FDA Submission

Udo Dehm

November 3, 2020

Name of Device: PneumoniaXNet

1 Algorithm Description

1.1 General Information

1.1.1 Intended Use Statement

Assisting radiologists in the detection of pneumonia on X-ray chest images. It is explicitly stated that the AI based algorithm PneumoniaXNet is intended to be used under the supervision of an expert like a radiologist or other clinician with expert knowledge.

1.1.2 Indications for Use

Automated detection of pneumonia from chest X-rays in non-emergency clinical settings. The device PneumoniaXNet is intended to use as a diagnostic tool if the following conditions are fullfilled:

- Predictions on male and female patients (see figure 7 for data distribution) between the ages of about 10 to 90 years (see figure 6 for data distribution).
- Chest X-ray view positions: posteroanterior (PA) view and/or erect anteroposterior (AP) chest view (see figure 8 for data distribution).
- Optionally (on rare occasions): Analysing chest X-rays in regions with inadequate access to diagnostic imaging specialists. Getting a rough idea of the disease of a patient is better than nothing.

Description of a possible clinical setting:

- Obtaining X-ray image of a patient's chest (PA or AP view).
- Sending scan in DICOM format to a remote server with installed PneumoniaXNet software for processing.
- Checking the DICOM file for compatibility (see 1.2.1).
- If the scan passes the compatibility check it is preprocessed (see 1.2.2).
- Feed the X-ray image into the machine learning algorithm PneumoniaXNet.
- The PneumoniaXNet classifier will output one of two diagnostical predictions: patient has pneumonia (1) or patient has no pneumonia (0).
- The result is sent to a radiologist who will validate the result and compile a final diagnosis.

1.1.3 Device Limitations

- The PneumoniaXNet device does not achieve 100% accuracy. Therefore it is advised that this classifiers predictions are only used as a supplementary diagnosis tool. The final diagnosis should always be compiled by an expert like a radiologist.
- PneumoniaXNet should be run on a CUDA capable GPU (e.g. local server or cloud server with GPU access). This is especially important if the algorithm is used in situations where getting a result quickly is important

1.1.4 Clinical Impact of Performance

- Enhancing workflow by providing fast and reliable pneumonia detection from X-ray images
- If algorithm predicts a positive pneumonia case, a radiologist can prioritize to analyse this case more urgently. The patient could be treated sooner.
- It is strongly recommended that the device is used as an assisting device for an imaging specialist. If the algorithm predicts a false positive (FP) the radiologist can always intervene and prevent a patient from a useless pneumonia treatment. The only downside of a false positive should be some wasted time. If the algorithm predicts a false negative

(FN) this could lead to a loss of time for the patient's treatment. A trained radiologist should detect the pneumonia disease in this sirious case. False negatives are much more severe than false positives. Therefore the PneumoniaXNet device tries to prevent false negatives by adjusting the classifier threshold (see section 1.3.2).

1.2 Algorithm Design and Function

1.2.1 DICOM Checking Steps

Before a DICOM file is preprocessed and fed to the PneumoniaXNet algorithm it is checked if it contains the correct properties. The following things are checked:

- Modality must be 'DX'.
- Body part must be 'chest'.
- Patient position must be 'PA' or 'AP'.

1.2.2 Preprocessing Steps

The following preprocessing steps are performed on each image before it is fed into the PneumoniaXNet algorithm:

- resizing image to spatial size 224x224 pixels
- convert image from grayscale to RGB with dimensions 1x224x224x3 (the first dimension is the batch size, multiple images can be fed simultaneously in batches into the algorithm)
- DenseNet121 specific preprocessing steps: input pixel values are scaled between 0 and 1 and each channel is normalized with respect to the ImageNet dataset

1.2.3 CNN Architecture

describe the architecture of the classifier We use the Keras implementation of the DenseNet121 model as base model. We cut the last few (dense) layers of the net and replace them with the following additional layers:

- Global average pooling layer
- dense layer with 1024 neurons and ReLu activation with dropout 0.5

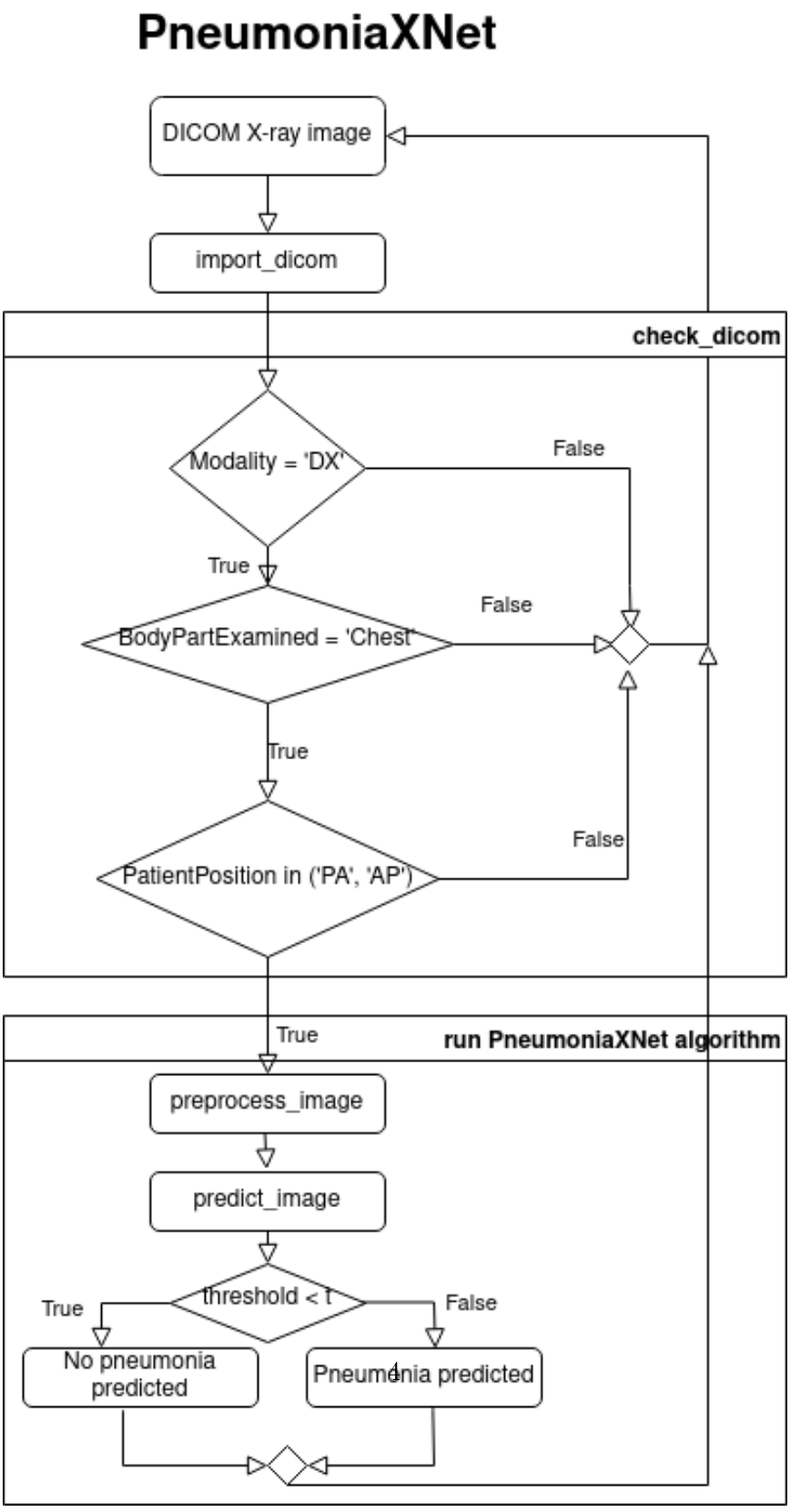

Figure 1: Flowchart of the PnuemoniaXNet algorithm (inference).

- dense layer with 512 neurons and ReLu activation with dropout 0.5
- dense layer with 256 neurons and ReLu activation
- dense layer with 1 neuron and sigmoid activation (output layer)

We freeze all trainable parameters of the base net except of the parameters of the last 7 layers. These parameters in addition to the parameters of the additional layers above sum up to the trainable parameters. In total we have 1,872,129 trainable parameters. This so called transfer learning is a common and efficient way of training convolutional neural networks. The imported DenseNet121 model has previously been trained on the ImageNet dataset.

1.3 Algorithm Training

1.3.1 Parameters used for training:

- Batch size: 16 images
- Optimizer learning rate: Adam optimizer with initial learning rate of 0.001
- image augmentation used during training:
  - horizontal flip: useful because we have X-ray images from both viewing positions PA and AP. The algorithm will be able to predict pneumonia presence from both viewing positions.
  - rotation range 10 degrees
  - height shift range 0.1
  - width shift range 0.1
  - shear range 0.1
  - zoom range 0.15

The model performance and training progress can be seen in figures 2, 3, and 4

1.3.2 Final Threshold and Explanation

Our goal is to minimize the false negative (FN) predictions. Since the F1 score is proportional to 1/FN we are interested in maximizing the F1 score. Furthermore, the F1 metrics enables us to compare our results with previous work (see 1.6.3).

To maximize the F1 score we chose the prediction threshold 0.216 (see figure 5).

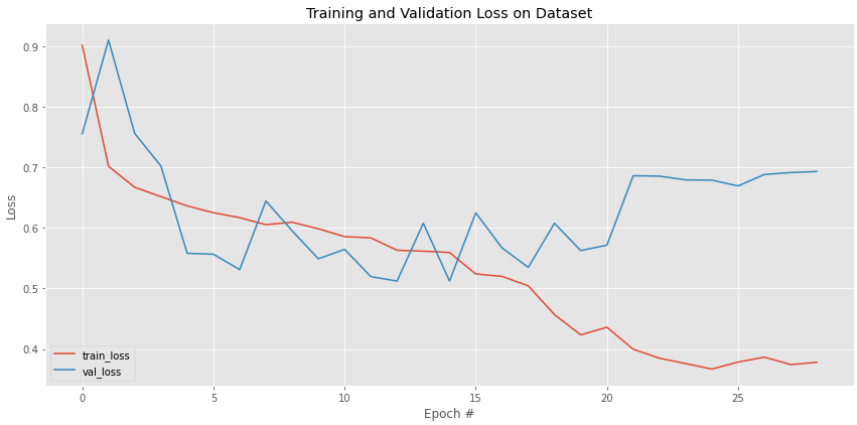

Figure 2: PnuemoniaXNet training performance: Training and validation losses.

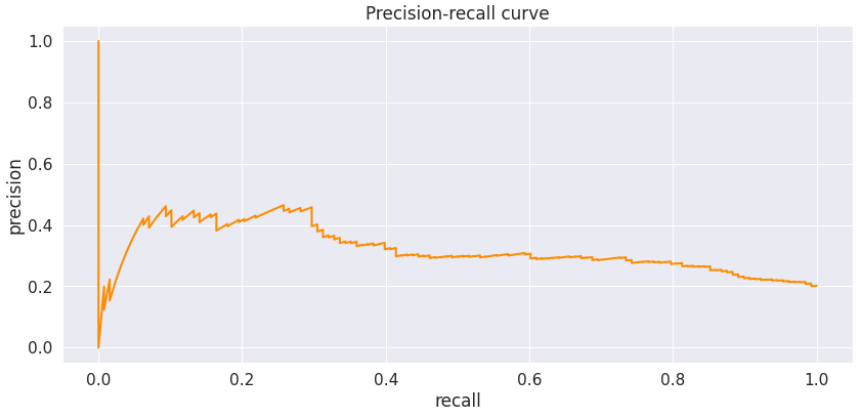

Figure 3: Testing set precision-recall curve

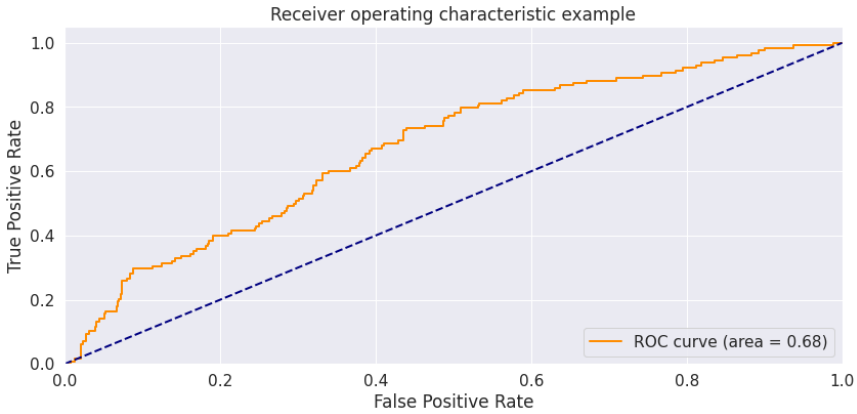

Figure 4: Testing set ROC curve

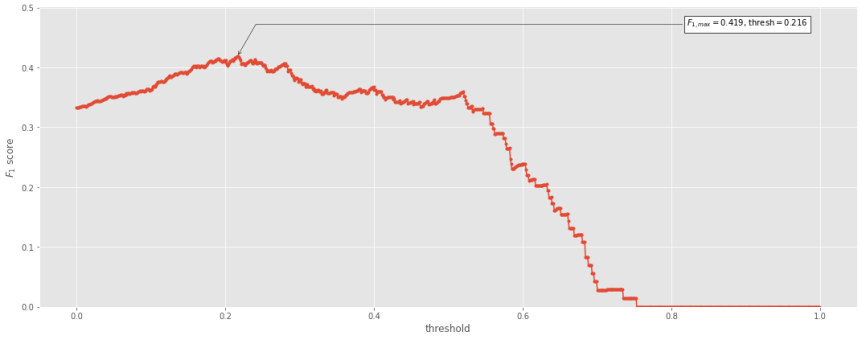

Figure 5: Threshold selection based on F1 metric.

1.4 Databases

We train the PneumoniaXNet algorithm on the National Institutes of Health Chest X-Ray Dataset. This dataset is comprised of 112,120 X-ray images from CT scans with disease labels from 30,805 unique patients. It was not designed specifically for detecting pneumonia disease. It also contains other diseases. Patients might have multiple diseases simultaneously (see figures 9 and 10). Altogether, it contains 15 classes with 14 diseases and one class for 'no findings' (no disease of the 14 diseases in this dataset). The patients' age, gender and viewing position of the X-ray images are depicted in figures 6, 7 and 8. The prevalence of the most common diseases in the dataset is visualized in figure 9. When randomly splitting the dataset into training, validation and testing sets it is ensured that the ratios in all demographics are roughly maintained.

We preprocess the NIH chest X-ray dataset before we split it. For this we convert all patient ages to the unit year and delete all patients with age > 100 from the dataset (this is the case for 16 data points). The adjusted dataset contains 1430 images with pneumonia disease labels which corresponds to ~1.3% of all data points. We can say that according to the

1.4.1 Description of Training Dataset

80% of the the patients in the NIH chest X-ray dataset are assigned to the (raw) training set. We split the dataset by patient to ensure that a patient can only be in one dataset (training, validation or testing set). The training set contains much more negative samples (no pneumonia) than positive samples (has pneumonia). We balance the training set by randomly chosing negative samples until we have the number of positive and negative samples. The rest negative samples in this dataset are discarded.

The final cardinality of the training dataset is 2302.

1.4.2 Description of Validation Dataset

10% of all patients are assigned to the validation dataset. This set is used for picking the best performing model (weights) during training. In a clinical setting we assume a higher incidence of X-ray images with pneumonia than in a dataset which maps a 'complete population'. Therefore we do not sample data points from the validation set until we reach 1.3% of pneumonia cases in the dataset. Instead, we randomly sample data points so that we get a ratio of 20%/80% of positive and negative samples.

The final cardinality of the validation dataset is 745.

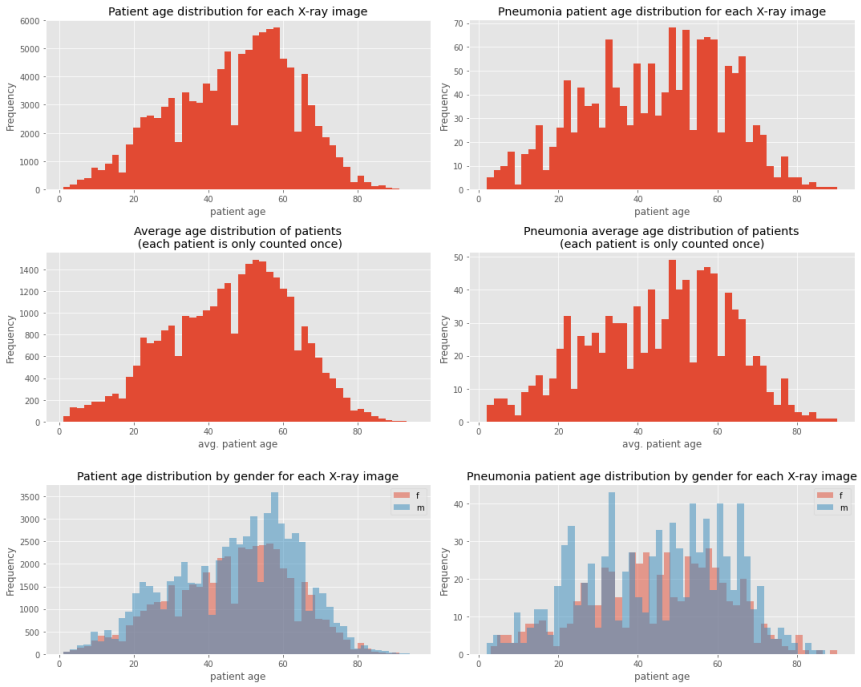

Figure 6: Age demographics

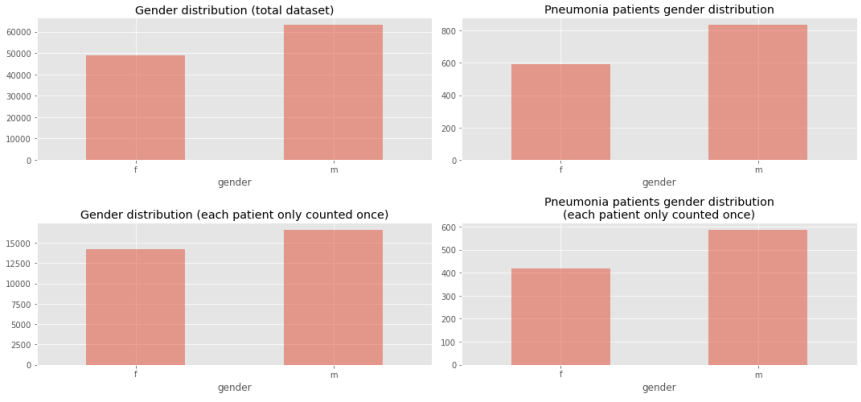

Figure 7: Gender demographics

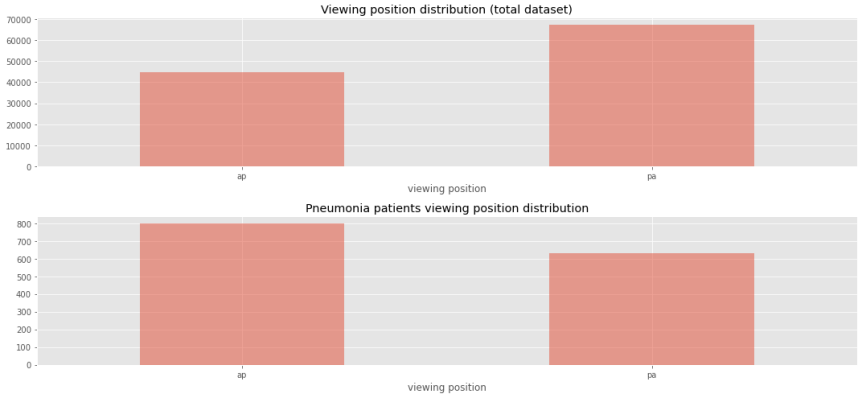

Figure 8: Viewing position of X-ray images

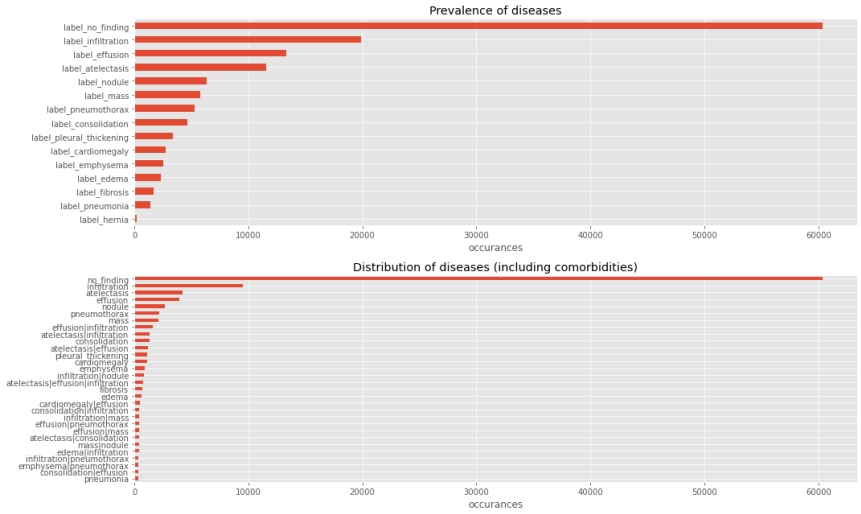

Figure 9: Prevalence of diseases

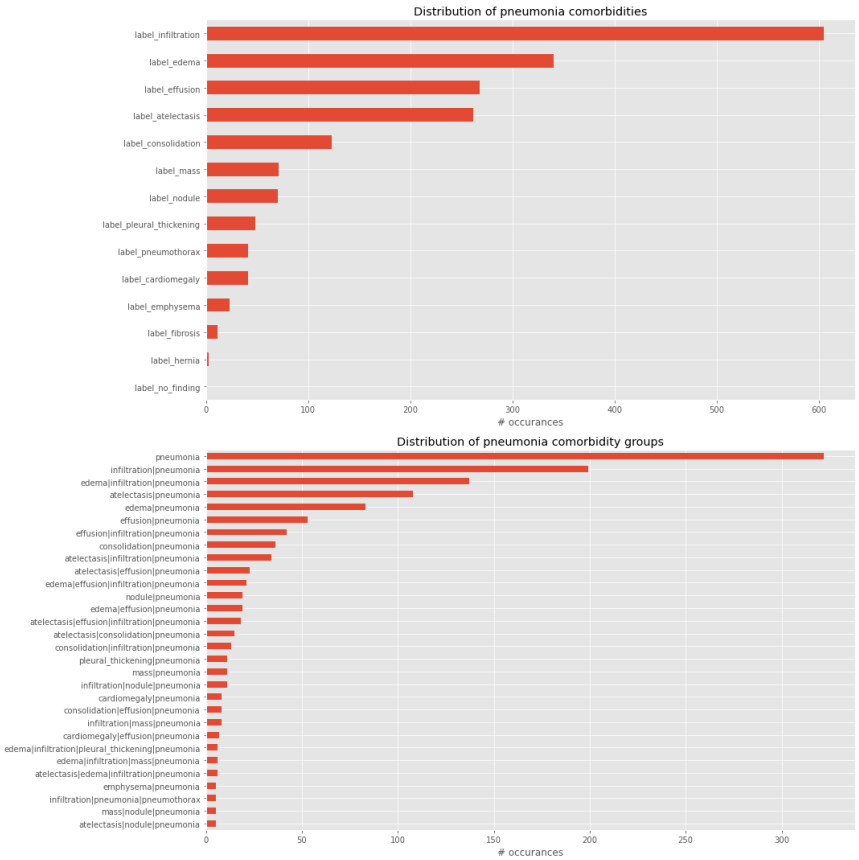

Figure 10: Pneumonia comorbidities

1.4.3 Description of Testing Dataset

10% of the patients are used for testing the performance of the PneumoniaXNet algorithm. This dataset is used to finally evaluate the model performance. It has the same ration between positive and negative samples as the validation dataset: 20%/80%.

The final cardinality of the testing dataset is 640.

1.5 Ground Truth

The ground truth for the used data was created by the NIH. They extracted the labels with the help of a NLP algorithm running over radiology reports which are not publicly available. This process is prone to some erroneous labels because the NLP algorithm might misinterpret complex sentence structures. The NIH reports a NLP labeling accuracy of $>$90%. The NIH states that they had to deal with uncertainties in the radiology reports (see also kaggle data source). Often they classified such uncertain cases as 'no finding'. The 'no finding' label can also contain diseases which are not considered in this dataset. This means that the 'no finding' label might still contain some diseases instead of being a scan of a healthy subject. All these limitations in data labels translate directly to the resulting algorithm which was trained on this data.

On the other hand the benefit of this method is to be able to label huge datasets in a very fast and cost efficient way.

1.6 FDA Validation Plan

1.6.1 Patient Population Description for FDA Validation Dataset

In this section we consider an ideal dataset that might be constructed by a clinical partner for the FDA validation dataset. The demographics of the FDA validation dataset should be similar to the NIH chest X-ray dataset on which the PneumoniaXNet algorithm was trained. This means:

- Age ranges: 10 to 90 years
- sex: male and females
- type of imaging modality: DX (digital radiology)
- body part imaged: chest
- prevalence of disease of interest: 20 % so that it matches the validation set used to evaluate the PneumoniaXNet algorithm

1.6.2 Ground Truth Acquisition Methodology

The most affordable and reliable method of acquiring ground truth labels is to get multiple experts, e.g. radiologists to label the images for presence of pneumonia. A majority vote for each image would reveal the ground truth. This is the silver standard approach.

The gold standard approach would be to take pathological samples of the tissue. This process is very time-consuming and expensive. If this method is available, even for a sub-sample of the available data, it will be valuable for evaluating the performance of the algorithm.

1.6.3 Algorithm Performance Standard

In a previous study done by Rajpurkar et al. the authors trained an algortihm for detecting pneumonia. They measured the performance of their model by comparing the F1 score of the model predictions with the averaged F1 score of four expert radiologists' predictions. To get comparable evaluation results we use the F1 score as performance metric. The F1 score is the harmonic mean of recall and precission. The four expert radiologists achieved an averaged F1 score of 0.387. We use this "radiologist-level value" as standard to beat.

With a F1 score of 0.41 the PneumoniaXNet algorithm performs at least as good as expert radiologists.