FDA Submission

Udo Dehm

October 29, 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 TODO: Algorithm Design and Function

« Insert Algorithm Flowchart or better architecture diagram of model »

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 TODO: Preprocessing Steps

describe the preprocessing steps

- Types of 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
  - TODO: describe augmentation and its parameters used

1.2.3 TODO: CNN Architecture

describe the architecture of the classifier

- we use transfer learning
- freeze most layers
- train only last few layers

1.3 Algorithm Training

1.3.1 TODO: Parameters used for training:

- Batch size
- Optimizer learning rate: TODO:Adam optimizer with initial learning rate of xxx
- Layers of pre-existing architecture that were frozen
- Layers of pre-existing architecture that were fine-tuned
- Layers added to pre-existing architecture

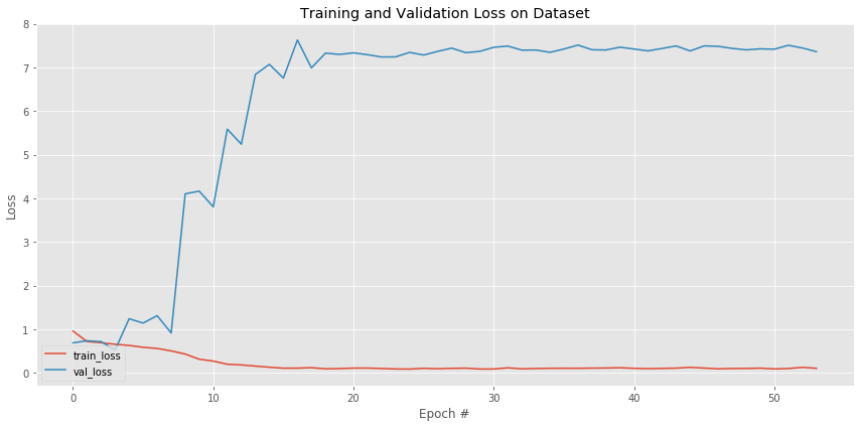

Figure 1: PnuemoniaXNet training performance: Training and validation losses.

1.3.2 TODO: Final Threshold and Explanation

describe the performance statistics and threshold used in final validation

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

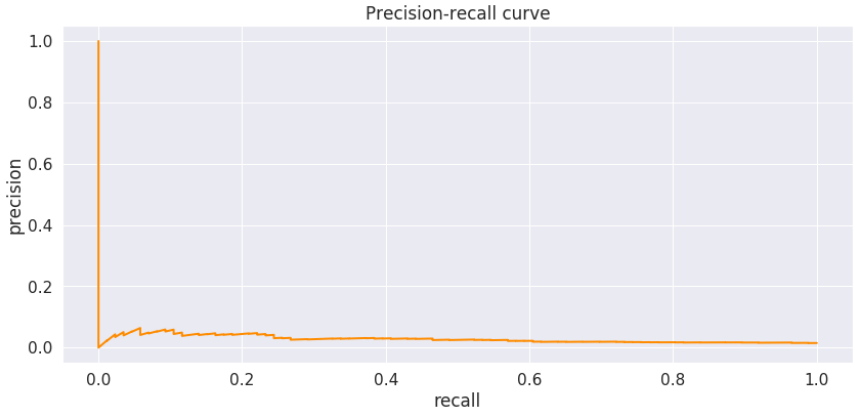

Figure 2: Validation set precision-recall curve

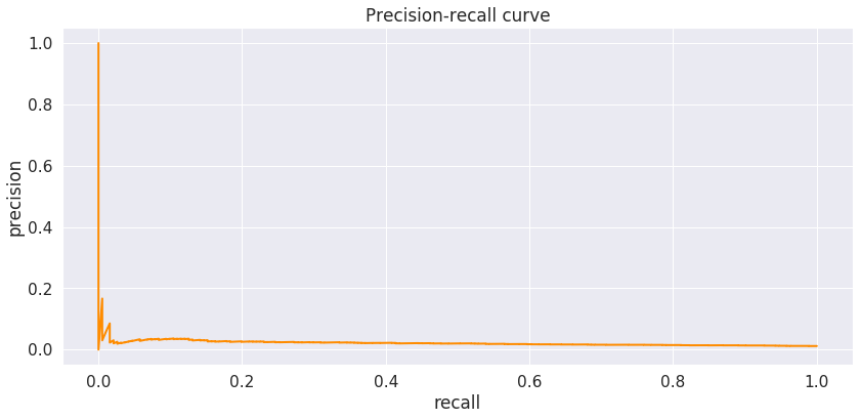

Figure 3: Testing set precision-recall curve

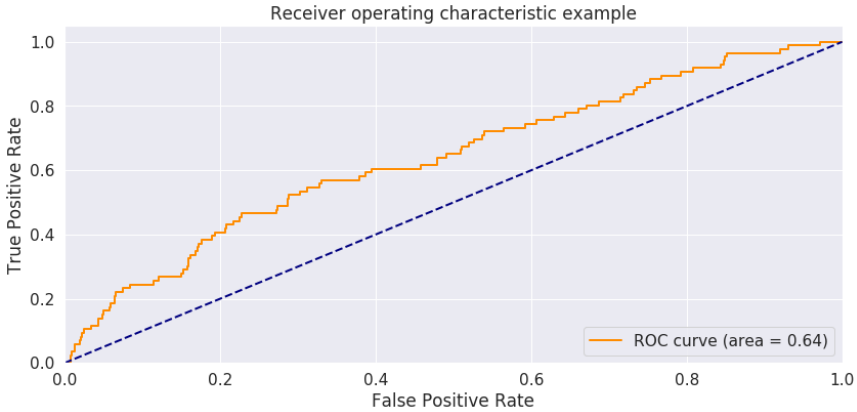

Figure 4: Validation set ROC curve

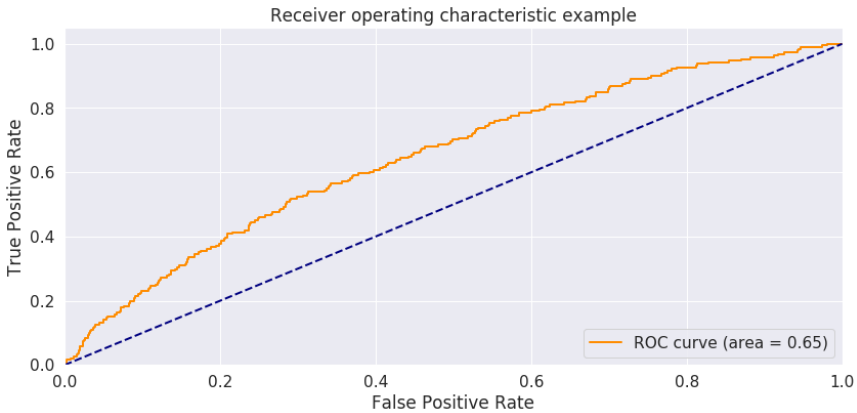

Figure 5: Testing set ROC curve

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.

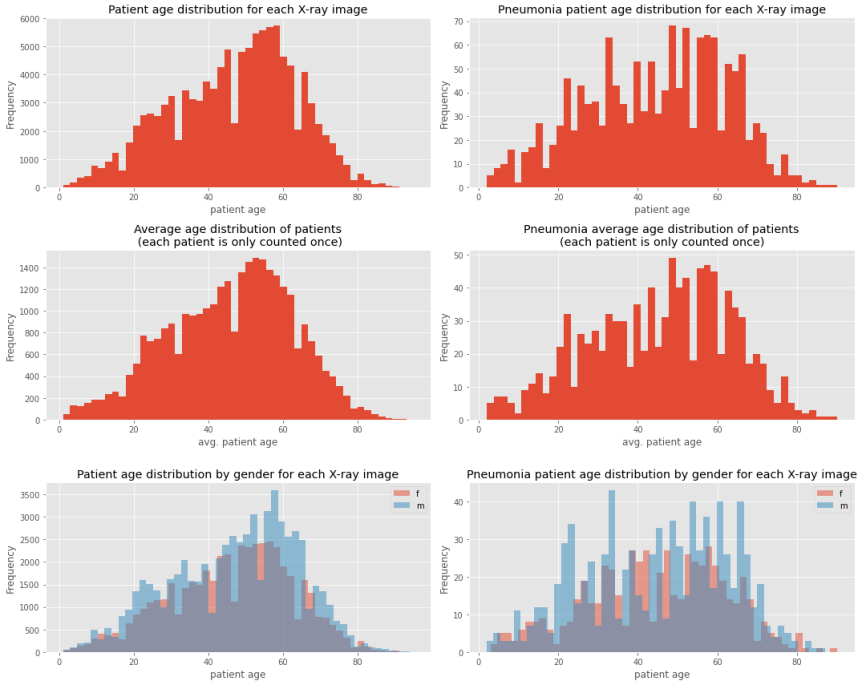

Figure 6: Age demographics

1.4.1 TODO: Description of Training Dataset

80% of the the patients in the NIH chest X-ray dataset are assigned to the training set. We split the dataset by patient to ensure that a patient can only be in one dataset (training, validation or testing set).

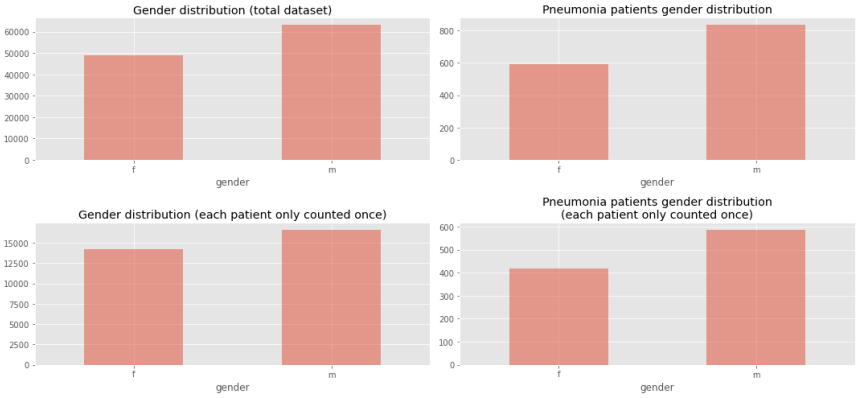

Figure 7: Gender demographics

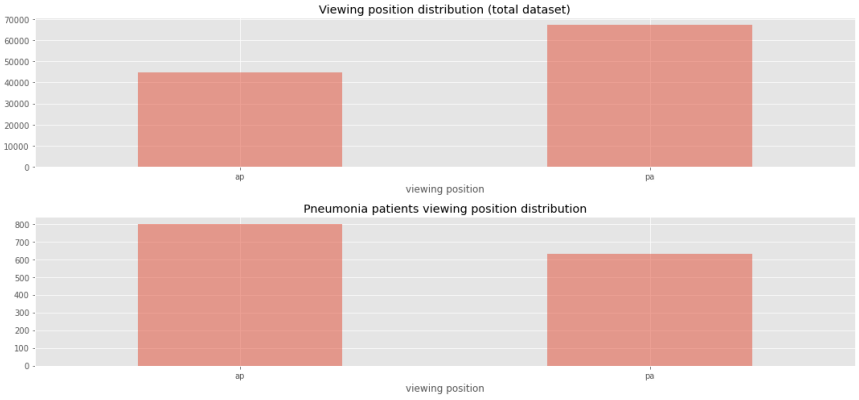

Figure 8: Viewing position of X-ray images

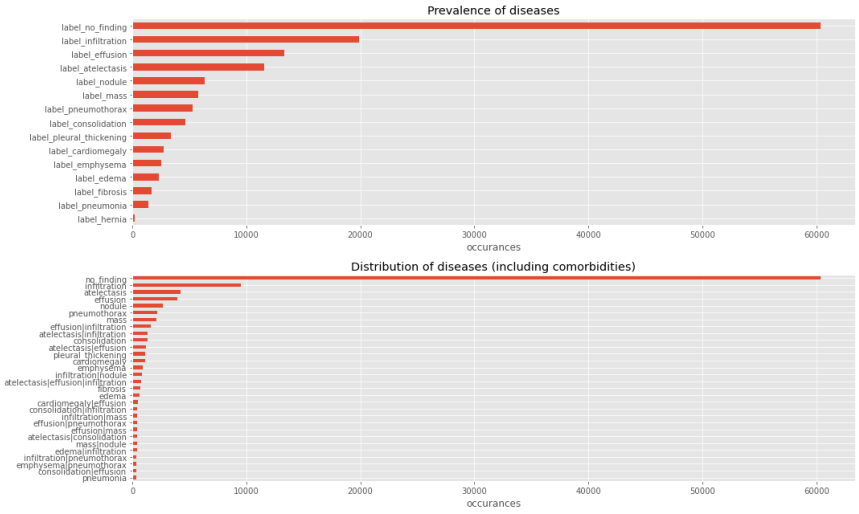

Figure 9: Prevalence of diseases

1.4.2 TODO: Description of Validation Dataset

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.

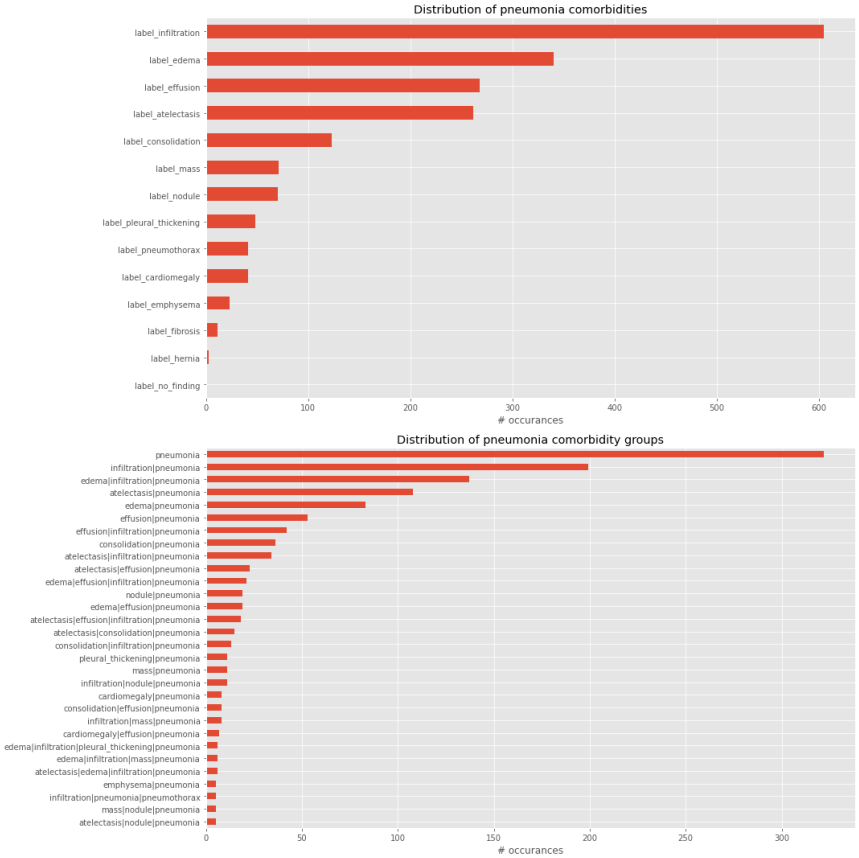

Figure 10: Pneumonia comorbidities

1.6 FDA Validation Plan

1.6.1 TODO: 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
- TODO: prevalence of disease of interest: ... % 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 TODO: 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.

TODO: write something about performance of the PneumoniaXNet algorithm and how it outperforms the radiologist-level performance.