PA
- There should be an equal balance in gender, age, race, and socioeconomic status.
- Diseases that are more commonly found with pneumonia should be present so that detection becomes easier.
- Diseases that are not at all found with pneumonia should be eliminated

Ground Truth Acquisition Methodology

- Silver standard; a weighted average of 3 radiologists diagnosis

Algorithm Performance Standard

- The algorithm was optimized for the highest F1 score
- From the research paper below, the radiologist's average F1 score was at 0.387 (0.330, 0.442) if we can optimize the model to F1 values within the radiologist's F1 scores that would be great.
- https://arxiv.org/pdf/1711.05225.pdf