FDA Submission

Your Name: Justin Sheldon

Name of your Device: Deep Neural Network Pneumonia Detector

Algorithm Description

1. General Information

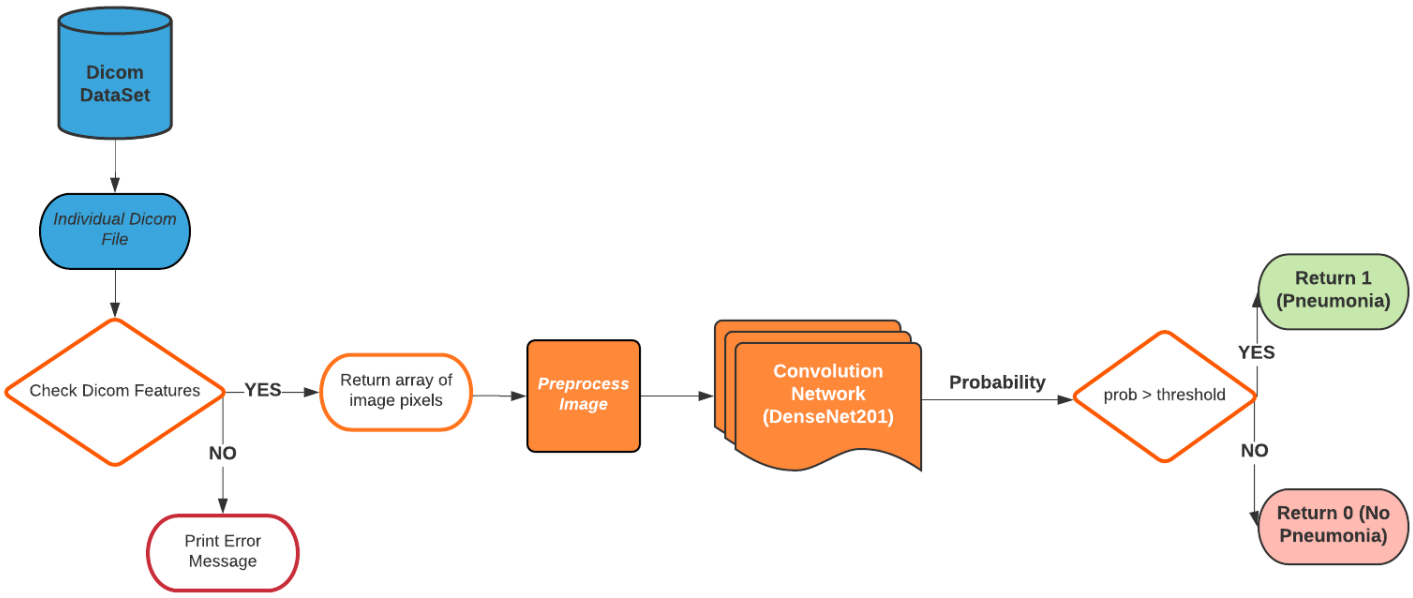
Intended Use Statement: Used to aid in detection of Pneumonia from Digital Radiography chest images (chest x-rays).

Indications for Use: This algorithm should be used to aid in detection of high risk patients for pneumonia so that these x-rays can be flagged as urgent for a radiologist to take a closer look. This algorithm could be employed as soon as the x-ray is processed. The algorithm must be used with Digital Radiography images that are of the patients chest area. It will work with both PA (Posterior->Anterior) and AP (Anterior->Posterior) views and can be used for patients of any age male or female. Details about the dataset that the algorithm was trained on can be found in the EDA notebook.

Device Limitations: Due to the severity of pneumonia we wanted to catch as many positive cases as possible without sacrificing too much precision. This lead to choosing a low threshold for classifying a case as positive. Due to this, the algorithm will have a high false positive rate (64.5% on our validation dataset). We believe this is an good trade off because this algorithm is only intended to be an aid to the radiologist so that they can view high risk cases as soon as possible. The radiologist will still be the one deciding if the patient has pneumonia or not, the algorithm just helps them see high risk cases sooner in their workflow.

Clinical Impact of Performance: Due to the severity of pneumonia it is important to detect high risk patients as soon as possible and this algorithm could help in alerting radiologist to high risk patients so that they can move these patients up in their queue of cases to look at.

2. Algorithm Design and Function



DICOM Checking Steps: To verify that the DICOM file is correct for our algorithm we need to make sure the image modality is equal to 'DX' (Digital Radiography), we need to check that the position of the patient during the scan is either 'PA' or 'AP', and we need to make sure that the body part examined is 'CHEST'.

Preprocessing Steps: The preprocessing for this model is very simple. All we do is rescale the pixel values from the range of 0-255 down to 0-1 by dividing each pixel by 255. This helps when training the network.

CNN Architecture: The model used is a pre-trained DenseNet201 architecture with a single fully connected output layer of size 1 to output a probability of pneumonia. The model was pre-trained on the 'imagenet' dataset. The entire DenseNet architecture is explained in this paper. An example of the model architecture from the paper is below.

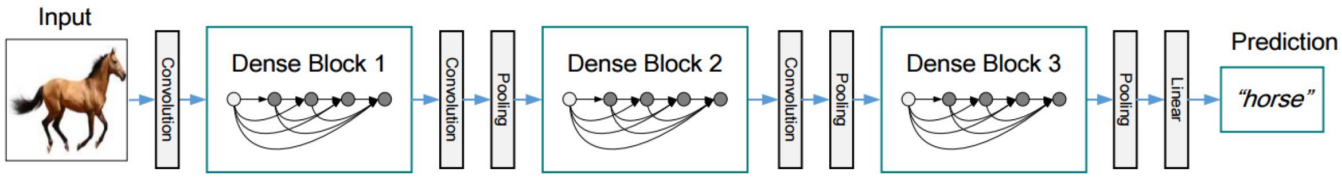
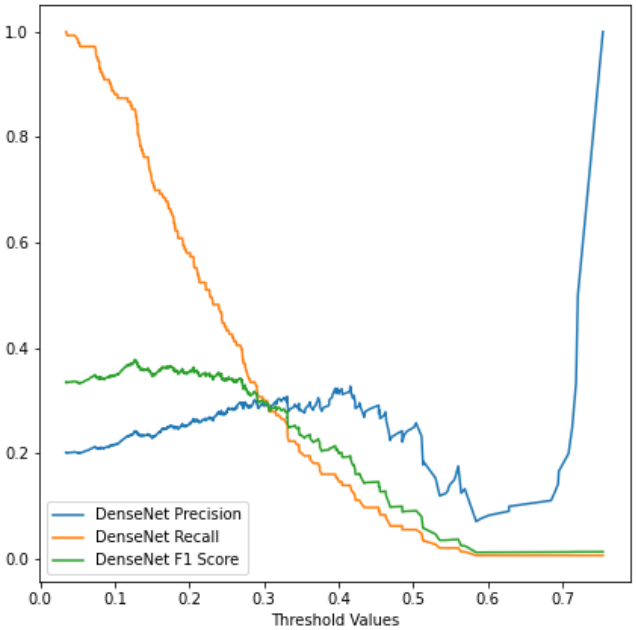


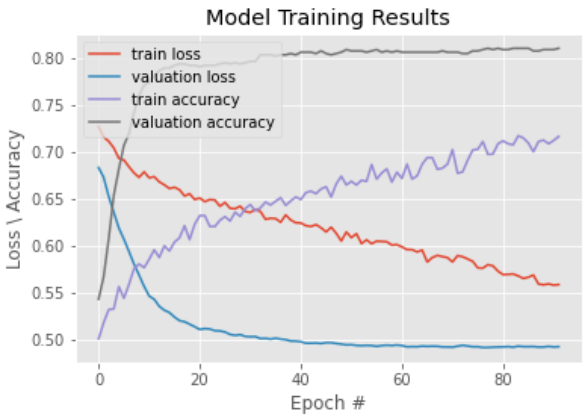
Figure 2. A deep DenseNet with three dense blocks. The layers between two adjacent blocks are referred to as transition layers and change feature map sizes via convolution and pooling.

3. Algorithm Training

Parameters:

- Types of augmentation used during training: horizontal flip, rotation_range=20, shear_range=0.1, zoom_range=0.15.
  - Horizontal Flip: This is similar to turning a 'PA' x-ray into a 'AP' x-ray.
  - rotation_range: Not all x-rays are going to have the body perfectly centered so this simulates that imperfection.
  - shear_range: Not all x-rays are going to have the body perfectly centered so this simulates that imperfection.
  - zoom_range: Not all x-rays are going to be the same distance from the scanner and so this simulates different distances.
- Batch size: 128
- Optimizer learning rate: 1e-5
- Layers of pre-existing architecture that were frozen: First 160
- Layers of pre-existing architecture that were fine-tuned: Last 40
- Layers added to pre-existing architecture: Single output layer changed from 1000 class classification to a single class classification.





Final Threshold and Explanation: Based on the indications of use statement, we would like this algorithm to have high recall in order to aid the radiologist with detecting high risk pneumonia cases early in their workflow. That being said we still would like some precision so that we are not sending too many false positives to the radiologist. Based on these factors and looking at the precision recall curves above, I used a threshold value of 1.35.

4. Databases

Dataset size: 112,120, Pneumonia Cases: 1431

Due to how uncommon Pneumonia is in the dataset we made some modifications to the training and validation sets. 90% of the positive cases went into the training set and the remaining 10% are in the validation set. The data to fill out the training and validation sets with negative cases was sampled at random from the original databases negative cases.

Description of Training Dataset: 50/50 split of positive/negative pneumonia cases.

- Training Set Size: 2576
- Positive Cases: 1288
- Negative Cases: 1288

Description of Validation Dataset: 20/80 split of positive/negative pneumonia cases.

- Validation Set Size: 715
- Positive Cases: 143
- Negative Cases: 572

5. Ground Truth

This NIH Chest X-ray Dataset is comprised of 112,120 X-ray images with disease labels from 30,805 unique patients. To create these labels, the authors used Natural Language Processing to text-mine disease classifications from the associated radiological reports. The labels are expected to be >90% accurate and suitable for weakly-supervised learning. Using NLP to create ground truth labels for the images is beneficial because of the massive volume of x-rays you can classify by just having the associated radiologist report. To do this by hand would be much more expensive and time consuming. The downside of this is that this method is not 100% accurate and you could end up with inaccurate classifications that confuse your models.

6. FDA Validation Plan

Patient Population Description for FDA Validation Dataset: The patient population to use for the validation set could be anyone that would typically go to hospital and be checked for pneumonia. This includes males and females of any age. We do not need to exclude any diseases from this population because the goal is to be able to detect pneumonia from a chest x-ray independent of other factors. It would be necessary to have at least 10% of the chest x-rays to include pneumonia so that we can evaluate the algorithms performance. In order for the model to work the images need to be chest x-rays (Digital Radiography of the chest area) because that is what the model was trained on.

Ground Truth Acquisition Methodology: To obtain optimal ground truth labels it would be very beneficial to have multiple radiologist classify the chest x-rays in the validation set for pneumonia. If using an odd number of radiologist go with the classification that gets the most votes. If using an even number of radiologist break a tie by considering years of experience as a radiologist. For example, if using 4 radiologist and the final vote is 2/2 then sum up the radiologist years of experience for each classification and choose the group that has more years of experience.

Algorithm Performance Standard: The goal is to be able to accurately classify pneumonia from only looking at the x_ray image with a reasonable level of accuracy. Finding a 'reasonable level' of accuracy for this task is more complicated than just considering how often the model correctly classifies pneumonia vs not pneumonia. Pneumonia is rare in our dataset (1.276%) so a naive model could just classify every image as not pneumonia do very well (98.724% Accurate). Therefore we need to use other metrics like Recall, Precision, and F1 Score to evaluate the model. This paper goes into detail on finding a good baseline accuracy. According to this paper the average radiologist F1 score for detecting pneumonia from x_ray images with no other information about the patients is 0.387 (based on the radiologist they used for this study). We will use this as our performance standard.

- F1 Score: 0.387, where F1 = (2 * precision * recall)/(precision+recall)
- Recall = True Positives/(True Positives + False Negatives)
- Precision = True Positives/(True Positives + False Positives)