FDA Submission

Johannes Rippberger

Name of your Device: Pneumonia-Checker - xRay edition

Algorithm Description

1. General Information

Intended Use Statement:

The AI solution at hand is intended to support radiologist decision making in the clinical setting. It should be applied to pre-qualify image data for radiologists to ease their diagnosis process. The "Pneumonia-Checker - xRay edition" is, as also indicated by its name, a solution specifically intended to support the diagnosis for Pneumonia from 2D xRay images. It is intended to be used on patients of age between 18 and 72.

Indications for Use:

"Pneumonia-Checker - xRay edition" could also be applied to prioritize patients in the clinical queueing of radiologists. As Pneumonia is a critical condition, potentially lethal especially for older patients, a quick treatment is key. Furthermore the solution could also be applied for patients outside the intended age scale.

Device Limitations:

The AI solution was trained on image data of patients where

- 90% are falling into the intended age limitations of 18 to 72 years of age, 10% outside those limitations.
- 58% of patients where male and only 42% female.

Device is not intended to be used for diagnosis of any other diseases than Pneumonia.
Device is not intended to be used for diagnosis from any other images than 2d-xray-medical images.

Clinical Impact of Performance:

The Pneumonia-Checkers good performance on Recall is specifically important, as it is critical for patients if occurrence of Pneumonia is not recognized by the solution (Case of a 'False Negative'). Here Pneumonia-Checker achieves over 75% of recall.

Precision in identifying positive cases and the identification of patients showing Pneumonia though they are not really having this disease (case of 'False Positives') is less important for Pneumonia-Checker, as the negative impact on patients is rather small.

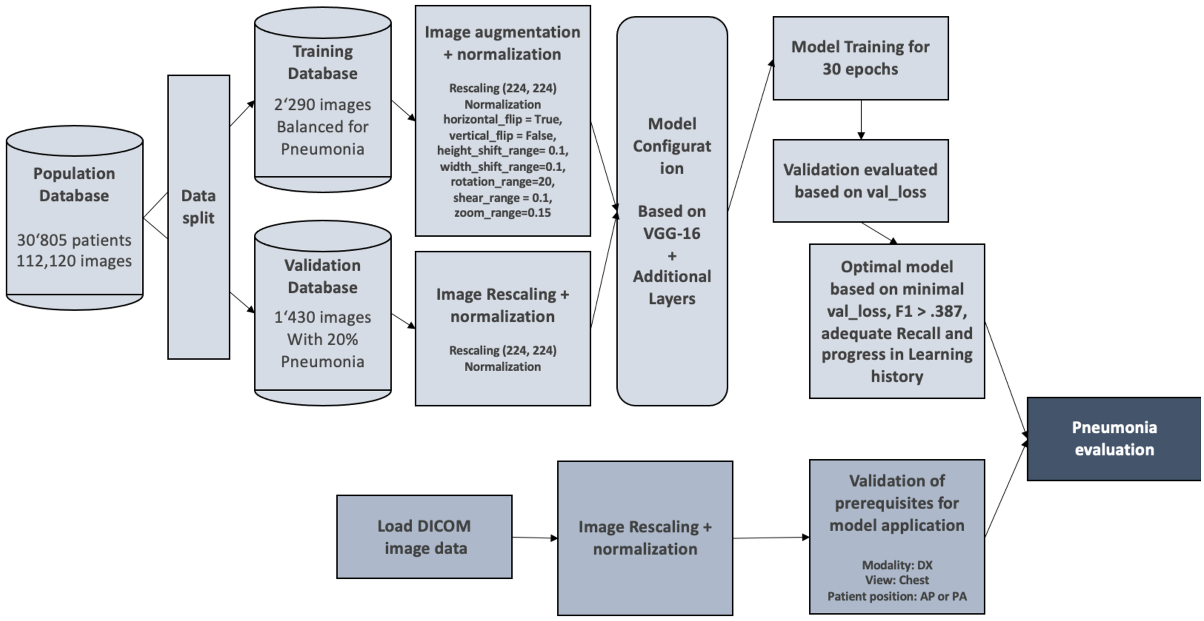
Therefore it is required to always have a radiologist reevaluate the results of Pneumonia-Checker. It should solely support Radiologists to ease their workflows.

Technical performance to be regarded:

AI solution can be operated on a computer with a standard CPU to fulfil its purpose, though more computing power (e.g. via a GPU) can speed up the computing process.

2. Algorithm Design and Function

Algorithm development flowchart:



DICOM Checking Steps:

Test-sample based validation of capability to ensure

- body part examined was CHEST (Error message: Body part XX, for Pneumonia-Check needs to be Chest)
- imaging modality was DX (Error message: Modality XX for Pneumonia-Check needs to be DX)
- view positions of picture taken were AP or PA (Error message: Patient position XX, for Pneumonia-Check needs to be AP or PA)

For Age distribution (see Intended Use Statement) a message is shown to draw users attention to the fact that age limitations are exceeded (Message: ATTENTION! Patient Age XX outside Pneumonia_Check required age range between 18 and 72).

Preprocessing Steps:

Preprocessing via

- Image normalization
- Image rescaling

CNN Architecture:

The model is based on the standard VGG16 model, trained on the ImageNet database.
The model uses the first 16 layers of the VGG16 model.
The VGG16 model output is flattened and passed through several additional dense and dropout layers to optimize performance (see table below)

```
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #   
=================================================================
model_2 (Model)              (None, 7, 7, 512)         14714688  
_________________________________________________________________
flatten_1 (Flatten)          (None, 25088)             0         
_________________________________________________________________
dropout_1 (Dropout)          (None, 25088)             0         
_________________________________________________________________
dense_1 (Dense)              (None, 1024)              25691136  
_________________________________________________________________
dropout_2 (Dropout)          (None, 1024)              0         
_________________________________________________________________
dense_2 (Dense)              (None, 512)               524800    
_________________________________________________________________
dropout_3 (Dropout)          (None, 512)               0         
_________________________________________________________________
dense_3 (Dense)              (None, 256)               131328    
_________________________________________________________________
dropout_4 (Dropout)          (None, 256)               0         
_________________________________________________________________
dense_4 (Dense)              (None, 1)                 257       
=================================================================
Total params: 41,062,209
Trainable params: 28,707,329
Non-trainable params: 12,354,880
```

3. Algorithm Training

Parameters:

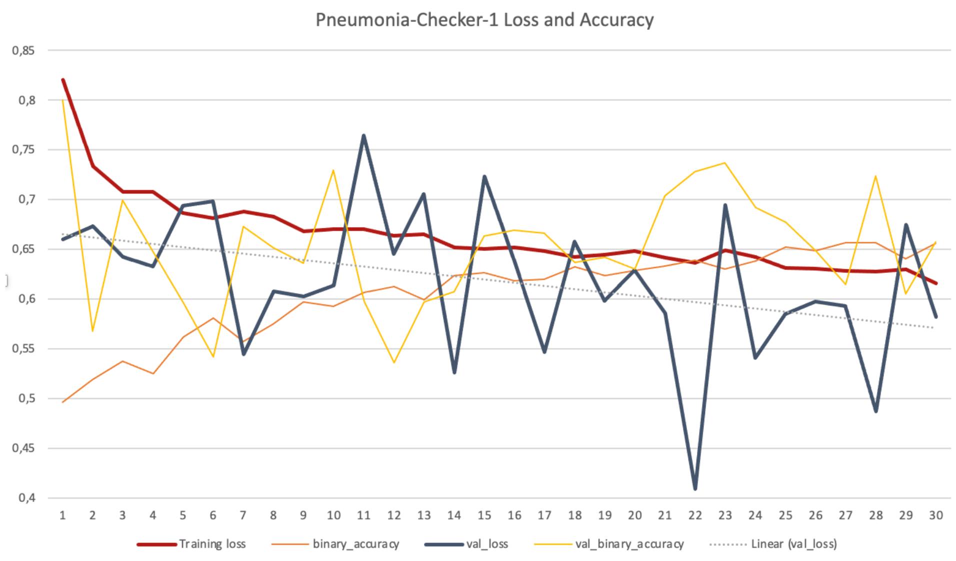
Image augmentation is applied for training data according to the following parameters:

- Horizontal flips: True
- Vertical flips: False
- Height shift range of 0.1,
- Width shift range of 0.1,
- Rotation range of 20,
- Shear range of 0.1,
- Zoom range of 0.15.

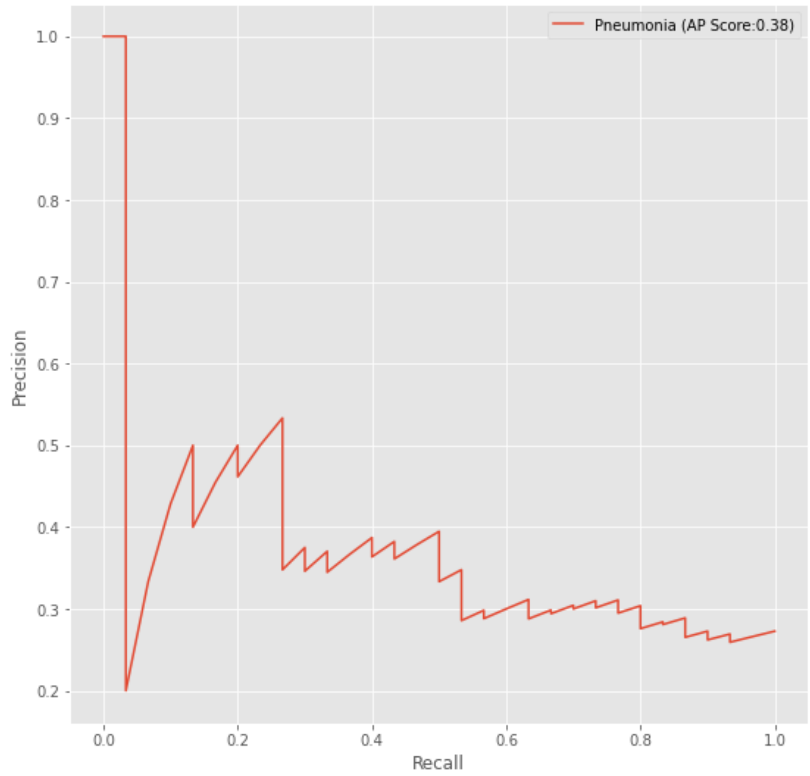
Batch size: 128 (for either Training and Validation)

Adam Optimizer with learning rate: 1e-4

As indicated in the visualization below the Validation_loss improved along training for 30 Epochs.
Pneumonia-Checkers's final val_loss was measured at 0.409.

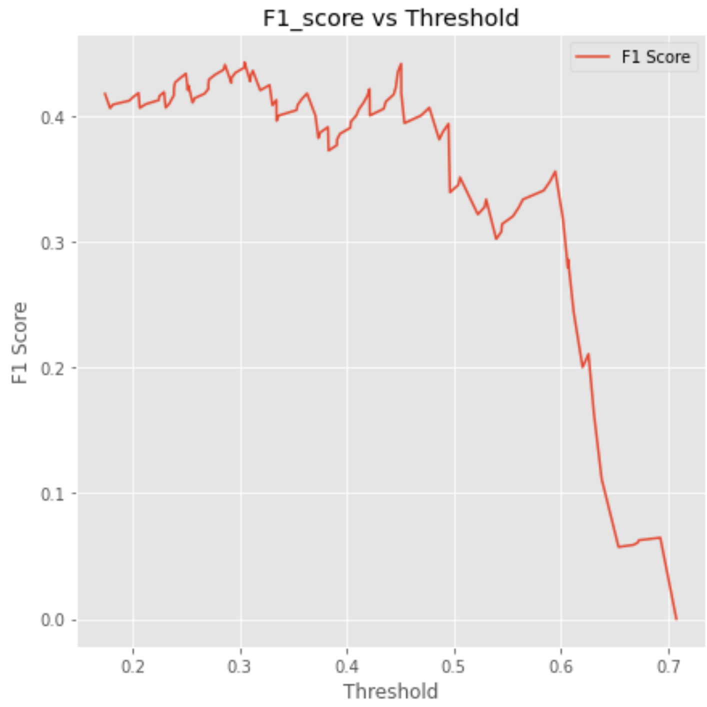


The model's Precision – Recall curve is shown below.



Final Threshold and Explanation:

Threshold: 0.31
F1-score: 0.441



The final threshold of 0.31 was based on the highest F1 Score of 0.441. Based on literature a baseline for average radiologist performance is an F1 Score of 0.387. This means Pneumonia-Checker is capable to achieve a better performance to identify existence of Pneumonia than the average radiologist.

For the threshold of 0.31 the Recall is measured at 0.767, which should allow us to be confident that many False Negative results can be avoided in our model. As indicated in "Clinical Impact of Performance" the avoidance of False-Negative results is critical, to ensure patients suffering from Pneumonia are treated accordingly.

4. Databases

Description of Training Dataset:
Training dataset was comprised of

- 2'290 training images sampled from 112'120 medical images of 30'805 patients
- Balanced 50-50 split between pictures labelled with Pneumonia and without showing Pneumonia (labelled No-Pneumonia).
- Training dataset was augmented before utilization (see "Algorithm Training")
- Training data was reflected to be adequate representation of population regarding
    - Age distribution
    - Gender distribution
    - Patient position on image taken
    - Co-Occurrence of diseases

Description of Validation Dataset:
Validation dataset was comprised of

- 1'430 testing images sampled from 112'120 medical images of 30'805 patients

- Real-Life focused dataset split 20-80 between pictures labelled with Pneumonia and without showing Pneumonia (labelled No-Pneumonia)
- Validation data was reflected to be adequate representation of population regarding
    - Age distribution
    - Gender distribution
    - Patient position on image taken
    - Co-Occurrence of diseases

Overlap of imaging data between samples for Testing and Validation data was avoided.

5. Ground Truth

Ground Truth was established from 112,120 images of 30,805 patients labeled for disease occurrence via a Natural Language Processing process from Radiologist reports. This has the following limitations:

- As Pneumonia and other chest-area diseases are rather challenging to identify form chest x-rays even for well-trained radiologists an error can occur in the initial diagnosis of the disease. Literature tells us that radiologists label imaging data correctly for Pneumonia in 0.387 of the cases in average. As Ground Truth is based on radiologists reports this needs to be considered.
- The NLP based extraction of disease labels from image data has an estimated accuracy of correct labelling of 90%.

6. FDA Validation Plan

Patient Population Description for FDA Validation Dataset:
Clinical validation of the "Pneumonie-Checker - xRay edition" should be based on the following criteria:

- Patient Age distribution between 18 and 72.
- Male patients 58%.
- Chest X-Ray taken in AP or PA position.
- Chest X-Ray image in DICOM format.

Ground Truth Acquisition Methodology:
As Ground Truth the silver standard approach of using several radiologists should be used since identifying Pneumonia is difficult even for well-trained radiologists.

Algorithm Performance Standard:
The algorithm’s F1 score should be more than the average performance of the radiologists with an F1 score of .387 (based on literature) to provide a valid improvement over today's situation.