# FDA Submission

**Laima Marcinkeviciute**

## Application for Pneumonia detection from Chest X-rays

**Algorithm Description**

1. General Information

**Intended Use Statement:**
The algorithm is intended to assist radiologists with chest X-ray screening for detecting Pneumonia.

**Indications for Use:**
- Patient Age: 10 to 75
- Patient Gender: M and F
- Chest X-Ray with view positions: AP and PA

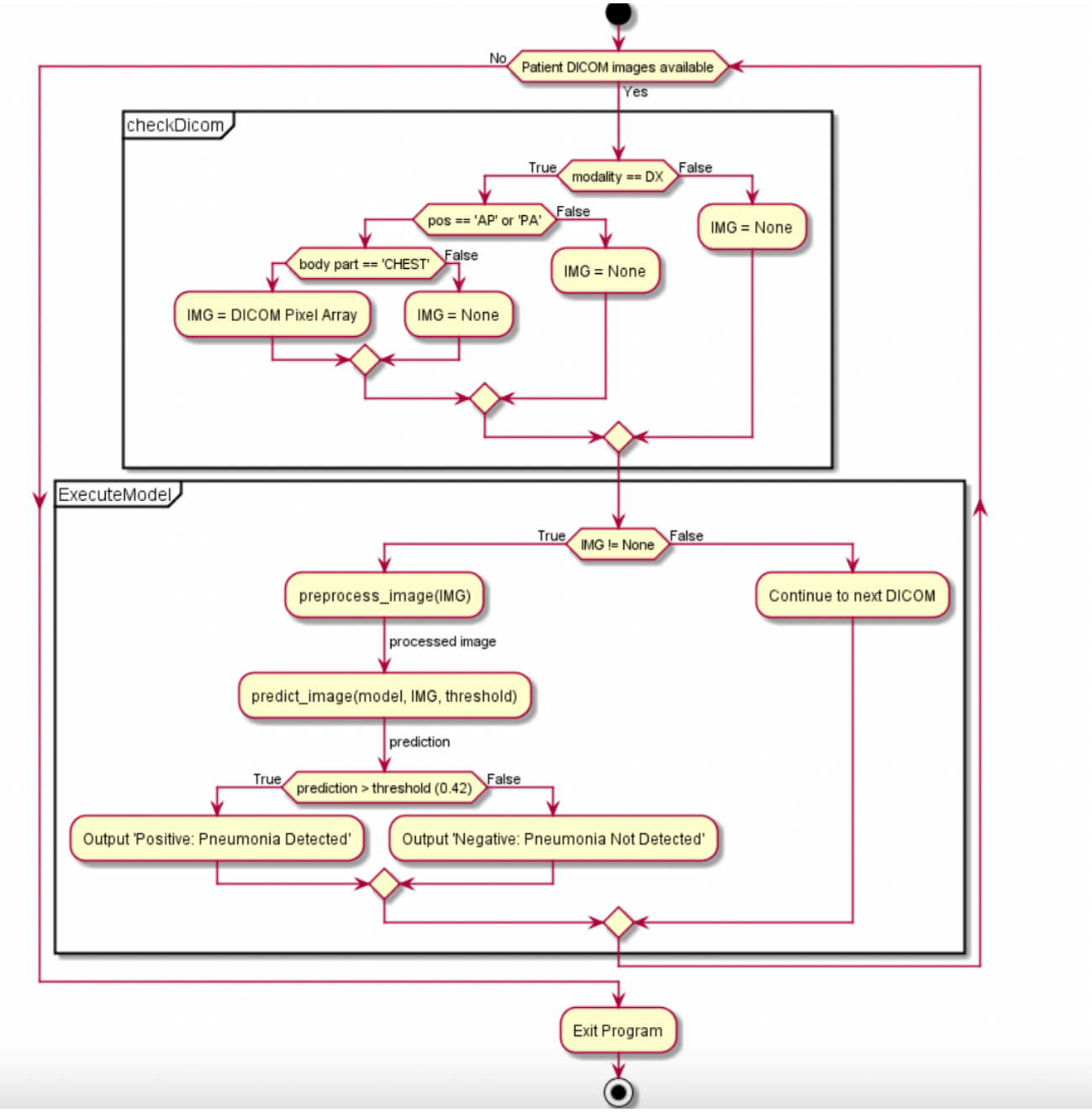
Clinical Setting: The algorithm is intended for integration into the workflow of diagnostic clinics.
This is not an emergency detection scenario, so chest x-ray images may be sent to a remote server for processing.
DICOM format following HIPAA rules must be used for all X-Ray images.
Each X-Ray DICOM metadata is first verified for correct patient information (see DICOM Checking Steps). If the X-Ray passes the DICOM verification step, the image is then pre-processed and input to the machine learning algorithm, yielding a prediction. After the prediction is complete, the result is sent to a radiologist. The radiologist will give the final diagnosis based on their own independent analysis.

**Device Limitations:**
It is recommended to run this application on a GPU-enabled workspace either locally or, more likely, in the cloud to speed up the model execution.
The device (algorithm) does not achieve 100% accuracy. Therefore, algorithm results must be only be used as supplemental data to an expert radiologist who will determine the final diagnosis. Algorithm results may not be trusted individually for final diagnosis.

**Clinical Impact of Performance:**
Reduces time of verification of pneumonia diagnoses. Runtime is well under a second per diagnosis. Thus, the CNN is much faster than human doctors.
The early detection algorithm attempts to minimize false negative rates in Pneumonia classification, and has a higher likelihood of a false positive reading. All readings, false positives and false negatives, must be reviewed by a radiologist.
- **False Positives** incorrectly indicate the presence of Pneumonia. Further review by a radiologist my waste time and resources for the patient and hospital, but a final diagnosis is confirmed by the radiologist and there is no life-threatening Pneumonia that has gone undetected in the patient.

- **False Negatives** incorrectly indicate no presence of Pneumonia. This is a life-threatening scenario for the patient and it is vital that a radiologist review the result and make a final diagnosis incorporating their own expert analysis. False negative occurrences are severe failures of the algorithm, so the algorithm has been optimized to attempt to minimize false negative readings.

2. Algorithm Design and Function



**DICOM Checking Steps:**
Verify DICOM metadata:
Body Part must be CHEST
Modality must be DX
Patient Position must be AP or PA

**Pre-processing Steps:** Image Pre-processing:
1. Resize image to 224x224 pixels
2. Convert from grayscale to RGB (Image dimensions: 224x224x3 pixels)
3. Apply ResNet image pre-processing which re-centers pixel values around zero (subtract pixel mean) but does not normalize range (do not divide by pixel standard deviation)

**CNN Architecture:** The ResNet50 model architecture is used as a basis for the Pneumonia detection model, with a flatten layer and two additional fully connected layers appended. Dropout and batch normalization layers are also included to improve training and to achieve a better model performance accuracy.

3. Algorithm Training

**Parameters:**

- Types of augmentation used during training:
    - horizontal flip
    - height/width shift with range 0.1
    - rotation range 20
    - shear range 0.1
    - zoom range 0.1
- Batch size
    - 32
- Optimizer learning rate
    - Adam optimizer, learning rate 1e-4
- Layers of pre-existing architecture that were frozen
    - All layers except final 2D convolutional layer and average pooling were frozen
- Layers of pre-existing architecture that were fine-tuned
    - Final 2D convolutional layer 'conv5_block3_3_conv' and following layers were fine tuned:

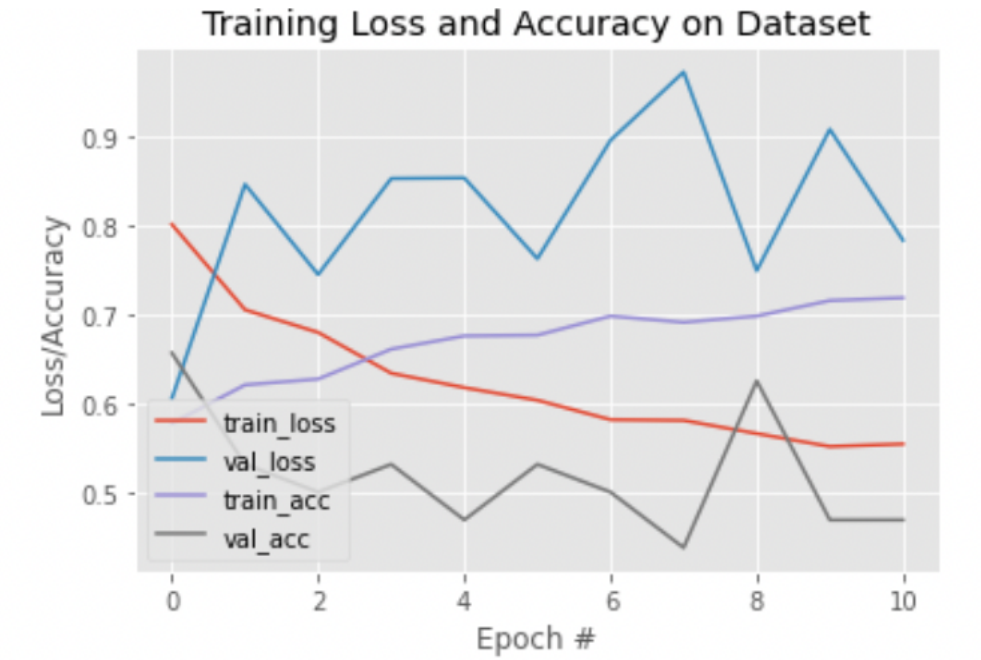
```
Layer (type)                 Output Shape              Param #
=================================================================
conv5_block3_3_conv (Conv2D) (None, 7, 7, 2048)        1050624
_________________________________________________________________
conv5_block3_3_bn (BatchNormali (None, 7, 7, 2048)     8192
_________________________________________________________________
conv5_block3_add (Add)       (None, 7, 7, 2048)        0


_________________________________________________________________
conv5_block3_out (Activation)  (None, 7, 7, 2048)      0
_________________________________________________________________
avg_pool (GlobalAveragePooling2 (None, 2048)           0
=================================================================
```

- Layers added to pre-existing architecture
    - Flatten, dropout, dense, batch normalize, dense

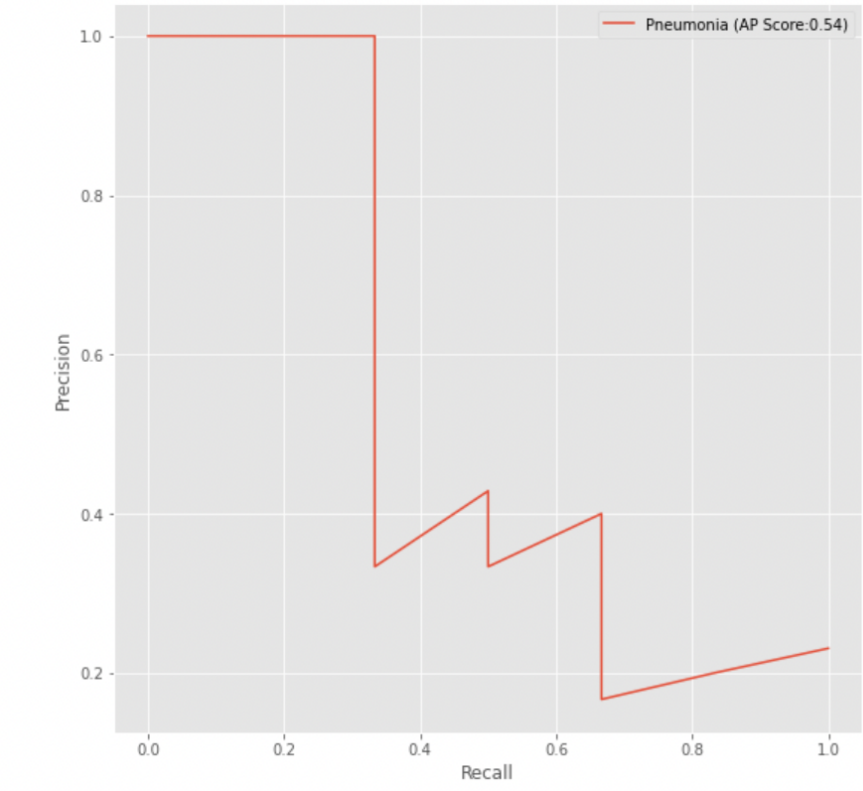
Model Architecture: "Pneumonia_Detection_Model"

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
model_1 (Model)              (None, 7, 7, 2048)        23587712
_________________________________________________________________
flatten_1 (Flatten)          (None, 100352)            0
_________________________________________________________________
dropout_1 (Dropout)          (None, 100352)            0
_________________________________________________________________
dropout_2 (Dropout)          (None, 100352)            0
_________________________________________________________________
dense_1 (Dense)              (None, 64)                6422528
_________________________________________________________________
batch_normalization_1 (Batch (None, 64)                256
_________________________________________________________________
activation_1 (Activation)    (None, 64)                0
_________________________________________________________________
dense_2 (Dense)              (None, 1)                 65
=================================================================
Total params: 30,010,561
Trainable params: 7,477,441
Non-trainable params: 22,533,120
```

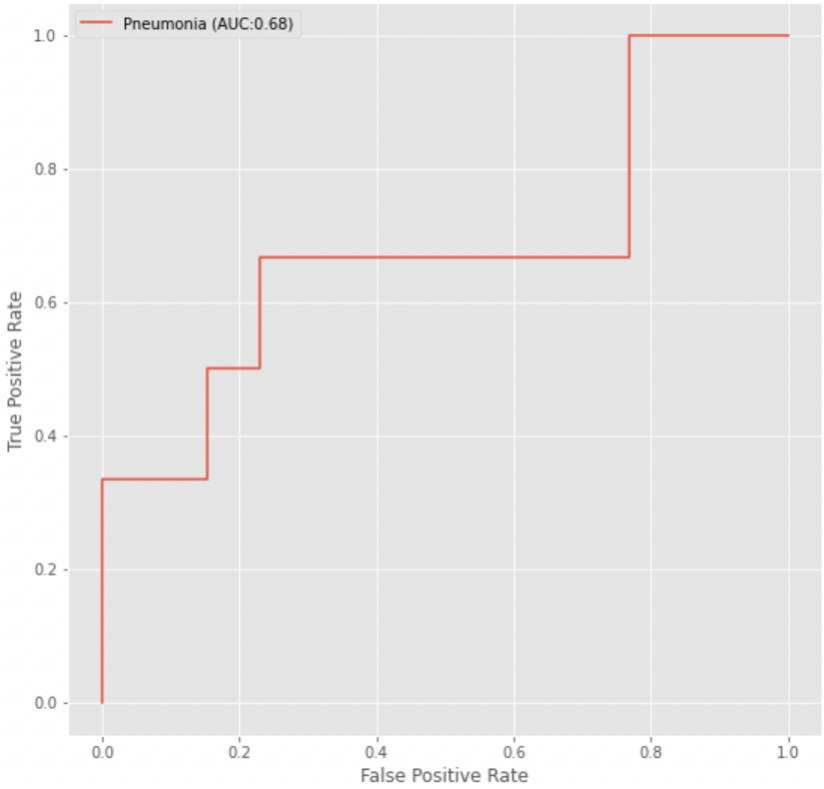
**Training performance:**




**Precision-recall curve:**



**ROC curve:**




**Final Threshold and Explanation**: It is preferred that the model performance favours recall, so that the model has a fewer number of false negatives. All readings, false positives and false negatives, must be reviewed by a radiologist.

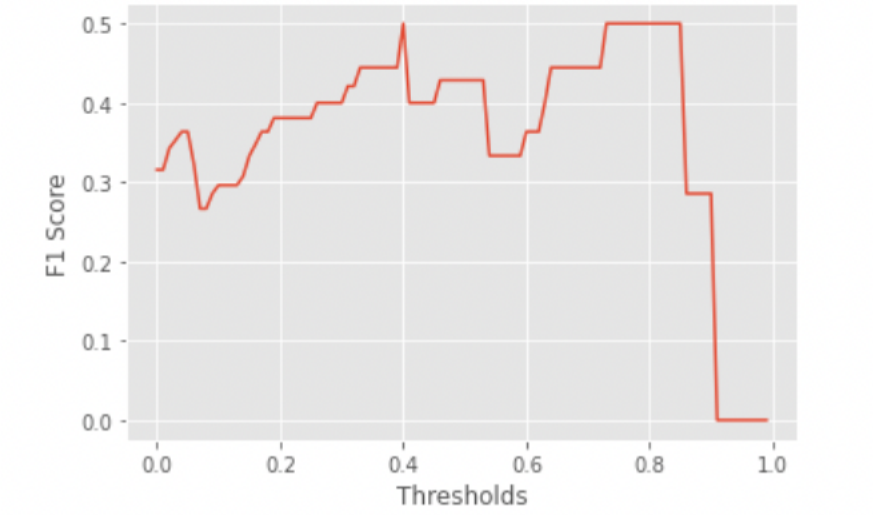
**False Positives** incorrectly indicate the presence of Pneumonia. Further review by a radiologist my waste time and resources for the patient and hospital, but a final diagnosis is confirmed by the radiologist and there is no life-threatening Pneumonia that has gone undetected in the patient.

**False Negatives** incorrectly indicate no presence of Pneumonia. This is a life-threatening scenario for the patient and it is vital that a radiologist review the result and make a final diagnosis incorporating their own expert analysis. False negative occurrences are severe failures of the algorithm, so the algorithm has been trained to attempt to minimize false negative readings. The classification threshold is determined as follows: For each threshold in the range of thresholds from 0.0 to 1.0 at increments of 0.01: Make predictions using threshold and the validation set, and calculate and save the F1 score Save the best (maximum) F1 score and associated threshold.
With this method, we choose a final classification threshold of **0.4**.

At a threshold of 0.4,
F1 score is maximized with a value of 0.5,
Recall is favoured with a value of 0.66,
Precision maintains a value of 0.4.

F1 Score vs Threshold is shown below:

```
Best threshold: 0.4
Precision: 0.4
Recall: 0.6666666666666666
Max F1 Score: 0.5
```




4. Databases

In the NIH master dataset, there are 1431 images with Pneumonia present.
The training and validation datasets are developed by splitting the NIH dataset and stratifying by the presence of Pneumonia.
The rate of Pneumonia in the total population (NIH dataset) is around 1%.
All sample cases with Pneumonia are included within Training and Validation sets.

**Description of Training Dataset:**

The training dataset contains 80% of the available samples with Pneumonia, and holds an equal number (50/50 split) of image samples with and without Pneumonia sampled from the NIH dataset. The training set size is: (1145 samples with Pneumonia + 1145 samples without Pneumonia) = 2290 total samples.

**Description of Validation Dataset:**

The validation dataset contains the remaining 20% of the available samples with Pneumonia, and holds a 20/80 split of image samples with and without Pneumonia sampled from the NIH dataset, e.g. the probability of randomly sampling a non Pneumonia case from the validation dataset it 80%. This 20/80 split is designed to mimic a clinical setting where it would be expected that roughly 20% of the scans may present with Pneumonia, even though the total population rate of Pneumonia is far lower (around 1% in NIH dataset). No samples from the validation dataset are included in the training dataset.
The validation set size is: (286 samples with Pneumonia + 1144 samples without Pneumonia), 1430 samples in total.

## 5. Ground Truth

The data is obtained from the NIH Chest X-ray Dataset.
The consensus labels of three U.S. board-certified radiologists (the majority of votes of three radiologists) were used as the reference standard of "ground truth".
This dataset was not specifically acquired for pneumonia analysis, so the percentage of samples in this dataset is very small:
Samples with Pneumonia: 1431
Samples without Pneumonia: 110689
Fraction: (1431 / 110689) = 0.013
The ground truth labels are developed using Natural Language Processing software, which may incorrectly categorize some image labels.
Additionally, the NIH dataset samples show many comorbidities for Pneumonia, primarily Infiltration and Edema.
Since the model is designed to predict any occurrence of pneumonia including cases where Pneumonia is present alongside other diseases, this ground truth dataset is acceptable.
However, the model would likely be improved by training on a dataset with many more cases of Pneumonia present and without the presence of comorbidities.

## 6. FDA Validation Plan

**Patient Population Description for FDA Validation Dataset:**
The ideal data set to receive from a clinical partner for FDA validation would meet the following requirements:

- body part examined: 'CHEST'
- modality: 'DX'
- patient position: 'AP' or 'PA'
- age range: 10 to 75 years
- gender distribution: balanced
- prevalence of pneumonia (and other diseases): 20%

**Ground Truth Acquisition Methodology:**
The consensus labels of three U.S. board-certified radiologists (the majority of votes of three radiologists) were used as the reference standard of "ground truth".
**Algorithm Performance Standard:**

Minimum acceptable F1-score should be 0.387 because I found in the CheXNet paper (https://stanfordmlgroup.github.io/projects/chexnet/ that this is the average radiologist F1-score.
The proposed algorithm achieves an F1 score of 0.5 on the validation dataset, which exceeds the radiologist performance standard.
Additionally, the algorithm is optimized for recall for use in early detection of Pneumonia.