FDA Submission

Your Name: Matthias Kirmse

Name of your Device: XPneumo

Algorithm Description

1. General Information

Intended Use Statement:
The algorithm is intended to be used for the identification of pneumonia from chest X-ray images.

Indications for Use:
This algorithm is intended for use on women and men in the age between 1 and 95, which have an X-ray of the chest from a posteroanterior or anteroposterior viewing angle.

Device Limitations:
The algorithm can only be used for the specified condition and populations. It was not trained to detect other diseases or pneumonia from other viewing angels.
The detection performance is mostly independent from other conditions except there are lower recall rates for patients with mass or nodules and the f1 score is also lower for patients with pneumothorax due to the lower precision value.

Clinical Impact of Performance:

The algorithm should be applied in the clinical workflow after DICOM images are generated. The high recall configuration we chose makes it possible to use the algorithm as a prioritization tool, putting cases positively identified on the top of the queue ranked by their probability. Additionally, the radiologist would be given the classification and probability value after his own diagnosis.

In this setting, false positives would mean that the radiologist would potentially have an even closer examination, which can be beneficial but also costs additional time. In the worst case false positives could lead to unnecessary treatment. False negatives on the other hand could have more severe consequences if not caught, naimly potentially missing critical treatment. For this reason, the algorithm should only be used as a supporting tool to the radiologist but not be solely relied on.

2. Algorithm Design and Function

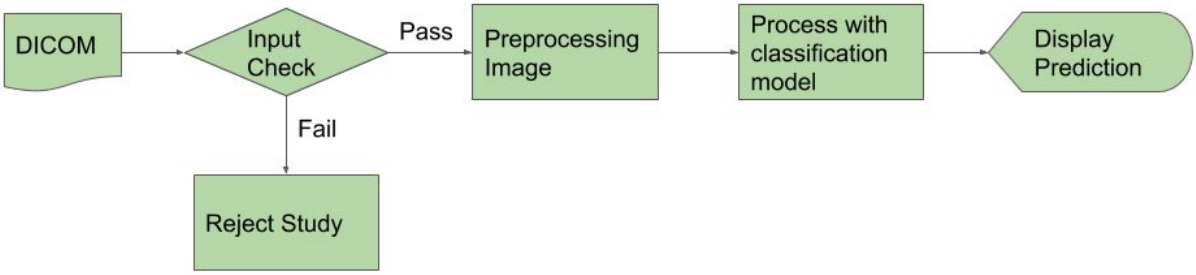

Figure 1. Algorithm Flowchart

DICOM Checking Steps:
The algorithm first reads in the DICOM file and checks if the right modality, age, position and body part are present. Otherwise the study is rejected.

Preprocessing Steps:
The images are then normalized using the pixel mean and standard deviation of the image.

CNN Architecture:
The normalized image is subsequently processed by a deep learning model outputting the probability for pneumonia. More precisely, we use a VGG16 model pretrained on ImageNet. For this network, the first 17 layers were not updated during training. On top there were 3 fully connected layers with 1024, 512, 256 and 1 node with intermediate dropout layers.

3. Algorithm Training

Augmentation:
We used the following image augmentation parameters to increase the train variability:

- horizontal_flip = True
- height_shift_range = 0.05
- width_shift_range = 0.05
- rotation_range = 5
- brightness_range = (0.95,1)
- shear_range = 0.05
- zoom_range= 0.05

Model Architecture:
The first 17 layers of the pretrained VGG-16 network were frozen and only the last convolutional layer (Block5) was used as transfer layer and fine tuned. Additionally, we added a layer to flatten the input followed by 1024, 512 and 256 sized dense layers with ReLu activation. These were connected with dropout layers with a rate of 0.5. Finally, a dense layer followed with size 1 and a sigmoid activation function providing the class probability.

Training Parameter:
As batch size we used 64 as a tradeoff of memory requirements and training stability. The optimizer was trained with a constant learning rate of 1e-5. We applied early stopping based on the validation loss with a patient parameter of 10 epochs.

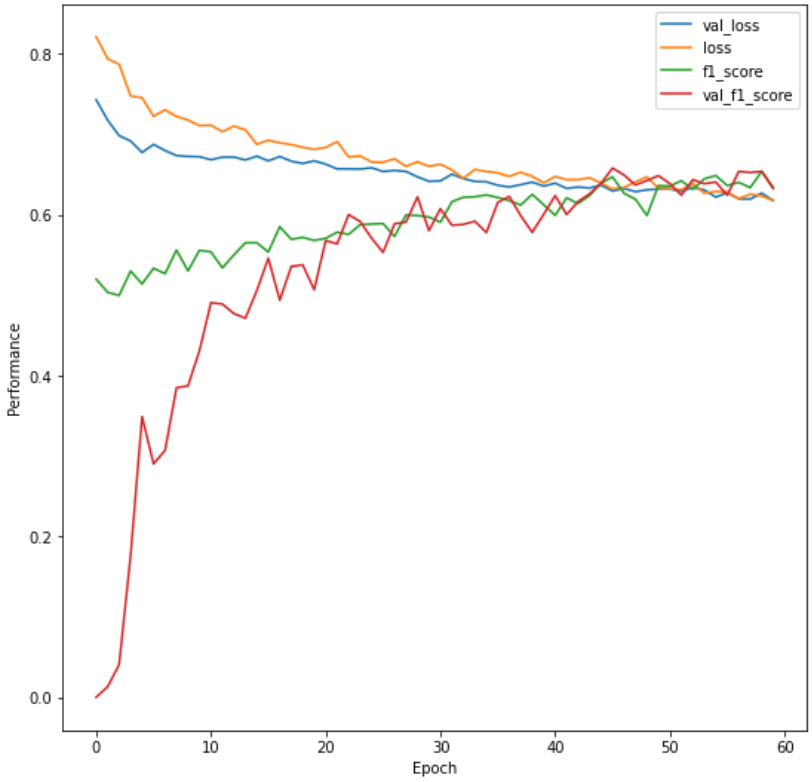

Figure 2. Model training history

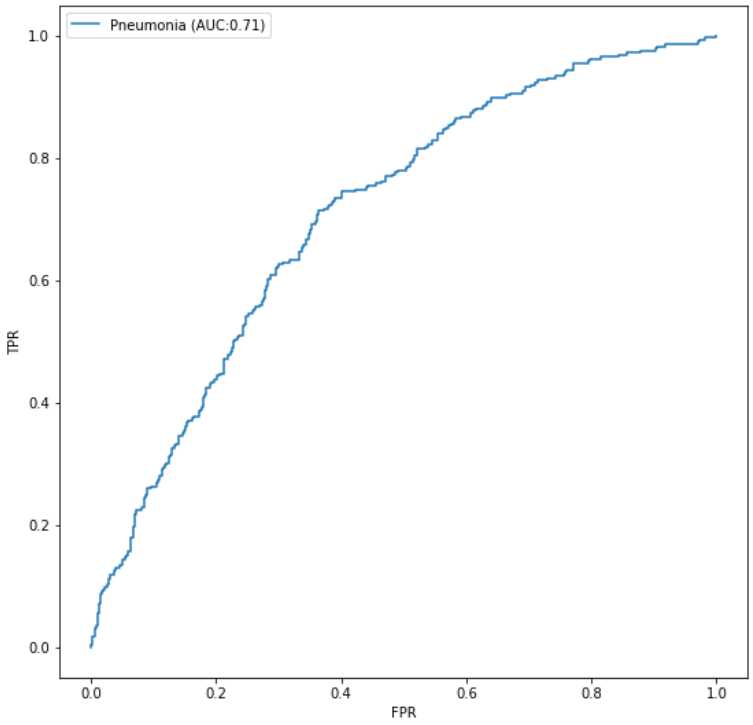

Figure 3. ROC curve with AUC on test set.

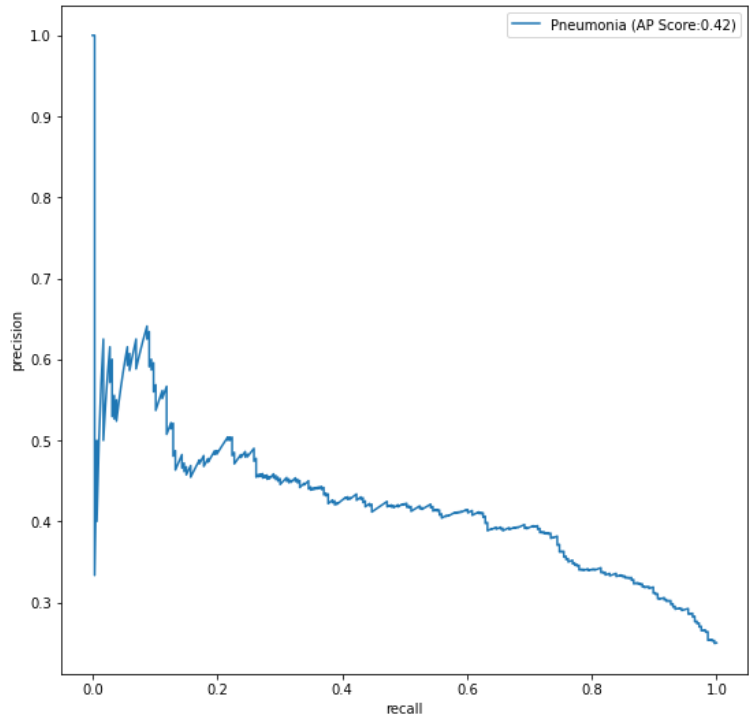

Figure 4. Precision-recall curve on test set.

Final Threshold and Explanation:

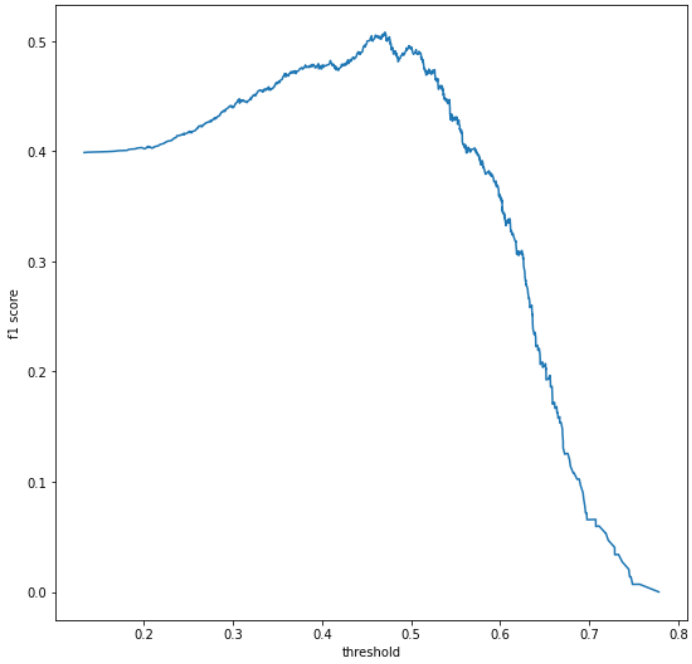

Figure 5. Threshold vs f1 score.

After finishing the training we took the model with the best loss on the validation set of 0.636. This was then applied to the test set to get an unbiased estimate. The resulting test set AUC was 0.71.

To choose the final threshold we first looked for the optimal f1 score of 0.508, which would yield a recall of 0.713 and a precision of 0.395. However, to be more useful in our scenario we chose the threshold 0.353, increasing the recall to 0.9 while giving a precision of 0.311 and overall f1 score of 0.462.

4. Databases

As base for our training and validation we used the NIH chest X-ray dataset containing 112,120 X-ray images from 30,805 patients. After preprocessing the data set, the patient population consisted of male and female patients between 1 and 95 years old. The images were captured using Digital Radiography (‘DX’) with posteroanterior and anteroposterior view angles. Out of all X-rays (after preprocessing) 1430 were labeled as pneumonia and 110674 not. Thereby, pneumonia findings were mainly comorbid with infiltration(605), edema(340), effusion(268), atelectasis(262) and consolidation(123).

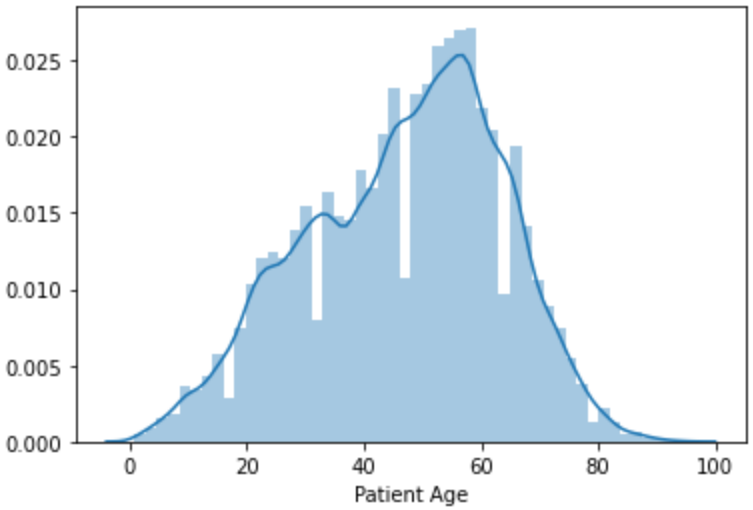

Figure 5. Age distribution in NIH dataset.

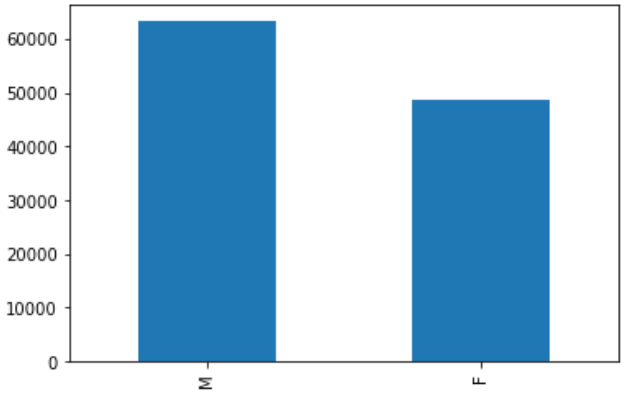

Figure 6. Gender distribution in NIH dataset.

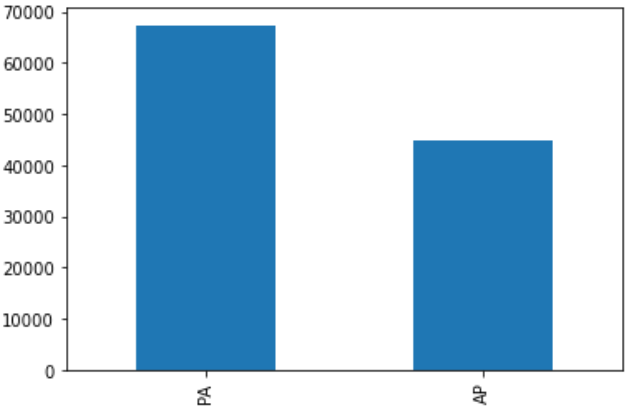

Figure 6. Distribution of viewing angle in NIH dataset.

Description of Training and Validation Dataset:

We randomly split the data set in 80% used for training and 20% used for later testing. The random split was stratified based on the pneumonia label guaranteeing an equal distribution. Then we balanced the training set to have equally many positive and negative examples. For image augmentation we used the parameters described above.
The training set was again split in a 90% portion used for training the model and a 10% independent validation set to determine the early stopping criterion.

Description of Test Dataset:
20% of the data were used for the independent test set in order to obtain an unbiased estimate of the algorithm performance. Here, we balance the class distribution 1:3 as expected in a clinical setting. For the test set, no image augmentation was applied.

5. Ground Truth

The image labels for the NIH dataset were extracted from radiologist reports using NLP methods. This accuracy is estimated to be >90%.

6. FDA Validation Plan

Patient Population Description for FDA Validation Dataset:
The FDA validation set should include digital radiography of male and female patients between the age of 1 and 95. They should be captured from posteroanterior and anteroposterior view angles.

Ground Truth Acquisition Methodology:
As described in the CheXNet paper (https://arxiv.org/pdf/1711.05225.pdf), the ground truth could be obtained by getting a group of practicing radiologists to label the examples and take for example a majority vote weighted by their experience. Additionally the patient's clinical history as well as sputum cultures could be used to further improve the ground truth.

Algorithm Performance Standard:
The algorithm should be rated based on its f1-score compared to the average f1-score of the radiologists used to acquire the ground truth. If it is within a 95% confidence interval, the algorithm could be seen to perform equally well than an average radiologist on this task.