FDA Submission

Name: Marta Kaczmarska

Name of the Device: Pneumonia Detector

Algorithm Description

1. General Information

Intended Use Statement:

Assist a radiologist in detecting pneumonia on chest X-ray images.

Indications for Use:

Serve as additional confirmation of presence/absence of pneumonia. While used in emergency setting the device could aid in prioritizing the images reading queue. The device is indicated for use with both males and females (aged 20-70) radiographs obtained in AP or PA position.

Device Limitations:

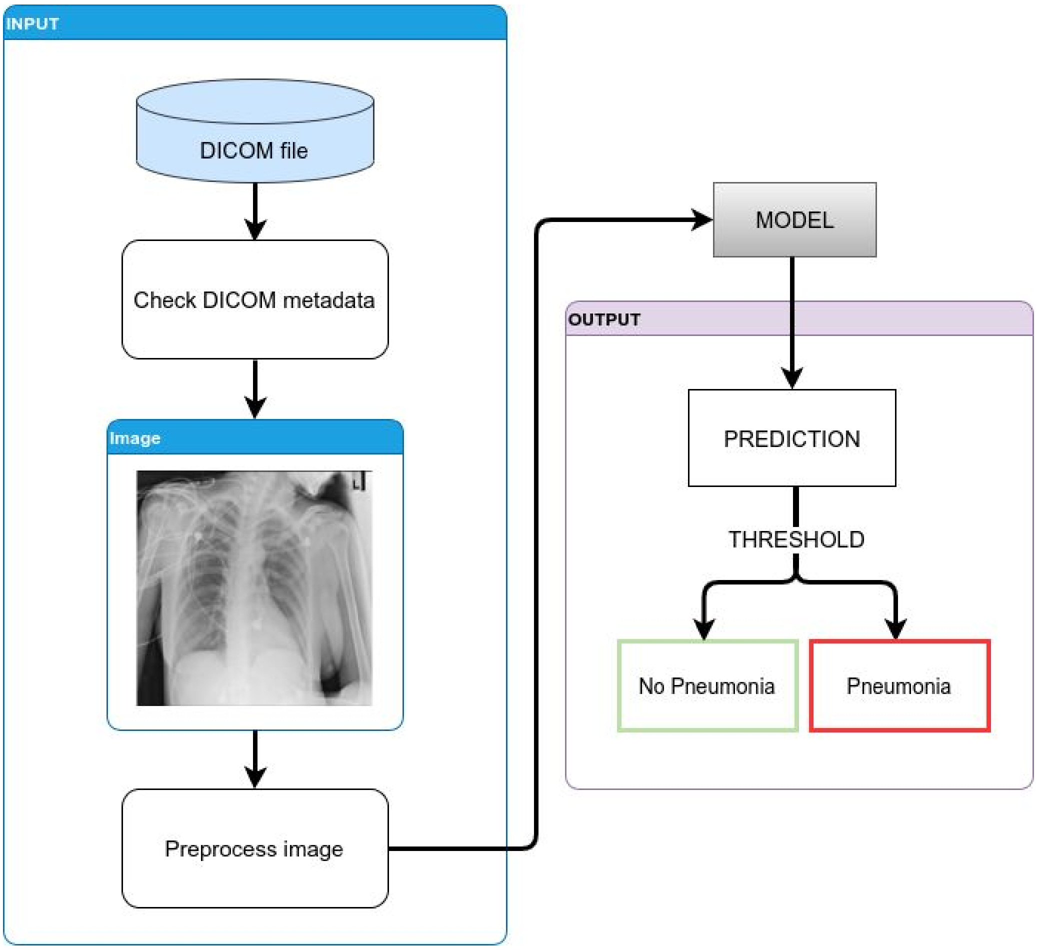
Device is not suitable for images other than chest radiographs obtained in AP or PA position. Caution should be taken while used for patients younger than 20 and older than 70.
While used in emergency setting, additional computational resources may be taken into consideration. A single outcome of the device is produced fast from the human perspective; exemplary results for one image are:
CPU time: total: 3 µs
wall time: 8.34 µs
However constant flow of images may cause a bottleneck, which in consequence may extend waiting time for the results significantly.
The device has high False Positives rate, which may impact hospital resources (too many patients require bed/medicaments), however it is safer for the patient to assume disease at first and further examine the case keeping the patient for observation.
Pneumonia might be confused with other thoracic diseases. These include especially Infiltration, Consolidation, Edema, Effusion, Fibrosis, which may present similar image features as Pneumonia.

Clinical Impact of Performance:

Device would assist radiologist, which in turn would faciliate pneumonia detection and shorten reading time as well as lower mistakes, especially pneumonia marked as healthy patient (False Negatives). Patient with suspected pneumonia could be further examined or treated sooner. Probability displayed along with the Pneumonia/No Pneumonia label would further help clinician with assessing the diagnosis. That would help for example with confirmation if patient really have pneumonia (avoiding False Positives), which in turn could help save the hospital resources and save the patient from unneccesary hospitalization.

2. Algorithm Design and Function

Algorithm Flowchart:



DICOM Checking Steps:

1. Modality check: DX
2. Body Part Examined check: Chest
3. Patient Position check: AP, PA

Preprocessing Steps:

After getting pixel data from DICOM file:

1. Resize image to 224 x 224 px
2. Stack grayscale image along 3rd axis to obtain 3-channels image
3. Add dimension to obtain 4-dimensional tensor
4. Preprocess image according to the VGG16 model architecture requirements:
   - convert from RGB to BGR
   - zero-center each color channel, without scaling

CNN Architecture:

Layer (type)	Output Shape	Param #
input_1 (InputLayer)	(None, 224, 224, 3)	0
block1_conv1 (Conv2D)	(None, 224, 224, 64)	1792
block1_conv2 (Conv2D)	(None, 224, 224, 64)	36928
block1_pool (MaxPooling2D)	(None, 112, 112, 64)	0
block2_conv1 (Conv2D)	(None, 112, 112, 128)	73856
block2_conv2 (Conv2D)	(None, 112, 112, 128)	147584
block2_pool (MaxPooling2D)	(None, 56, 56, 128)	0
block3_conv1 (Conv2D)	(None, 56, 56, 256)	295168
block3_conv2 (Conv2D)	(None, 56, 56, 256)	590080
block3_conv3 (Conv2D)	(None, 56, 56, 256)	590080
block3_pool (MaxPooling2D)	(None, 28, 28, 256)	0
block4_conv1 (Conv2D)	(None, 28, 28, 512)	1180160
block4_conv2 (Conv2D)	(None, 28, 28, 512)	2359808
block4_conv3 (Conv2D)	(None, 28, 28, 512)	2359808
block4_pool (MaxPooling2D)	(None, 14, 14, 512)	0
block5_conv1 (Conv2D)	(None, 14, 14, 512)	2359808
block5_conv2 (Conv2D)	(None, 14, 14, 512)	2359808
block5_conv3 (Conv2D)	(None, 14, 14, 512)	2359808
block5_pool (MaxPooling2D)	(None, 7, 7, 512)	0
flatten_1 (Flatten)	(None, 25088)	0
dropout_1 (Dropout)	(None, 25088)	0
dense_1 (Dense)	(None, 2048)	51382272
dropout_2 (Dropout)	(None, 2048)	0
dense_2 (Dense)	(None, 1)	2049

3. Algorithm Training

Parameters:

- Types of augmentation used during training:
  - horizontal flip
  - height shift – range: 0.1 of total height
  - width shift – range: 0.1 of total width
  - rotatio – range: 10 degrees
  - zoom – range: 0.1

- Batch size
  - 32

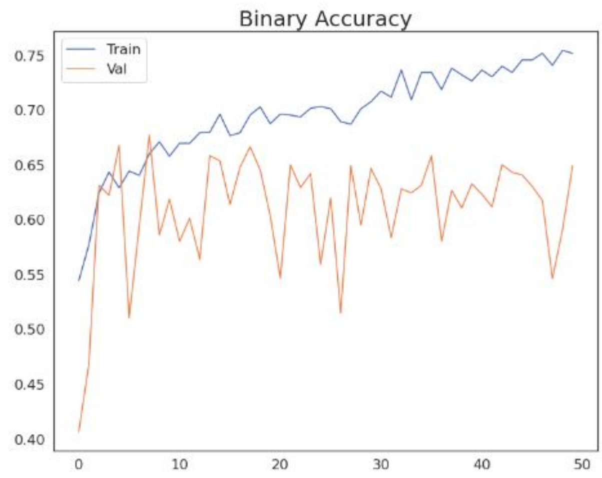
- Optimizer learning rate
  - 0.001

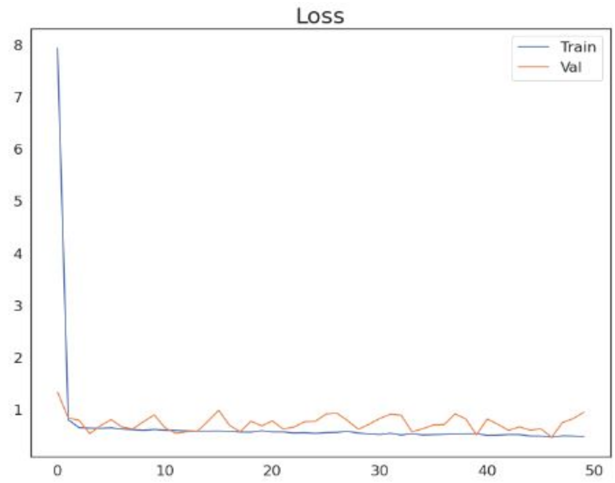
- Layers of pre-existing architecture that were frozen
  - input_1 (InputLayer)
  - block1_conv1 (Conv2D)
  - block1_conv2 (Conv2D)
  - block1_pool (MaxPooling2D)
  - block2_conv1 (Conv2D)
  - block2_conv2 (Conv2D)
  - block2_pool (MaxPooling2D)
  - block3_conv1 (Conv2D)
  - block3_conv2 (Conv2D)
  - block3_conv3 (Conv2D)
  - block3_pool (MaxPooling2D)
  - block4_conv1 (Conv2D)
  - block4_conv2 (Conv2D)
  - block4_conv3 (Conv2D)
  - block4_pool (MaxPooling2D)
  - block5_conv1 (Conv2D)
  - block5_conv2 (Conv2D)

- Layers of pre-existing architecture that were fine-tuned
  - block5_conv3 (Conv2D)
  - block5_pool (MaxPooling2D)

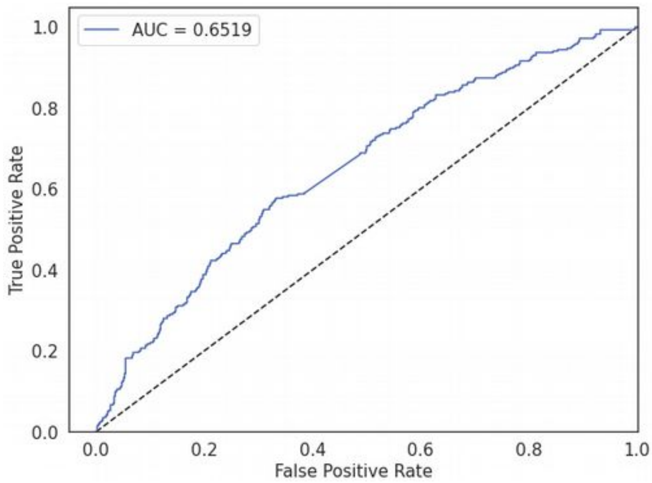
- Layers added to pre-existing architecture
  - flatten_1 (Flatten)
  - dropout_1 (Dropout 0.3)
  - dense_1 (Dense)
  - dropout_2 (Dropout 0.2)
  - classification layer: dense_2 (Dense)

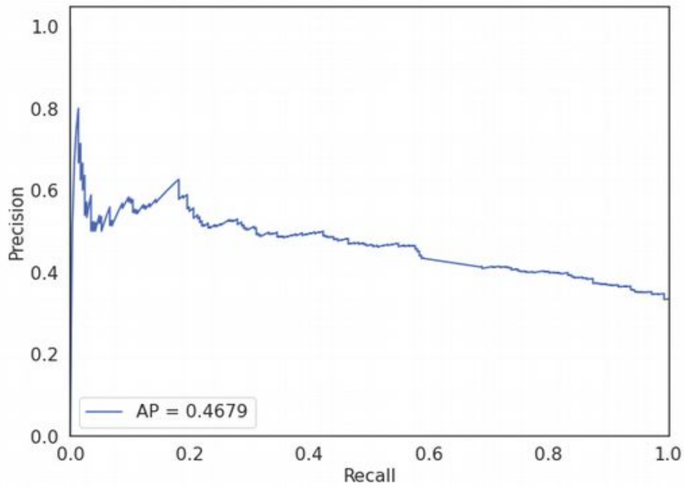
Algorithm Training Performance:





ROC and Precision-Recall Curves:

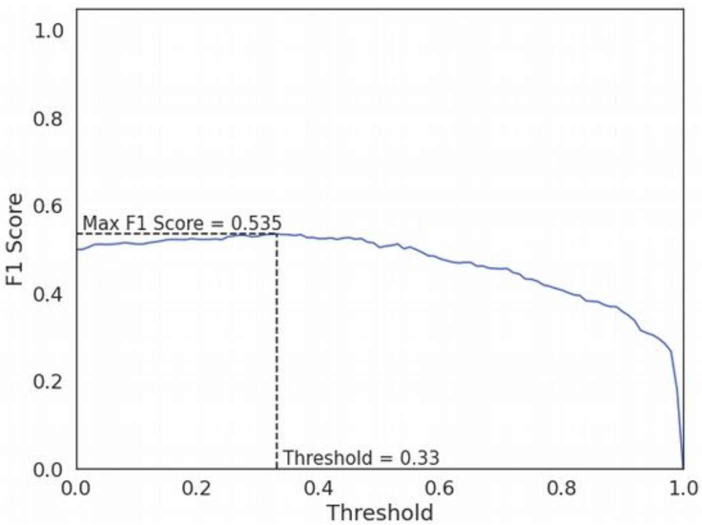




Final Threshold and Explanation:

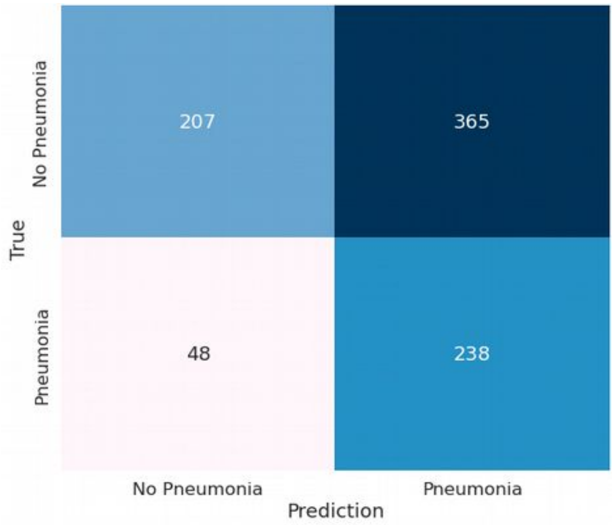
Threshold: 0.33

Chosen threshold maximizes f1 score allowing to have high Pneumonia recall at the same time. This results in low False Negative rate.



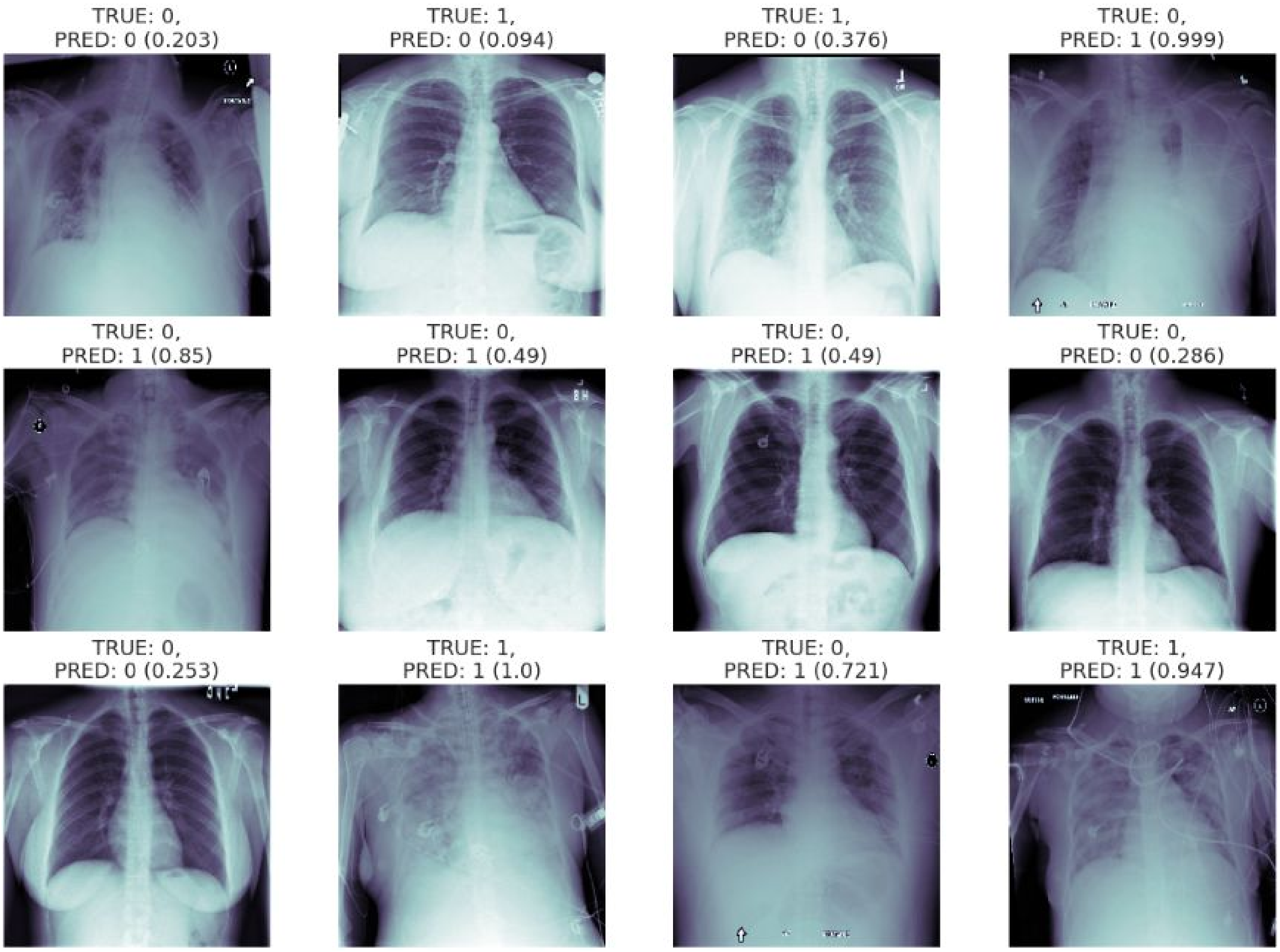
Classification Report and Confusion Matrix for the Chosen Threshold:

	Precision	Recall	F1-Score	Support
No Pneumonia	0.81	0.36	0.50	572
Pneumonia	0.39	0.83	0.54	286



Examples of the Algorithm Results Obtained after Incorporating the Chosen Threshold:

Number to the left denotes true label, number to the right denotes model prediction, where: 0 – No Pneumonia, 1 – Pneumonia

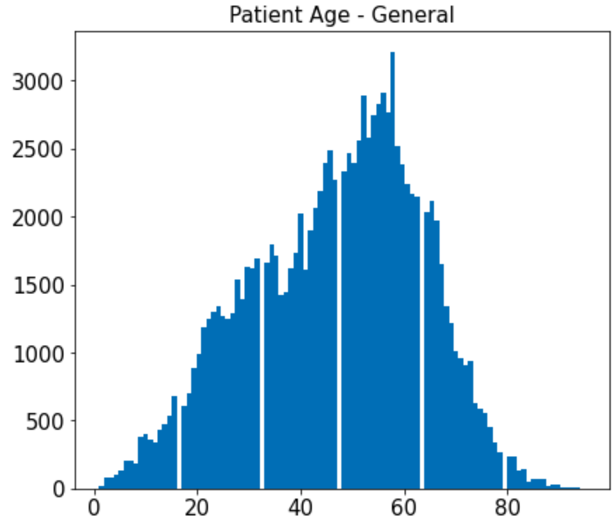


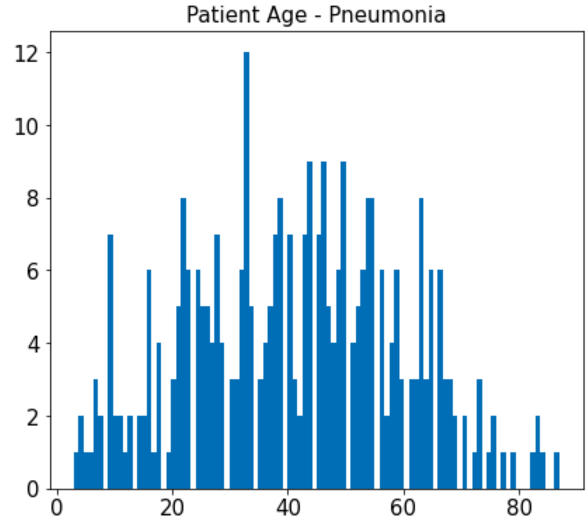
4. Databases

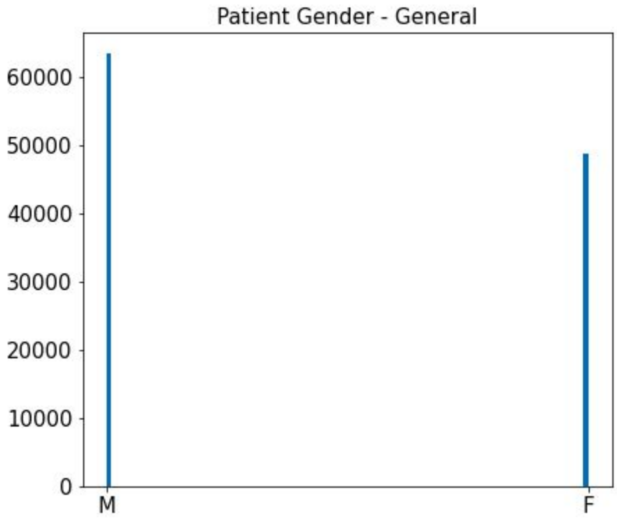
Description of the Whole Dataset:

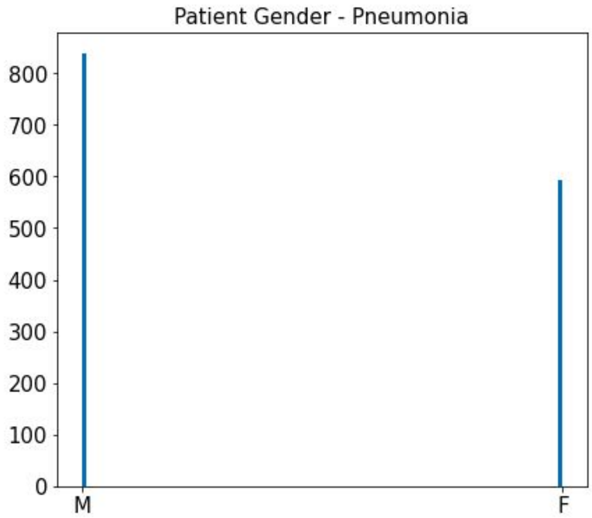
The NIH dataset contain 112120 X-ray images. There are 1431 cases with the disease of interest in this dataset.

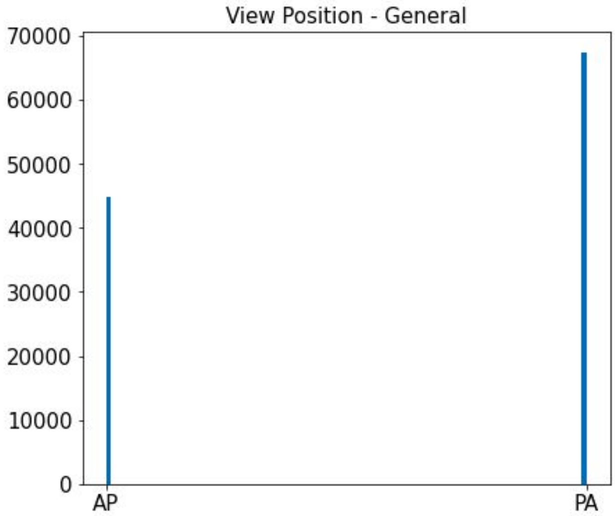
Some chosen characteristic of the whole dataset in comparison to the Pneumonia subset are shown on the following charts:

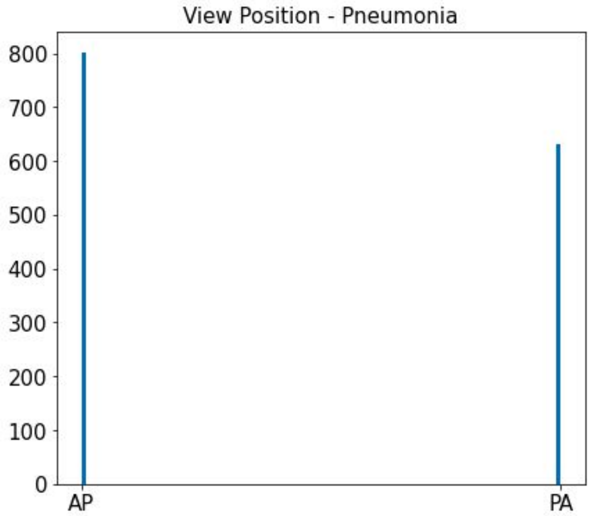


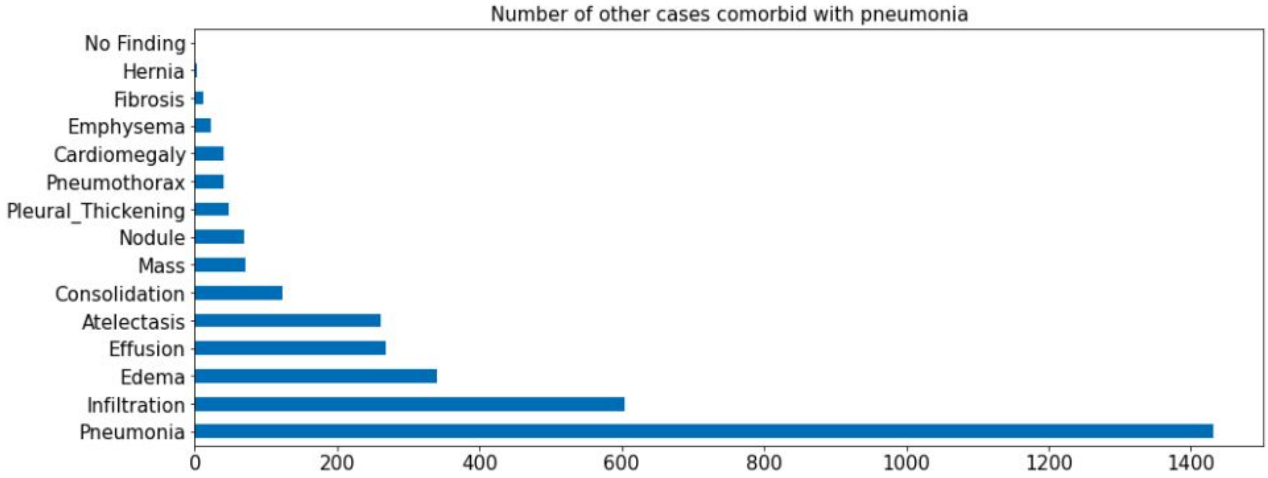




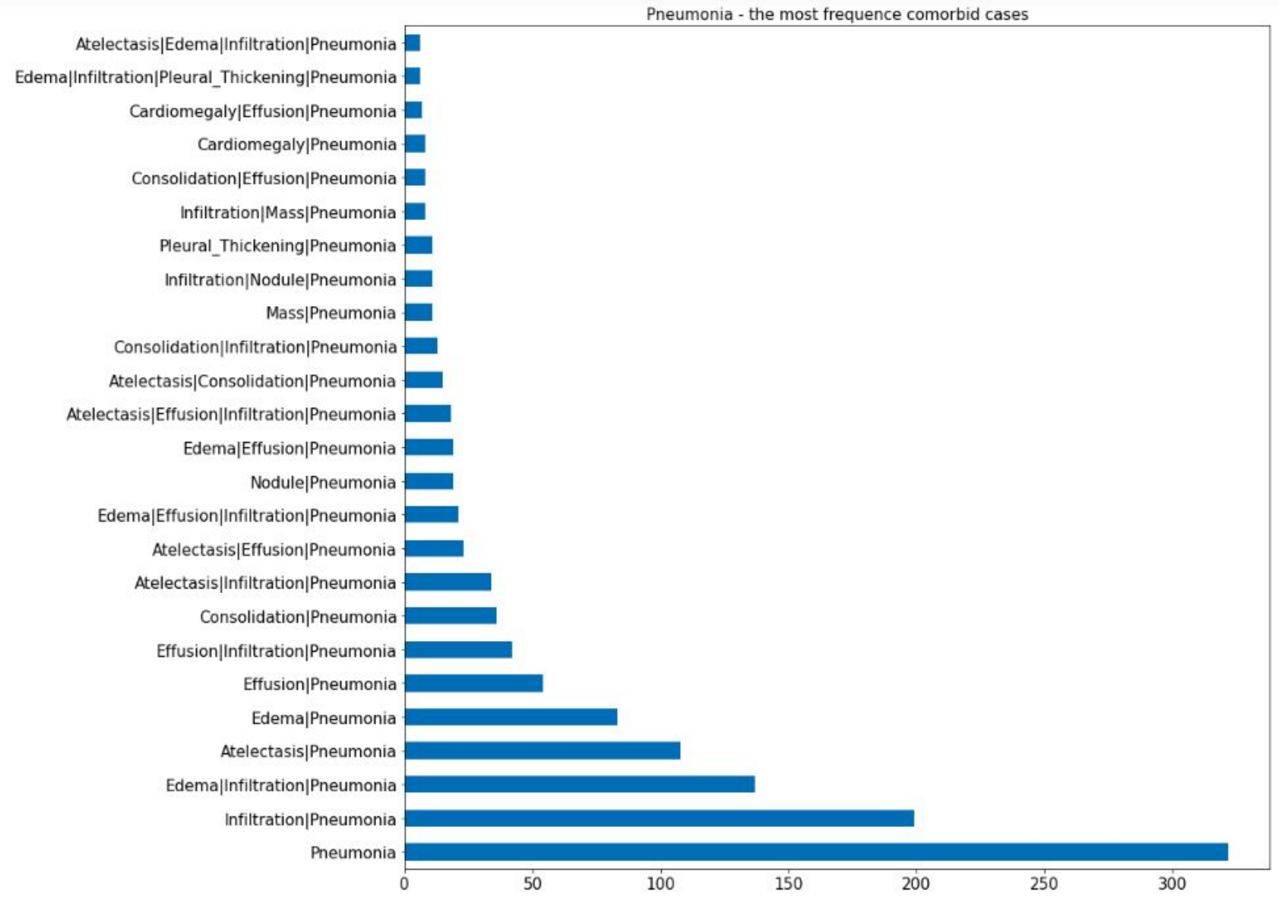








The most common comorbidity for Pneumonia are Infiltration and Edema. The least common is Hernia. On the plot below 25 most frequent labels are presented.



From the whole dataset the subsets of training and validation data were yielded.

To make the most out of the available pneumonia images, the sixteen outliers (patient age > 100) were not rejected. It was assumed that such mistakes do not affect the patient demographics significantly. What is more, in general deep learning models perform better while fed more data and such mistakes in DICOM metadata do not affect the image quality.

Description of Training Dataset:

Training dataset consist of 2290 cases; 1145 are pneumonia.

The best approach is to have equal amounts of positive and negative class in the training dataset. The instances were selected randomly as 80% of the whole NIH dataset and then some negative instances were rejected randomly to generate balanced dataset.

Positive class is considered to be Pneumonia or Pneumonia with one or more comorbidities. Negative class is absence of Pneumonia, that is no disease at all or presence of any other thoracic disease.

Description of Validation Dataset:

Validation dataset consist of 858 cases; 286 are pneumonia.

It was decided to have pneumonia in 1/3 cases in validation dataset. From 20% of the whole NIH dataset that was created after splitting into training (described above) and validation sets some negative instances were rejected to generate desired distribution of the classes in the resultant validation datset.

5. Ground Truth

First, the radiologist reports were produced by the clinicians. Then, the labels were extracted from these reports using NLP methods and this extracted information is considered the ground truth for training and testing the model.

Apart from mistakes in the reports such method of obtaining the true labels involve other limitations, such as less than 100% corectness of NLP method used.

6. FDA Validation Plan

Patient Population Description for FDA Validation Dataset:

Patient characteristics:

- patient age: 20-70
- patient gender: male or female

Image characteristics:

- imaging modality: X-ray
- imaging position: AP or PA
- body part examined: chest

Disease of interest and comorbidities:

- presence of pneumonia - ~40% of the validation dataset
- no disease - ~40% of the validation dataset
- presence of Infiltration and/or Edema in combination with Pneumonia - ~10% of the validation dataset
- other diseases present separately or in combination with each other and/or Pneumonia: Atelectasis, Cardiomegaly, Consolidation, Effusion, Emphysema, Fibrosis, Hernia, Mass, Nodule, Pleural Thickening, Pneumothorax - ~10% of the validation dataset

Ground Truth Acquisition Methodology:

Since pneumonia can be hard to spot, ground truth should be obtained as a majority vote from 3 radiologists of different years of experience. Each radiologist should examine the image independently.

In case of unavailability of radiologists true labels could be extracted from radiological reports using NLP methods similarly as true labels obtained for NIH dataset.

Algorithm Performance Standard:

Algorithm should be at least as good as the current standards. Comparison of the performance of radiologists, other deep learning solutions and the subject of this submission (as achieved on validation dataset) is presented in the tables below:

	F1 Score
Radiologist Average[1]	0.387
CheXNet[1]	0.435
Pneumonia Detector	**0.535**

	AUC
Wang et al.[2]	0.633
Yao et al.[3]	0.713
CheXNet	**0.768**
Pneumonia Detector	0.652

1 *Rajpujkar P., Irvin J. et al., CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning*, 2017, https://stanfordmlgroup.github.io/projects/chexnet/

2 Wang X. et al, *ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases*, 2017, https://arxiv.org/abs/1705.02315

3 *Yao L. et al., Learning to diagnose from scratch by exploiting dependencies among labels*, 2017, https://arxiv.org/abs/1710.10501