FDA Submission

Your Name: Jack Warshaw

Name of your Device: Pneumonia Detection System

Algorithm Description

1. General Information

Intended Use Statement: This algorithm is used to predict the presence or absence of pneumonia in a chest Xray in a clinical setting. This would be used to fast track individuals who may have presence of pneumonia so that clinicians could confirm this through their own diagnosis.

Indications for Use: This algorithm is intended for use on patients under the age of 80 who have been administered a chest Xray in the AP or PA positions.

Device Limitations: This algorithm performs poorly when in co-occurrence with Edema, Infiltration, and Atelectasis, as it is sometimes unable to differentiate between the different conditions.

Clinical Impact of Performance: If this algorithm fails to predict pneumonia, it is possible that an individual with the condition may go longer with it untreated. If this algorithm incorrectly predicts that an individual has pneumonia, they will be incorrectly moved to receiving sooner treatment, which will result in a faster correct diagnosis. By plotting out our ROC for this algorithm we have determined it has an AUC of 0.68. Given that it has a slightly higher tendency for False Positives than other systems of its caliber, I recommend its screening is used in environments which do not have extremely high volumes of patients in need of immediate assistance, as this could result in overloading the number of patients who need to seek a more pressing diagnosis due to classifying individuals with Pneumonia who do not currently have it.

2. Algorithm Design and Function

Model: "sequential_2"

Layer (type)	Output Shape	Param #
model_2 (Model)	(None, 7, 7, 512)	14714688
flatten_2 (Flatten)	(None, 25088)	0
dropout_4 (Dropout)	(None, 25088)	0

dense_5 (Dense)	(None, 1024)	25691136
dropout_5 (Dropout)	(None, 1024)	0
dense_6 (Dense)	(None, 512)	524800
dropout_6 (Dropout)	(None, 512)	0
dense_7 (Dense)	(None, 256)	131328
dense_8 (Dense)	(None, 1)	257

Total params: 41,062,209
Trainable params: 28,707,329
Non-trainable params: 12,354,880

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

DICOM Checking Steps: Check that the BodyPartExamined is “CHEST”, the PatientPostion is either “AP” or “PA”, and the Modality is “DX”.

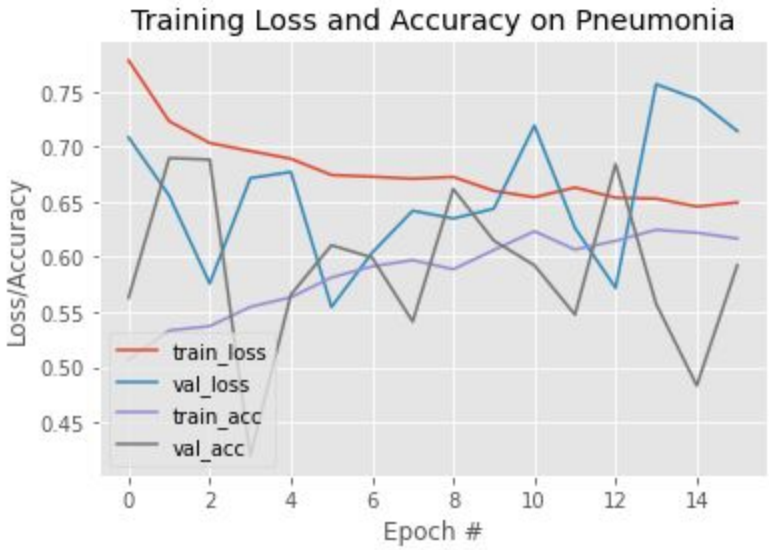
Preprocessing Steps: The images used on this algorithm must first be normalized using the image mean pixel intensity along with the standard deviation of the image’s pixel intensity. It is then put into a three dimensional array and reshaped to fit into the model.

CNN Architecture: The model is sequential, and has added VGG16 with a transfer layer of 'block5_pool'. There are 17 layers present in the premade model and additionally several layers have been fine tuned as well, starting with a Flatten layer and moving to an alternation of Dropout and Dense layers with relu activation (the values for the later of which goes from 1024 to 512 to 256), finishing with a Dense layer with sigmoid activation. None of the layers are frozen.

3. Algorithm Training

Parameters:

* Types of augmentation used during training:
    - Training data was rescaled, allows for horizontal flip (but not vertical flip), has a height shift range of 0.1, a width shift range of 0.1, a rotation range of 20, a sheer range or 0.1, and a zoom range of 0.1.
* Batch size:
    - 16
* Optimizer learning rate:
    - 1e-4
* Layers of pre-existing architecture that were frozen
    - None
* Layers of pre-existing architecture that were fine-tuned
    - 1, block5_pool
* Layers added to pre-existing architecture
    - 8, a Flatten layer and moving to an alternation of Dropout and Dense layers with relu activation (the values for the later of which goes from 1024 to 512 to 256), finishing with a Dense layer with sigmoid activation

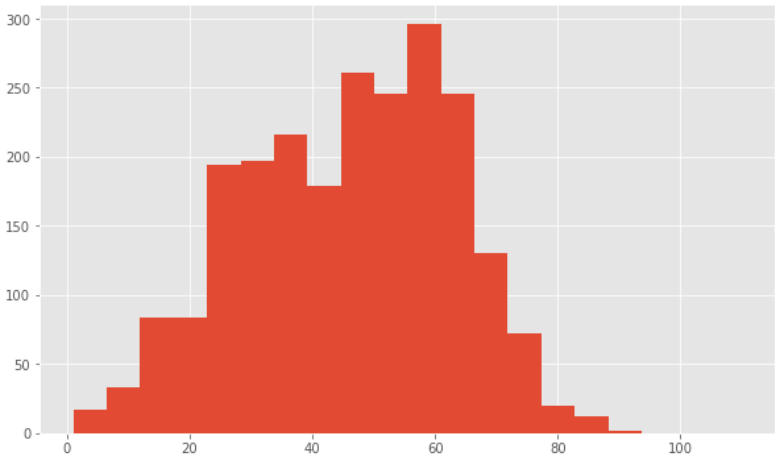


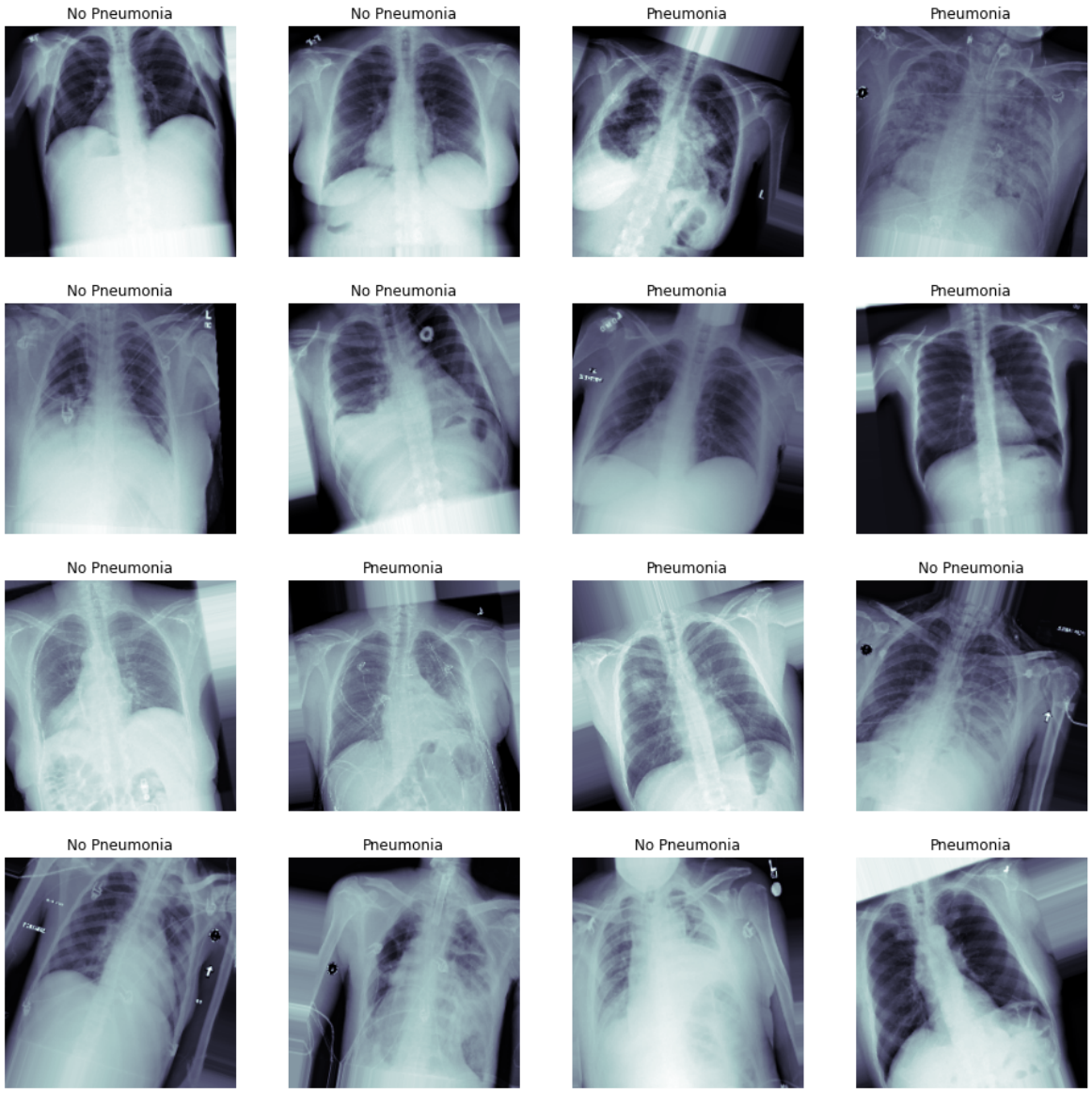
Final Threshold and Explanation: The final threshold is 0.5327289, which is linked to the f1 score of 0.4705882352941177.

4. Databases

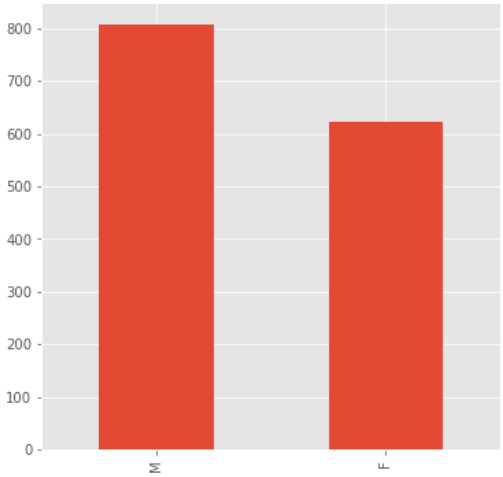
(For the below, include visualizations as they are useful and relevant)

Description of Training Dataset: The Training Dataset has undergone several alterations. This means that the view of the images in the training set may be slightly morphed in terms of image rotation, shifts in image size in the horizontal and vertical dimensions, and may also be slightly sheered or zoomed. All of the images in this dataset contain instances of pneumonia, and the dataset is comprised of 2290 validated images. This set contains an equal amount of positive and negative instances, with similar distributions of both age and gender as seen in our original dataset. This data was created by performing a train_test_split operation and then equalizing the number of both cases within the dataset. (Pictured below in order of appearance: graph of negative [0] vs positive [1] occurrences in training dataset, age distribution of training dataset, gender distribution of training dataset, and samples from training dataset)


300
250
200
150
100
50
0
0
20
40
60
80
100



Description of Validation Dataset: The Validation Dataset has not been altered other than a slight rescaling of 1/255. All of the images in this dataset contain instances of pneumonia. The validation dataset contains 1430 validated images. This set contains four times the number of negative instances than positive, with similar distributions of both age and gender as seen in our original dataset. This data was created by performing a train_test_split operation in which it was the test taking 0.2, and it was set to have 4 times the amount of negatives as positives at this juncture. (Pictured below in order of appearance: graph of negative [0] vs positive [1] occurrences in validation dataset, age distribution of validation dataset, and gender distribution of training dataset)



5. Ground Truth

The ground truth for this algorithm was determined by a unified weakly supervised multi-label image classification and disease localization framework, which was validated off a proposed dataset[1]. While this is a large source of verified data which means we have a larger array of data to work off within our data sets, any issues which may have been present within the classification of these images may be exacerbated in our own. For example, if certain types of images which may be under classified as pneumonia in the ground truth are used for our algorithm, it is possible that our algorithm may suffer the same pitfalls. However, the large number of classified and validated images allows for a more robust coverage of instances of pneumonia.

This dataset was originally gathered for use within a disease classifier known as ChestX-ray8, and in order to gather this information it utilized NLP on medical reports associated with said images. The NLP system which acquired this information had several frameworks in play for negation of these diseases in its setup, though this is still a challenge for NLP systems in general and may have affected elements of the data gathering (i.e. “the patient is clear of pneumonia” or “This is not normal inflammation”). Likewise, for the information gathering of this study, for an image to be marked as normal it cannot be containing any disease that would be marked on a report, not solely one of the 8 diseases of interest from this study. Therefore, we do not know what may happen if diseases beyond the 8 of interest in that study are passed through our model, as it was not trained to recognize such occurrences.

6. FDA Validation Plan

Patient Population Description for FDA Validation Dataset: This validation dataset should primarily include patients under the age of 80 who have been administered a chest Xray in the AP or PA positions. We would want a percentage of individuals with and without pneumonia in this set which is representative of the distribution of the disease in the total population.

Ground Truth Acquisition Methodology: The ground truth could be acquired through classifications performed by a group of radiologists to determine the presence of pneumonia in the validation dataset, to reduce error that may be present if a single radiologist were used.

Algorithm Performance Standard: Based on pre-existing literature[2], FDA regulations would require an F1 Score of at least 0.387, the radiologist standard average F1 score for pneumonia detection over four different radiologists sampled in the study for CheXNet (which showed an F1 score of 0.435). Other studies have shown F1 scores of 0.77 alongside Precision and Recall rates of 0.66 and 0.93 respectively, as seen in the ChestX-ray8 system.

---

[1]https://openaccess.thecvf.com/content_cvpr_2017/papers/Wang_ChestX-ray8_Hospital-Scale_Chest_CVPR_2017_paper.pdf

[2] https://arxiv.org/pdf/1711.05225.pdf