FDA Submission

Your Name: Nikos Sourlos

Name of your Device: Automatic Pneumonia Detection from X-ray Images

Algorithm Description

1. General Information

Intended Use Statement: Assist radiologists identify pneumonia in chest x-ray images

**Indications for Use: This algorithm is intended for use on patients of both genders from the ages of 1-95 who have been administered a chest X-ray study on a AP or PA view position. Images should be in DICOM format and the results should also be verified by a radiologist.
**

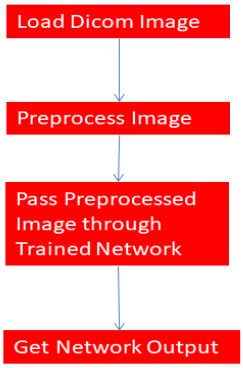
Device Limitations: Image testing could be done in a regular CPU. Time limitations should be taken into account. Moreover, since there is a correlation of pneumonia with infiltration and edema, presence of these diseases in the image could lead to poor results in terms of pneumonia detection. In addition to that, there is patient overlap during the splits which can make the model performance biased. Finally, the distribution of viewing positions is different for pneumonia patients from the total population.

Clinical Impact of Performance:

The model chosen with the highest f1 score has a relative high recall and a low precision. This means that there will be many false positives (image recognized as pneumonia while it corresponds to a healthy individuals) and only a few false negatives (image recognized as healthy while it has pneumonia). Since our tool aims to assist radiologists, it is important to having as few false negatives so that we miss as few patients with pneumonia as possible. On the other hand, since further examination will be performed, indidividuals who are healthy may be recognized by the radiologists in their examination.

2. Algorithm Design and Function

<< Insert Algorithm Flowchart >>



**DICOM Checking Steps: Image should be a Chest X-ray, with AP or PA position of patient, and modality should be digital radiography (DX) **

Preprocessing Steps: Image is normalized (mean subtracted and divided by its std), resized to 1*224*224*1, and then is repeated 3 times across the last dimension to finally get an image of 1*224*224*3, which is the input size for the VGG model.
CNN Architecture:
Below the original layers from the VGG model which are used in our network are presented. After them, the added layers are also depicted.
Model: "model_1"

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
=================================================================
Total params: 14,714,688
Trainable params: 2,359,808
Non-trainable params: 12,354,880

Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
model_1 (Model)              (None, 7, 7, 512)         14714688
_________________________________________________________________
average_pooling2d_1 (Average (None, 1, 1, 512)         0
_________________________________________________________________
flatten_1 (Flatten)          (None, 512)               0
_________________________________________________________________
dense_1 (Dense)              (None, 64)                32832
_________________________________________________________________
dropout_1 (Dropout)          (None, 64)                0
_________________________________________________________________
dense_2 (Dense)              (None, 1)                 65
=================================================================
Total params: 14,747,585
Trainable params: 2,392,705
Non-trainable params: 12,354,880
```

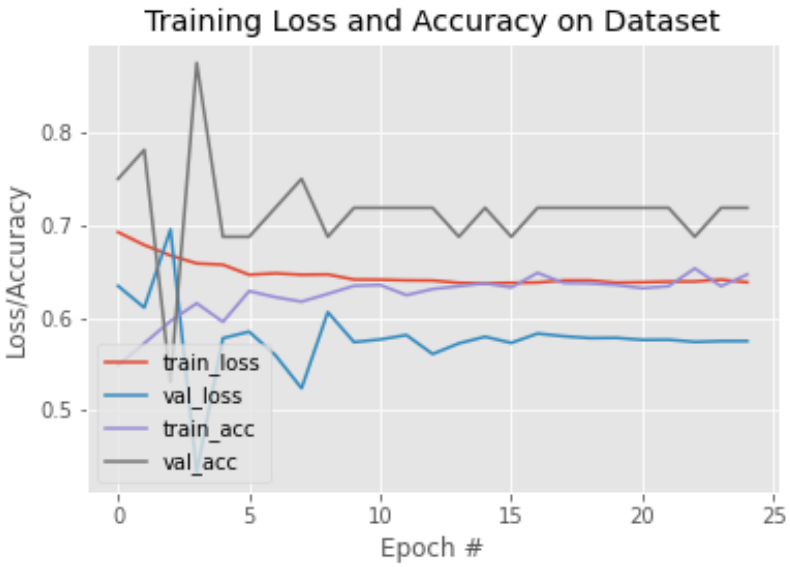
3. Algorithm Training

Parameters:

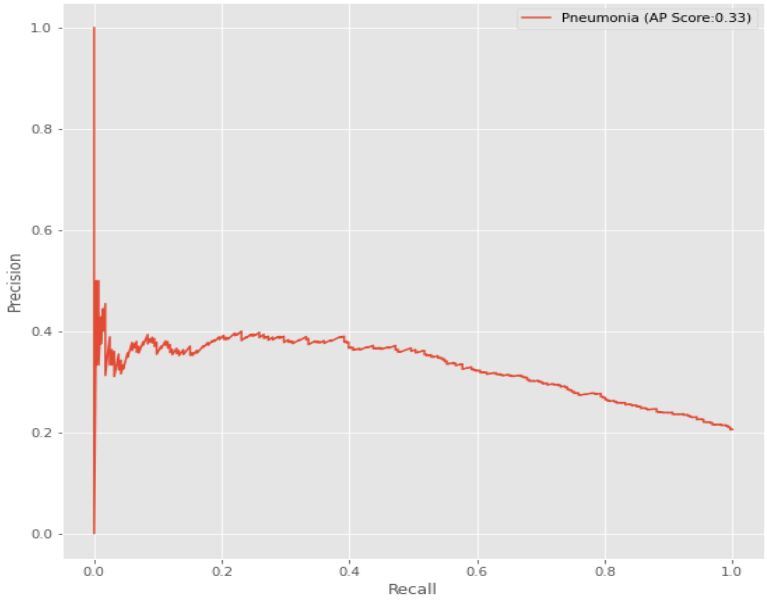
During training we used the following augmentation parameters:
samplewise_center=True, samplewise_std_normalization=True,
horizontal_flip = True, vertical_flip = False, height_shift_range= 0.1,
width_shift_range=0.1, rotation_range=10, shear_range = 0.1,
zoom_range=0.1.
Moreover, we choosed an initial learning rate of 0.0001 with a step decay of 0.25 every 5 epochs. We trained for a total of 25 epochs with a batch

size of 32. We only trained (fine-tuned) the last two layers of the VGG model along with the added ones from us (can be seen above). The rest 17 were remain frozen.

<< Insert algorithm training performance visualization >>



<< Insert P-R curve >>

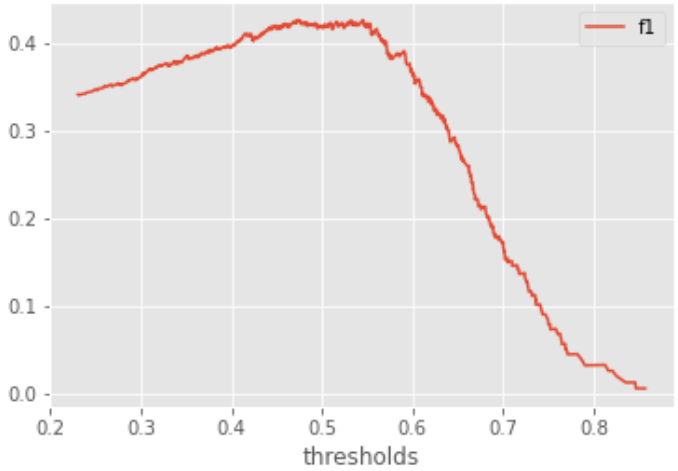


Final Threshold and Explanation:
Maximum f1: 0.4263392857142857

Threshold: 0.47376856207847595

We chose the threshold for which F1 score is maximized, as can be seen in the figure below. The chosen F1 score is better than the average radiologists score, as can be seen in the paper:

Rajpurkar, Pranav & Irvin, Jeremy & Zhu, Kaylie & Yang, Brandon & Mehta, Hershel & Duan, Tony & Ding, Daisy & Bagul, Aarti & Langlotz, Curtis & Shpanskaya, Katie & Lungren, Matthew & Ng, Andrew. (2017). CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning.



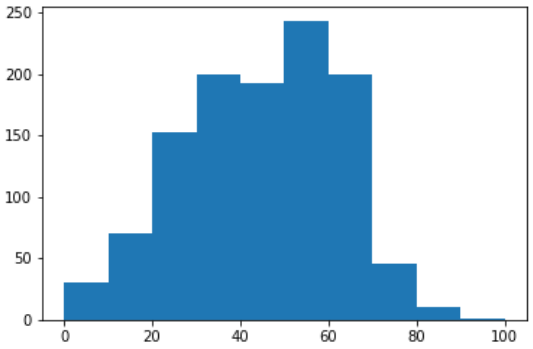
4. Databases
(For the below, include visualizations as they are useful and relevant)

Description of Training Dataset:
Description of Validation Dataset:

Since there are only 1431 pneumonia cases, we should make sure that 80% of these end up in the training set and 20% in the validation set. This means that 1144 pneumonia images end up in training set and the rest 286 in the validation set. Moreover, since there are 110677 non pneumonia cases, 80% of those end up in the training and 20% in the validation set. That is 88542 in the training and 22135 in the validation set. Since the training set must be balanced we only keep 1144 non pneumonia cases (maybe with other diseases). Since the validation set should be representative of the distribution of diseases in the real world, and assuming that 20% of patients have pneumonia, we keep in total 286 pneumonia and 1144 non pneumonia (maybe with other diseases) cases.

For all other variables in our dataset such as age, sex, view position, the distribution of both training and validation sets should follow the same distribution as our original full dataset. For example, we should have more males than females and more PA than AP view positions in both of these sets. We assume that these distributions are preserved during the above splits. This was confirmed for the gender only.

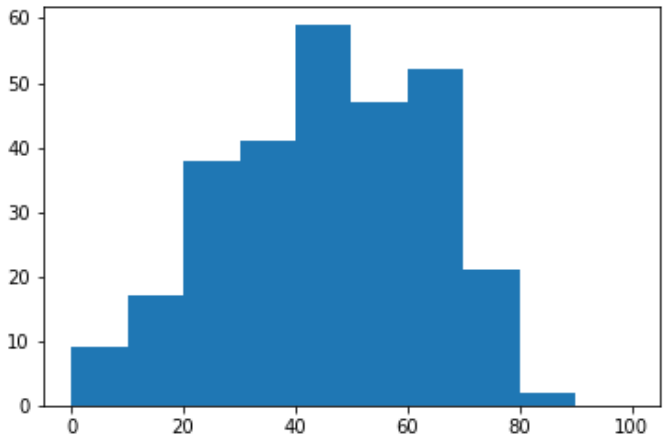
In the training set we have 59% males and 41% females. The Age distribution is shown below:



Moreover, the percentages of each label are the following:

Pneumonia	0.500000
Infiltration	0.293706
No Finding	0.265734
Effusion	0.155594
Atelectasis	0.142483
Edema	0.124563
Consolidation	0.067745
Nodule	0.055070
Mass	0.052885
Pneumothorax	0.041521
Pleural_Thickening	0.031469
Cardiomegaly	0.024913
Emphysema	0.019231
Fibrosis	0.013986
Hernia	0.001311

For the validation set we have 55% males and 45% females. The age distribution is shown below:

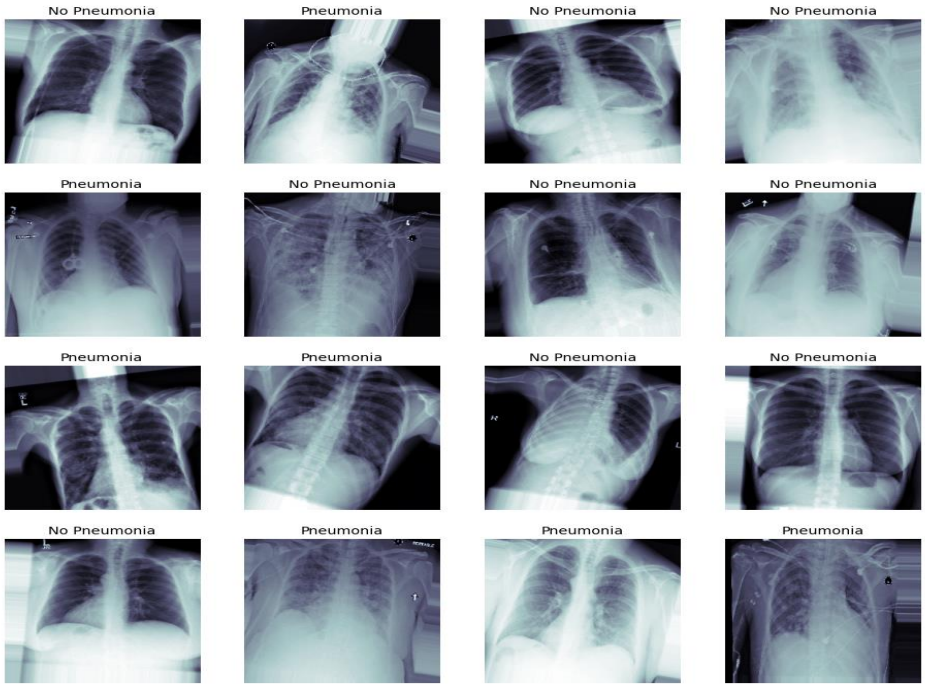


The percentages of each label are the following:

No Finding	0.457343
Infiltration	0.209091

Pneumonia	0.200000
Effusion	0.119580
Atelectasis	0.106993
Edema	0.059441
Nodule	0.052448
Mass	0.047552
Consolidation	0.045455
Pneumothorax	0.031469
Cardiomegaly	0.025874
Emphysema	0.025874
Pleural_Thickening	0.025874
Fibrosis	0.015385
Hernia	0.001399

Some examples of pictures in our training set are shown below:



5. Ground Truth

There are 112,000 chest x-rays with disease labels acquired from 30,000 patients in our dataset. There are 14 diseases in total (may more than one present each time) along with an extra class for non diseas ed individuals. The disease labels are:

No Finding
Infiltration
Effusion
Atelectasis
Nodule

Mass
Pneumothorax
Consolidation
Pleural_Thickening
Cardiomegaly
Emphysema
Edema
Fibrosis
Pneumonia
Hernia

It should be noted that the labels extracted automatically with NLP are suboptimal to those that can be extracted manually by radiologists, which in turn are suboptimal of the tissue biopsy labels. Therefore, the accuracy of our model can be suboptimal as well.

6. FDA Validation Plan

Patient Population Description for FDA Validation Dataset:

Similar as the indications for use statement. The imaging modality we used is 'DX', and we examined only Chest x-rays in AP or PA position. Images of other positions may also included in an ideal dataset.

In our ideal dataset, we should have images of patients of all ages and genders (preferably equal number of images for each age group and equal number of males and females). Moreover, the prevalence of pneumonia would be much greater than what we have in our dataset now (at least 20% or even 50%). Finally, the remaining images should not consist of images of edema and infiltration, diseases which have a strong correlation with pneumonia and which will affect the results of our AI algorithm.

Ground Truth Acquisition Methodology:
Based on the paper presented above (CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning), the silver standard is the weighted average of a few radiologists. The gold standard would be tissue biopsies which were not obtained for our data.

Algorithm Performance Standard:
Based on the F1 score, our algorithm should perform better than the average F1 score of the radiologists, which is true, as shown below. The source is again taken from the paper mentioned above.

	F1 Score (95% CI)
Radiologist 1	0.383 (0.309, 0.453)
Radiologist 2	0.356 (0.282, 0.428)
Radiologist 3	0.365 (0.291, 0.435)
Radiologist 4	0.442 (0.390, 0.492)
Radiologist Avg.	0.387 (0.330, 0.442)