FDA Submission

Raphael Morel:

The Algorithm:

Algorithm Description

1. General Information

Intended Use Statement:
The Algorithm's intended use is to detect abnormalities in chest radio images in order to help the radiologist make a decision on the presence or absence of pneumonia (data from the EDA).

Indications for Use:
Use the algorithm with only radio images of CHEST in DICOM format following the HIPAA rules.
Patients must have between 1 and 95 years old.
After the X-Ray is completed, the data is sent to the Algorithm. It will check the initial criteria.
If it satisfies them, then it will make a prediction and then send its prediction as well as the X-Ray image to a radiologist who will be the final decision-maker.

Device Limitations:
The patient's position must be AP or PA.
Pleural thickening and fibrosis can decrease the performance of the model as the pixel intensity distribution is quite similar to the pneumonia one and the algorithm will not be able to identify pneumonia correctly.

Clinical Impact of Performance:
There is a tradeoff between precision and recall. If the algorithm is evaluated on precision, then, clinically, the number of correct results will increase whereas the number of true positives will decrease. For a best performance, the threshold is set to 0.44.
In case of FP, the algorithm will not detect an existing pneumonia whereas if a FN occurs the algorithm will detect a pneumonia while the patient has no pneumonia. This could significantly influence the decision making of the medical staff.

2. Algorithm Design and Function

DICOM Checking Steps: DCMread and use of pixelArray. Then resizing to fit it in the model. Use DICOM to validate the model and address issues with patient's position, type of image and examined part of the body.

Preprocessing Steps: Normalization of DICOM image and resizing into (224,224,3) are mandatory to pass the image into the algorithm.

CNN Architecture: VGG16 with added layers.

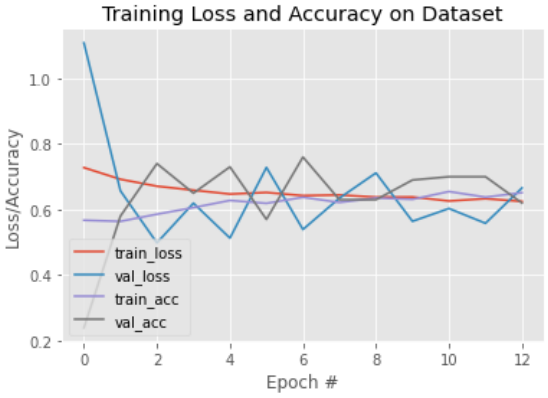
Layer (type)	Output Shape	Param #
input_2 (InputLayer)	(None, 224, 224, 3)	0
block1_conv1 (Conv2D)	(None, 224, 224, 64)	1792
block1_conv2 (Conv2D)	(None, 224, 224, 64)	36928
block1_pool (MaxPooling2D)	(None, 112, 112, 64)	0
block2_conv1 (Conv2D)	(None, 112, 112, 128)	73856
block2_conv2 (Conv2D)	(None, 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

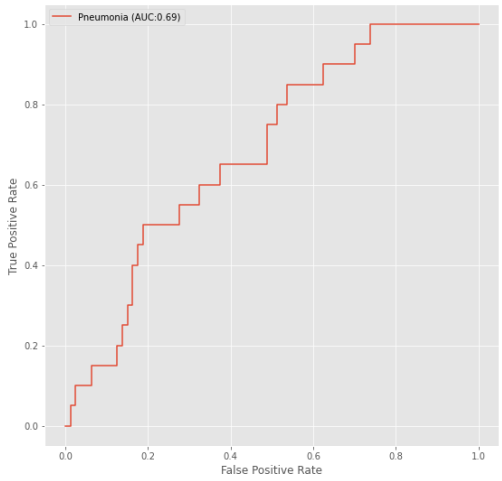
3. Algorithm Training

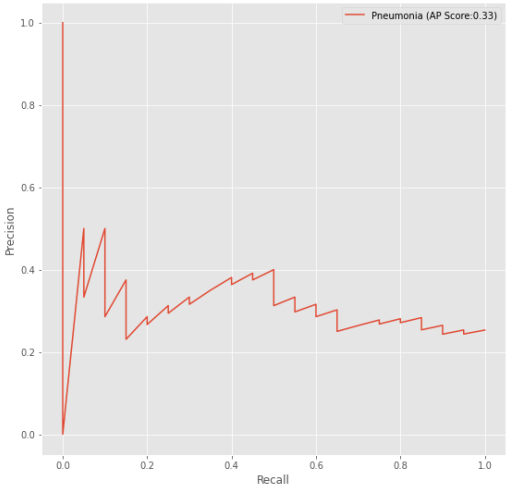
Parameters:
* Types of augmentation used during training: rescaling, horizontal_flip, height_shift, width_shift, rotation, shear and zoom.
* Batch size: 30
* Optimizer learning rate: adam, lr = 1e-4
* Layers of pre-existing architecture that were frozen transfer: layer is "block5_pool"
* Layers of pre-existing architecture that were fine-tuned: output layer
* Layers added to pre-existing architecture: Flatten, Dense (1024) with relu activation, Dense (512) with relu activation, Dense(1) with sigmoid activation.

<< Algorithm training performance visualization >>

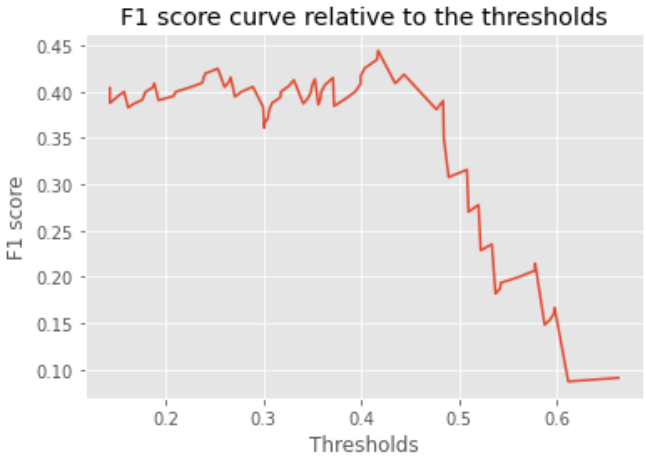


<< Insert P-R curve >>





Final Threshold and Explanation:
The threshold of 0.405 gives the best F1 score of 0.45. This means that if prediction is above 0.44 then we can say that it's a pneumonia. In fact, the trade-off between precision and recall is important because it depends on the clinical settings. High precision means that the number of relevant results is high, people detected with pneumonia will effectively have pneumonia in most cases. High recall on the other hand means that the number of FN will be low thus the number of patients having pneumonia that are detected is high.

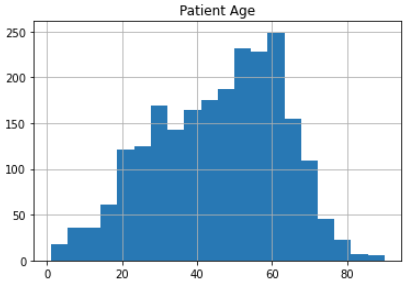


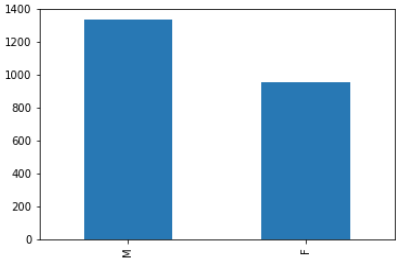
4. Databases

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

Description of Training Dataset:

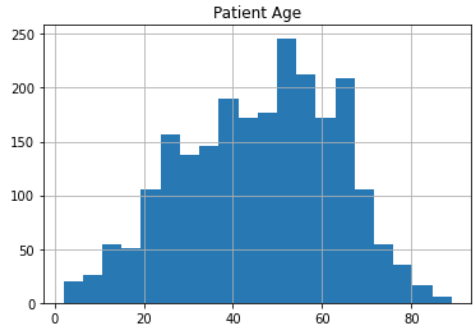
The validation dataset contains 2290 rows with information on the finding labels, the #of follow-ups, patient ID, age, gender, the view position and images parameters such as Width and Height. The training dataset was balanced to have an equal amount of positive and negative cases of Pneumonia.

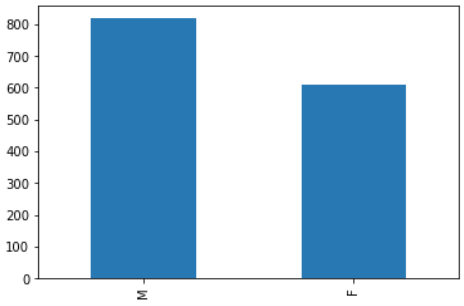




Description of Validation Dataset:

The validation dataset contains 1430 rows with information on the finding labels, the #of follow-ups, patient ID, age, gender, the view position and images parameters such as Width and Height. In the validation set, we balanced the dataset to have 20% of pneumonia case to have a 80/20 classic division.





5. Ground Truth

The ground truth was obtained using Natural Language Processing (NLP) to mine the associated radiological reports. The labels include 14 common thoracic pathologies such as Pneumonia, Pneumothorax, Consolidation and Infiltration. The biggest limitation of this dataset is that image labels were NLP-extracted so there could be some erroneous labels but the NLP labeling accuracy is estimated to be >90%.

6. FDA Validation Plan

Patient Population Description for FDA Validation Dataset:
A dataset representing the gender and age clinical representation for the FDA validation dataset. This dataset should also include a bigger number of comorbidities with pneumonia to investigate each comorbidity factor more in details. Finally, the algorithm worked poorly on pleural thickening and fibrosis thus a validation dataset shouldn't include these diseases

Ground Truth Acquisition Methodology:
Ground truth can obtained through biopsy labelling and checking to have a gold standard ground truth. However, due to time and cost constraints, a silver standard through a voting system of radiologists can be proved sufficient.

Algorithm Performance Standard: CheXNet on the same dataset obtained a F1 score of 0.435. The metric is used was the binary cross entropy loss to train the model. The algorithm obtained a F1 score of 0.44 which is better than the performance score of the radiologists with an average of 0.38.