Validation Plan

ChainRad

AI for Healthcare Hackathon
2021

Table of Contents

1. General Information

Intended Use Statement:

The software assists radiologists by giving the possibilities of trained diseases. However, this software does not contact the patients or with the patient's internal organs, nervous- or cardiovascular system, it’s a diagnostic tool for internal organs.

Indications for Use:

Indications for use is screening of ‘PA’ or ‘AP’ positioned chest X-ray images from a target population and from both biological sex with no prior history of lung defectiveness or lung surgery. The X-ray images shall be equals or higher than 224 x 224 pixels.

Device Limitations:

Software is recommended to diagnose chest x-ray images of patients who has both lung. It can detect only 14 different diseases:

- Atelectasis
- Cardiomegaly
- Consolidation
- Edema
- Effusion
- Emphysema
- Fibrosis
- Hernia
- Infiltration
- Nodule
- Mass
- Pleural Thickening
- Pneumonia
- Pneumothorax

Radiologists should review the result when a patient has a type of disease that has not been part of training. The histograms of consolidation, edema, effusion, emphysema are similar to pneumonia, which is a really common disease. The COVID-19 also has a very similar histogram spectrum. These similarities can lead to false positive diagnosis. It is the known source of error.

Related to some diseases, the model has lower F1 score than an average human radiologist. The following diseases are affected:

- emphysema
- fibrosis
- infiltration
- nodule
- pleural thickening
- pneumothorax

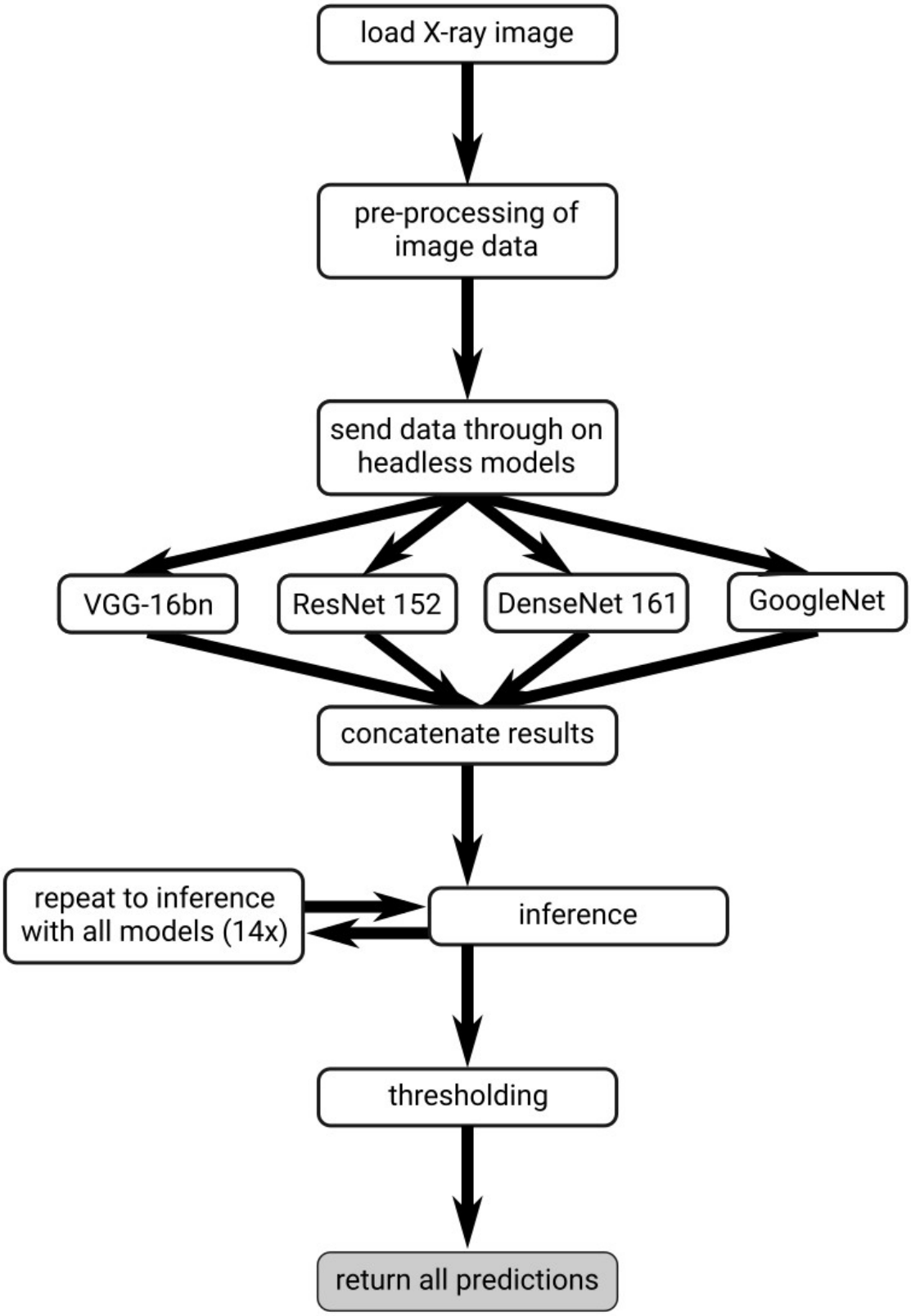
The model works well with negative cases, but fails with positive ones. The age of the patient shall be between 21 and 90 years, that's why the model cannot be used on X-rays of child patients. The structure of the human body is changing over the years and the model wasn't trained on young lungs. Older patients regularly have several diseases and the model wasn't trained on them.

The model is written in PyTorch, so the running machine has to meet the system requirements of the pytorch framework. Model cannot be trained or fine tuned on the machines of hospitals. The model is trained on Intel(R) Core i3-4160 3,6GHz CPU and Nvidia GT1060 GPU with 4 GB GPU-RAM. The computational time of image pre-processing depends on the performance of CPU, the time of inference depends on CPU and GPU.

Clinical Impact of Performance:

Software is recommended for assisting a radiologist screening images. The model predict negative cases well, but has some troubles with positive cases. The software can help to prioritize cases or in differential diagnosis. However, predicted negative cases shall be reviewed by a radiologist. As an other use-case according to the model performance, that is showed later in this document, it is advised to use the software as a second check. It is useful when there is no possible way to get a second opinion within a short time range.

2. Algorithm Design and Function



Preprocessing Steps:

The pre-processor function of algorithm makes the following preprocessing steps.

- Resize the image from original size to 224 x 224 pixels. This size is necessary as input of the headless pretrained networks.
- Normalize the values based on the standard method for headless pretrained models with the following parameters: mean= [0.485, 0.456, 0.406], std=[0.229, 0.224, 0.225]. These values refer to the 3 dimension tensor.

CNN Architecture:

The algorithm has 2 major parts, like base models and head model. The base models contain headless pretrained networks such as: VGG-16 batchnorm, ResNet 152, DenseNet 161, GoogleNet. The headless means each network includes every layer from first till the last pooling layer without the last classification layers, since they are removed. The results of each headless network are concatenated to get only one tensor for input of head model.

The head model is connected to base model via directly, without any pooling layer. After the flatten mechanism, there are 4 blocks of fully connected layers activated with relu function and with dropout.

The network output is a single node with a sigmoid activation function, that gives a probability value. This value is a core of binary classification.

3. Algorithm Training

Parameters:

Types of augmentation used during training. All of them is random:

- horizontal flip: 0.1
- translate (equivalent with height and width shifts): 0.1, 0.1
- rotation angle range between 0 and 13 degrees
- shear range: 0.1
- zoom range between 0.9 and 1.1

Batch size: **128**

Optimizer learning rate: **5e-6**

Layers data

Layers of pre-existing architecture that were frozen

All layers from the pretrained networks (VGG-16 batchnorm, ResNet 152, DenseNet 161, GoogleNet) without their own classification layers. The last classification blocks are removed.

Layers of pre-existing architecture that were fine-tuned

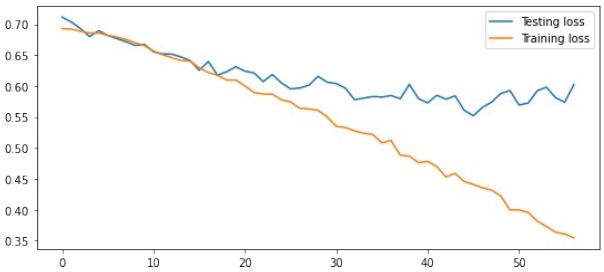
None of the layers from pretrained networks is fine-tuned.

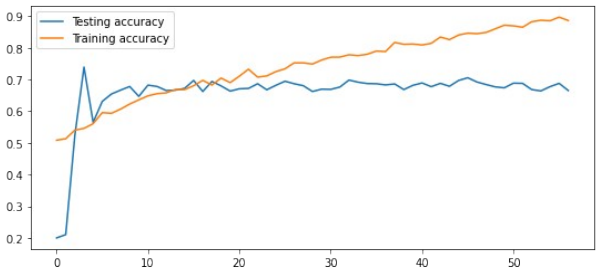
Layers added to pre-existing architecture

- Concat
- Flatten()
- Linear(30368, 2048)
- Dropout(p=0.5, inplace=True)
- ReLU(inplace=True)
- Linear(2048, 256)
- Dropout(p=0.5, inplace=True)
- ReLU(inplace=True)
- Linear(256, 32)
- Dropout(p=0.5, inplace=True)
- ReLU(inplace=True)
- Linear(32, 1)
- Activation('sigmoid')

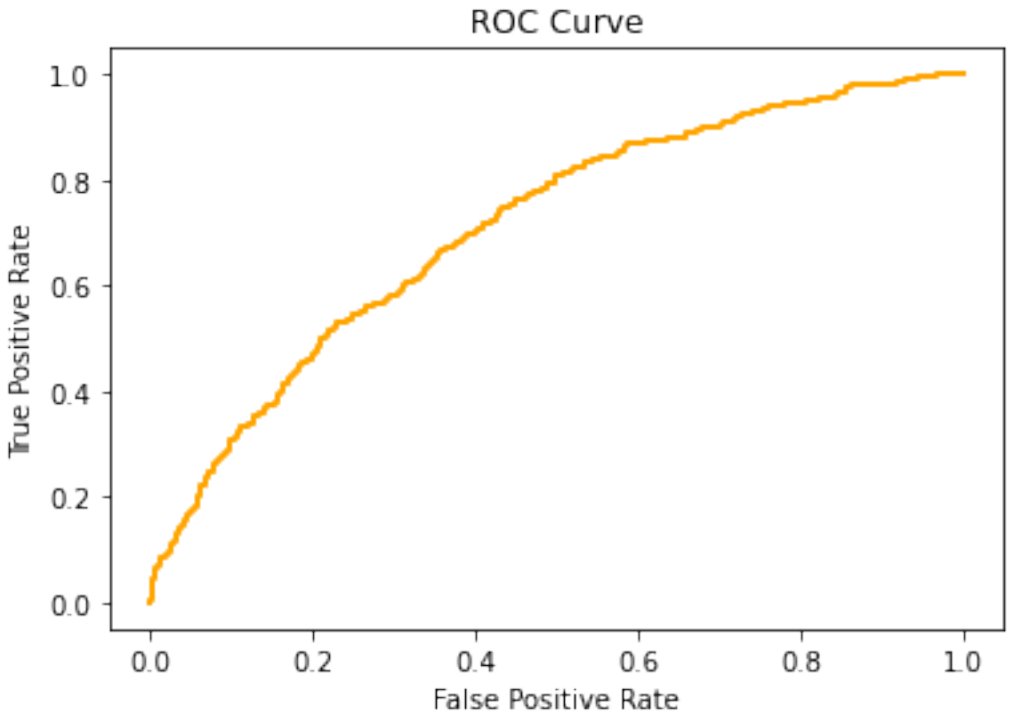
Atelectasis

Algorithm training performance visualization

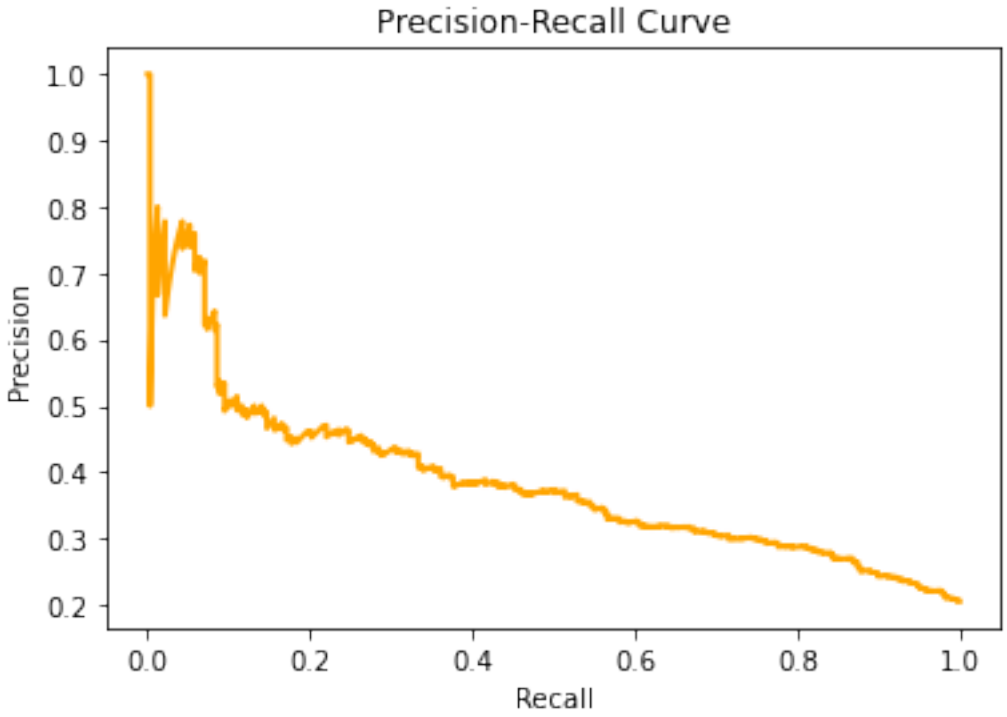


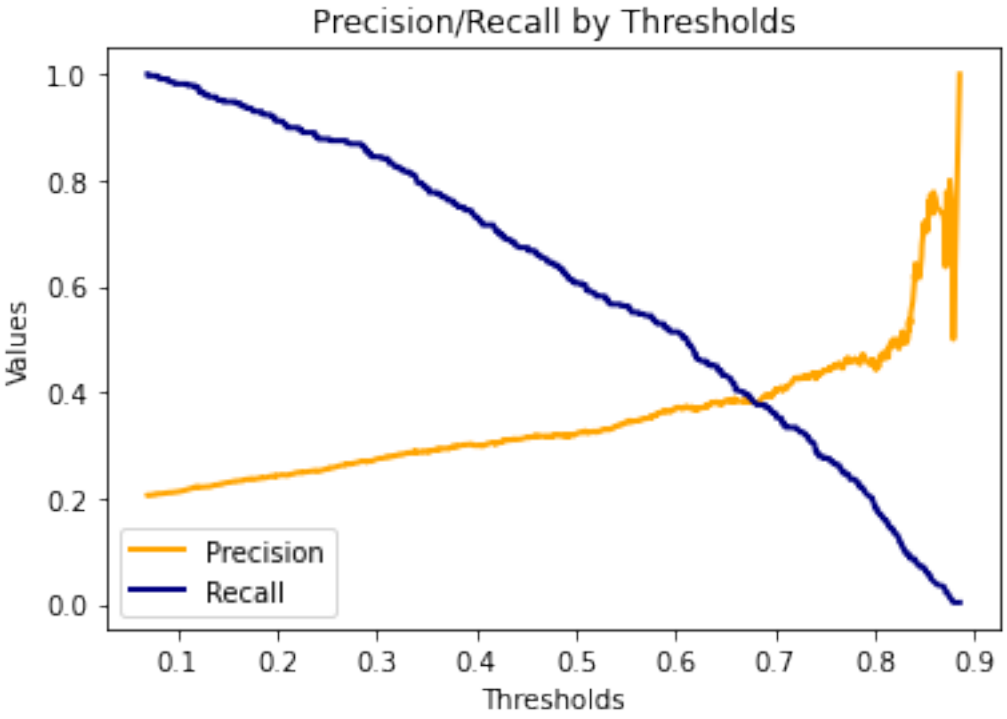


ROC curve shows the model learned something from training dataset. It will be better with smaller learning rate and longer training. This is the part of a future experiment.

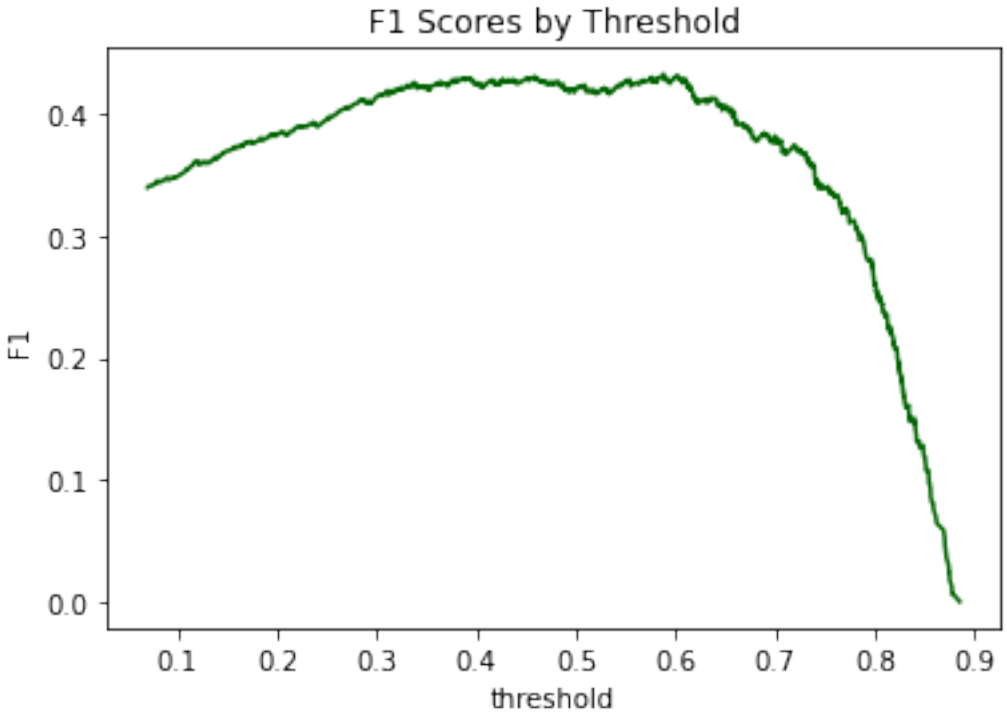


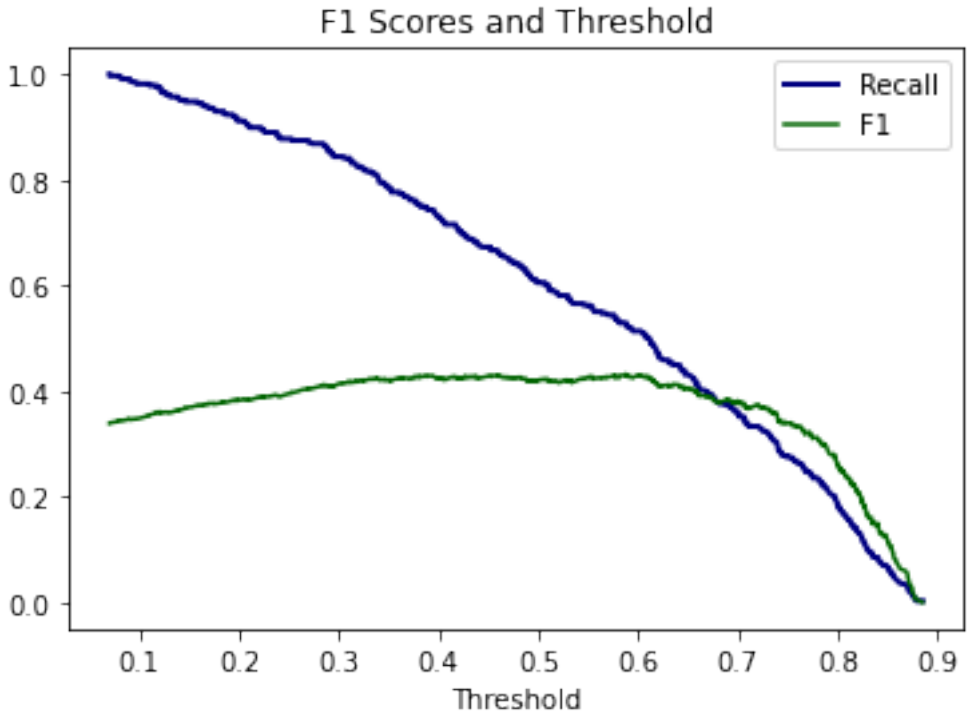
P-R curve


Precision-Recall Curve
Precision
Recall
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.0
0.2
0.4
0.6
0.8
1.0


Precision/Recall by Thresholds
Values
Thresholds
Precision
Recall
1.0
0.8
0.6
0.4
0.2
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9

Final Threshold and Explanation:





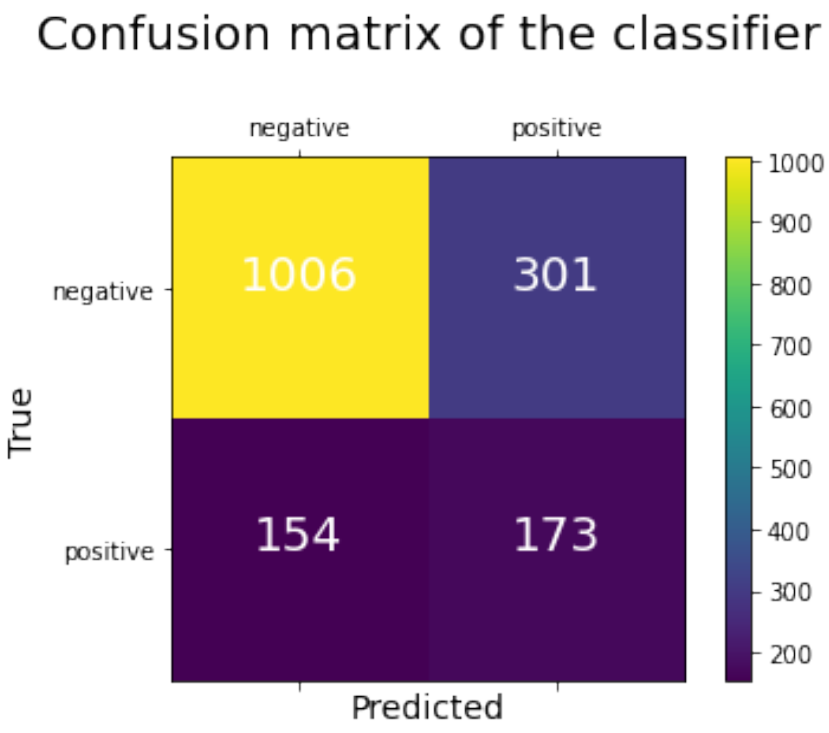
The final threshold is **0.5875**. This is the point where the F1 score graph reaches its maximum value.

Person / Algorithm	F1	95% CI min	95% CI max
Radiologist 1	0.383	0.309	0.453
Radiologist 2	0.356	0.282	0.428
Radiologist 3	0.365	0.291	0.435
Radiologist 4	0.442	0.390	0.492
Radiologist average	0.387	0.330	0.442
CheXNet	0.435	0.387	0.481
ChainRad	**0.433**		

The average F1 score of human radiologists is 0.387 according to paper of CheXNet, that is available here: https://arxiv.org/pdf/1711.05225.pdf

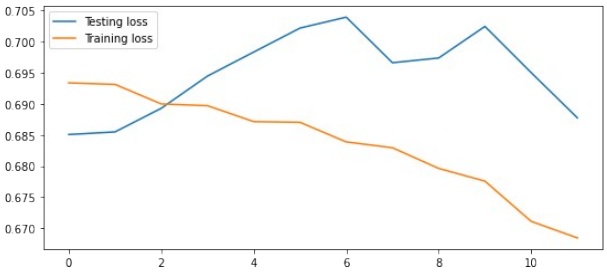
The main goal is for the model to reach the average F1 score of human radiologists. There are the major measures related to the final threshold.

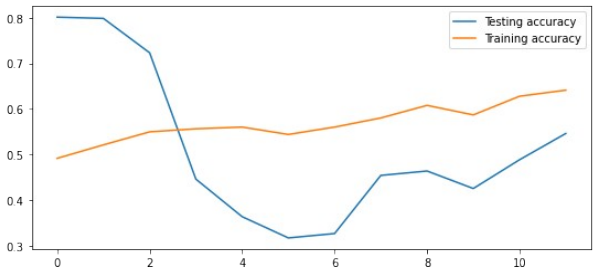
- F1 score max: **0.4325**
- Threshold: **0.5875**
- Precision: **0.3650**
- Recall: **0.5291**



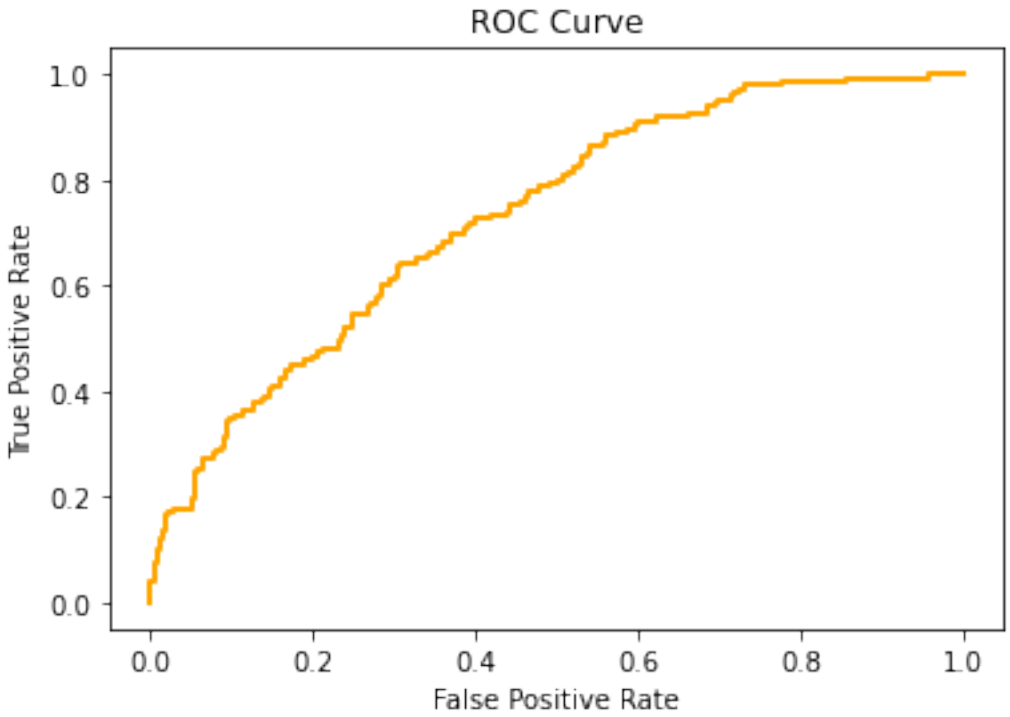
Cardiomegaly

Algorithm training performance visualization

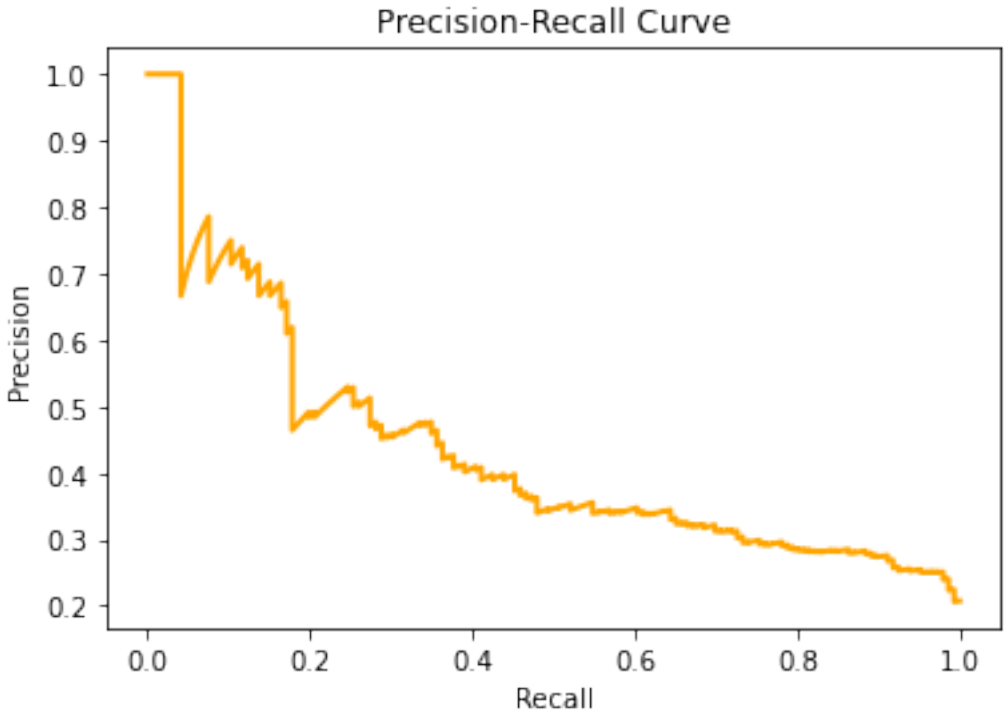


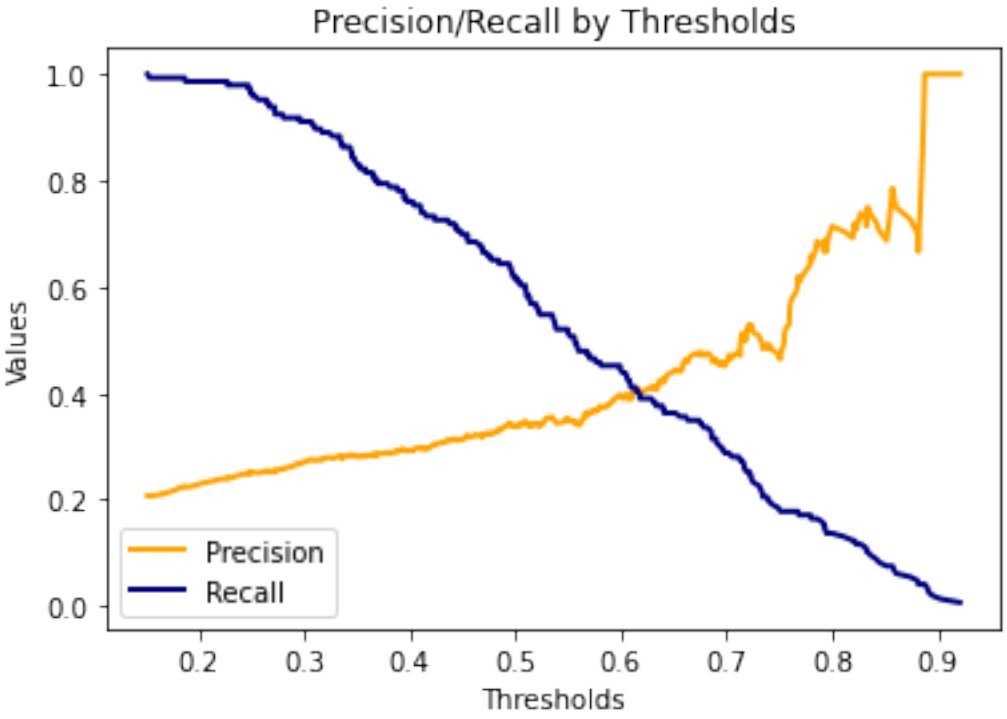


ROC curve shows the model learned something from training dataset. It will be better with smaller learning rate and longer training. This is the part of a future experiment.

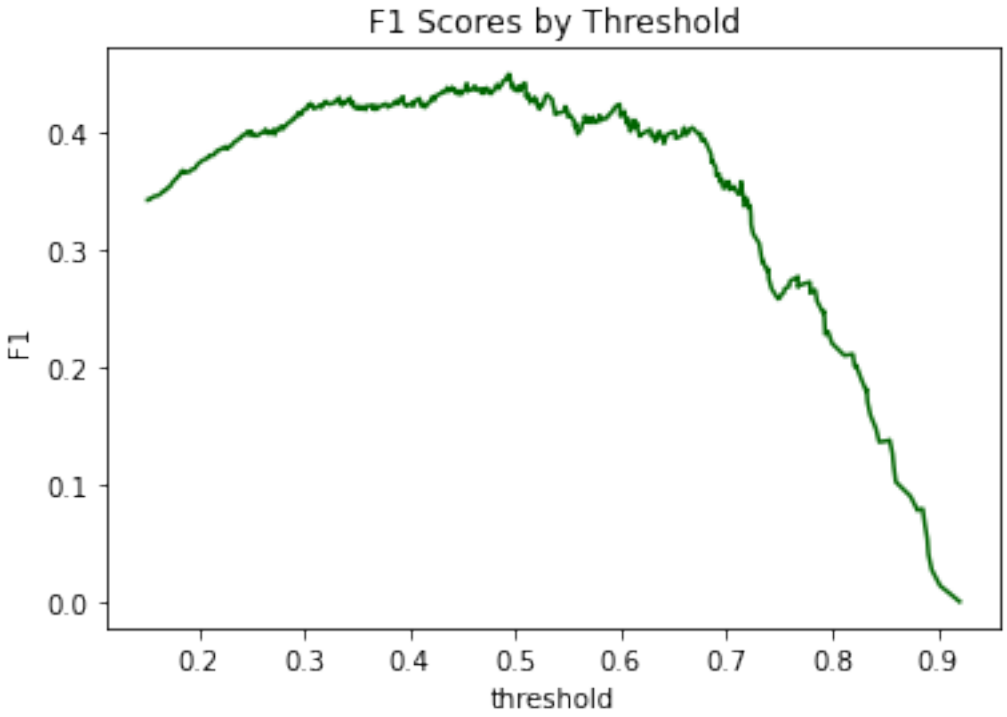


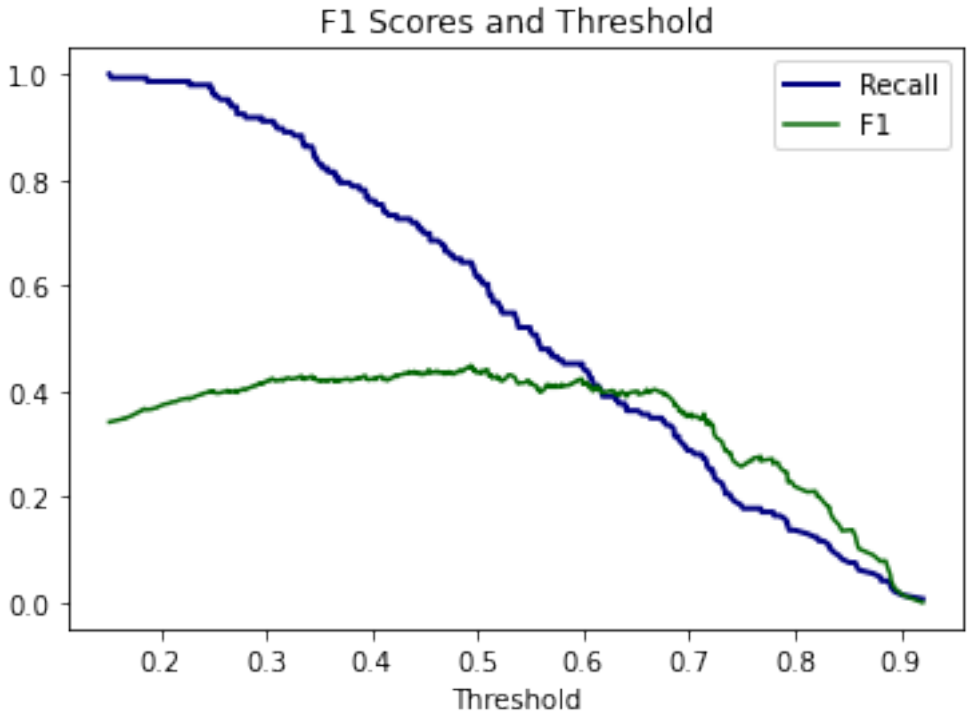
P-R curve


Precision-Recall Curve
Precision
Recall
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.0
0.2
0.4
0.6
0.8
1.0


Precision/Recall by Thresholds
Values
Thresholds
Precision
Recall
1.0
0.8
0.6
0.4
0.2
0.0
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9

Final Threshold and Explanation:





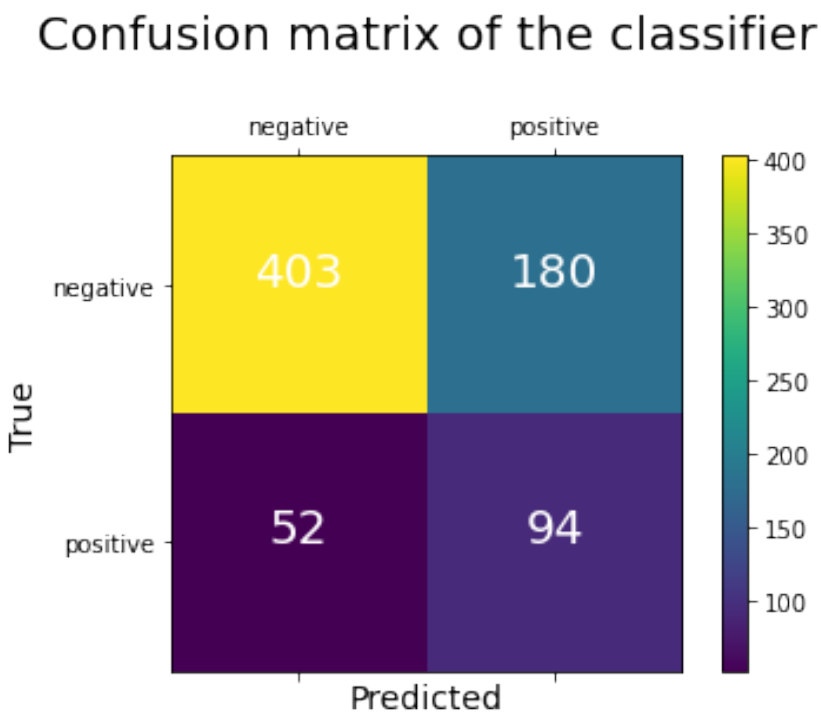
The final threshold is **0.4940**. This is the point where the F1 score graph reaches its maximum value.

Person / Algorithm	F1	95% CI min	95% CI max
Radiologist 1	0.383	0.309	0.453
Radiologist 2	0.356	0.282	0.428
Radiologist 3	0.365	0.291	0.435
Radiologist 4	0.442	0.390	0.492
Radiologist average	0.387	0.330	0.442
CheXNet	0.435	0.387	0.481
ChainRad	**0.449**		

The average F1 score of human radiologists is 0.387 according to paper of CheXNet, that is available here: https://arxiv.org/pdf/1711.05225.pdf

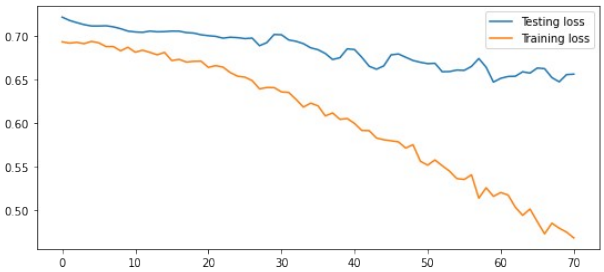
The main goal is for the model to reach the average F1 score of human radiologists. There are the major measures related to the final threshold.

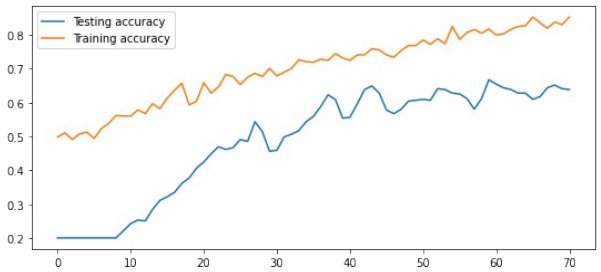
- F1 score max: **0.4487**
- Threshold: **0.4940**
- Precision: **0.3431**
- Recall: **0.6438**



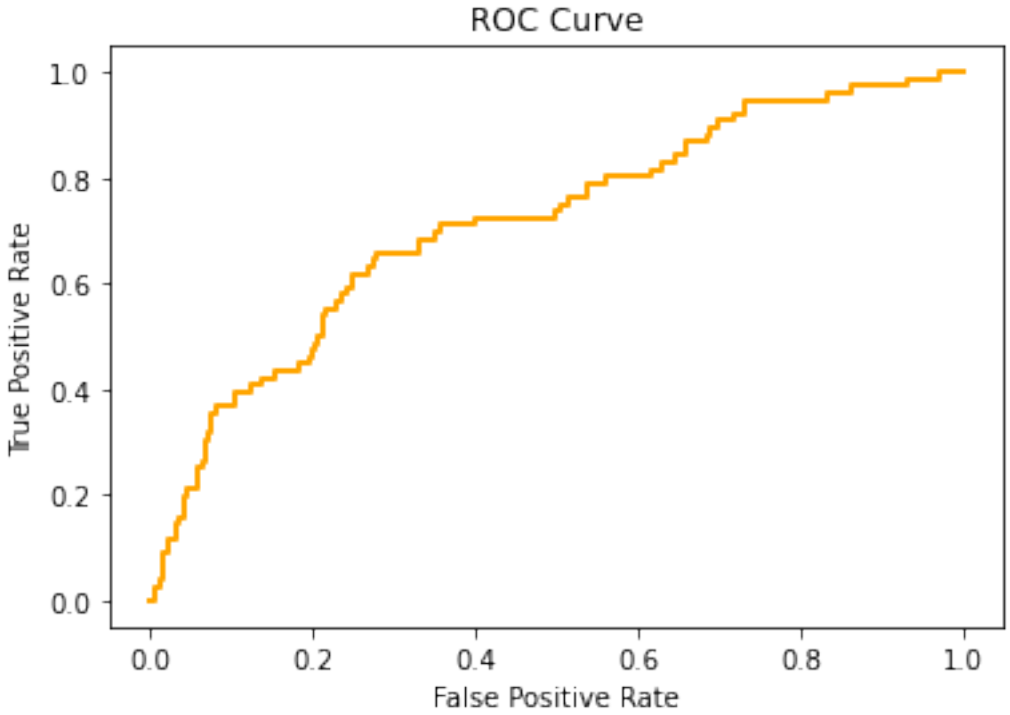
Consolidation

Algorithm training performance visualization

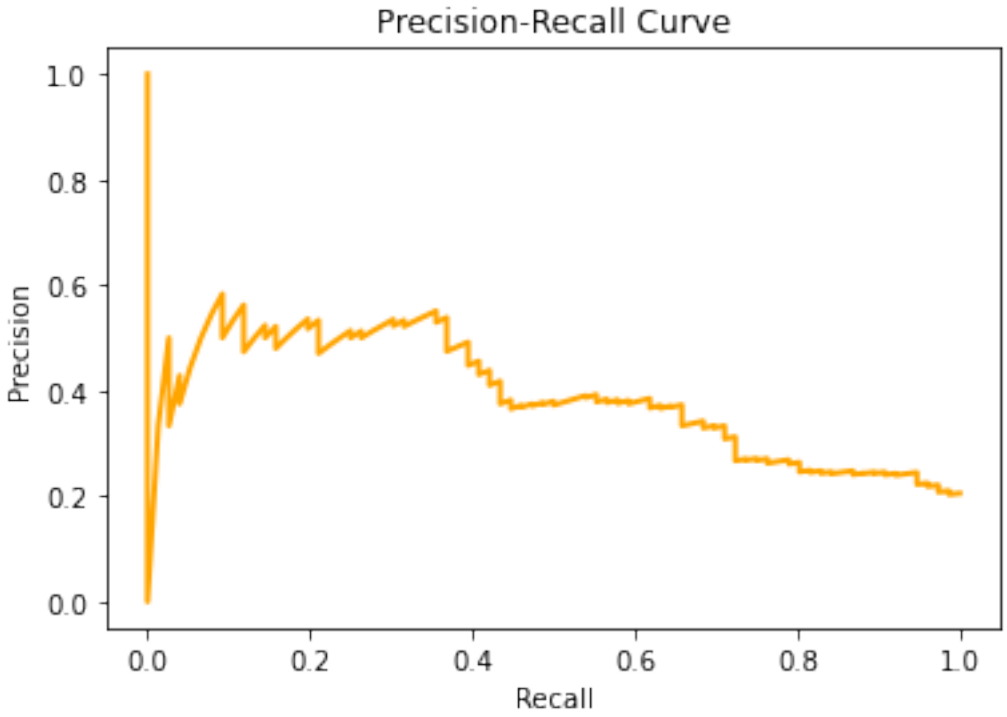


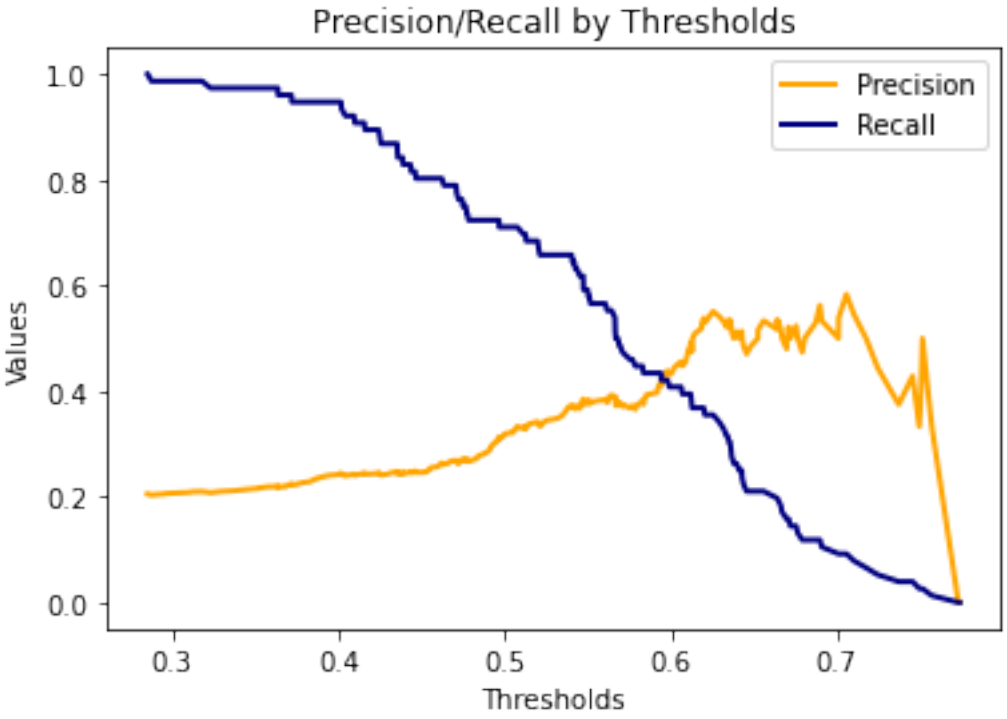


ROC curve shows the model learned something from training dataset. It will be better with smaller learning rate and longer training. This is the part of a future experiment.

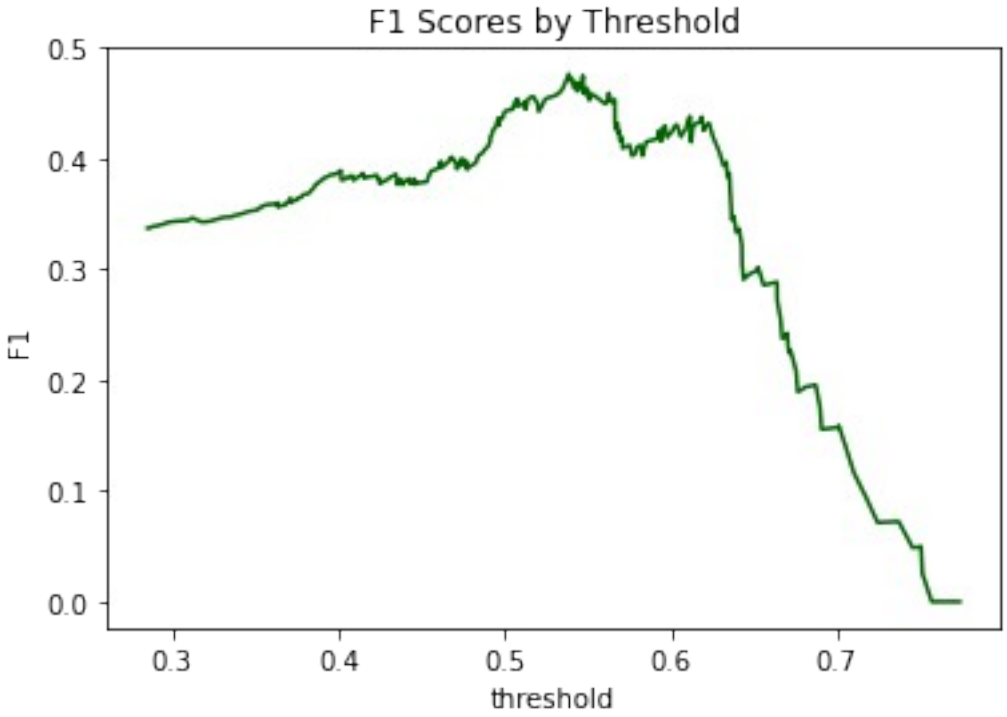


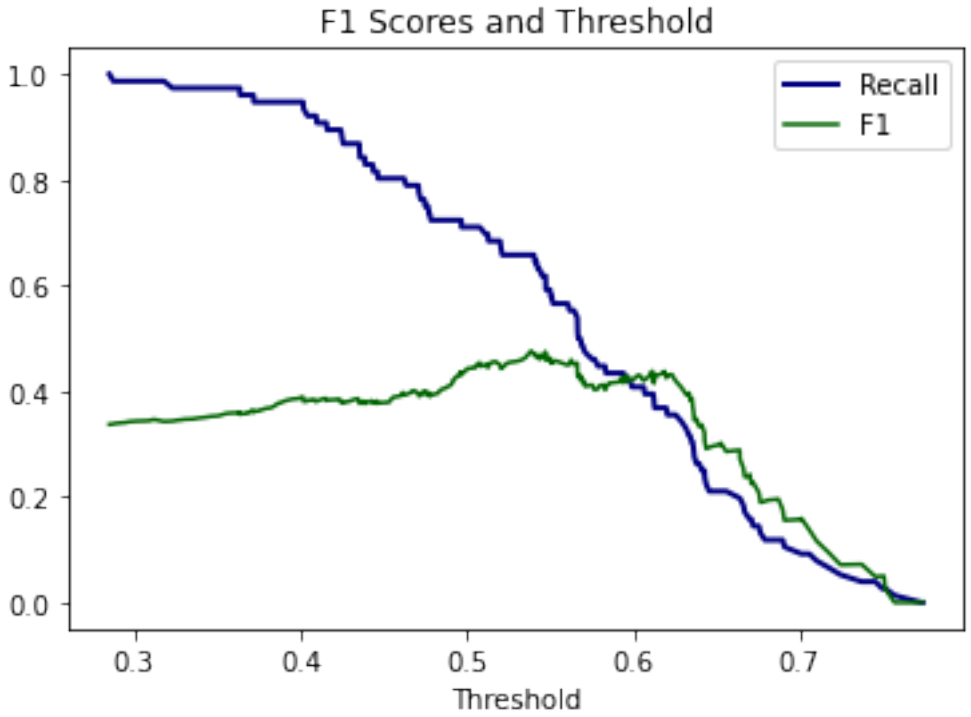
P-R curve


Precision-Recall Curve
Precision
Recall
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0


Precision/Recall by Thresholds
Precision
Recall
Values
Thresholds
1.0
0.8
0.6
0.4
0.2
0.0
0.3
0.4
0.5
0.6
0.7

Final Threshold and Explanation:





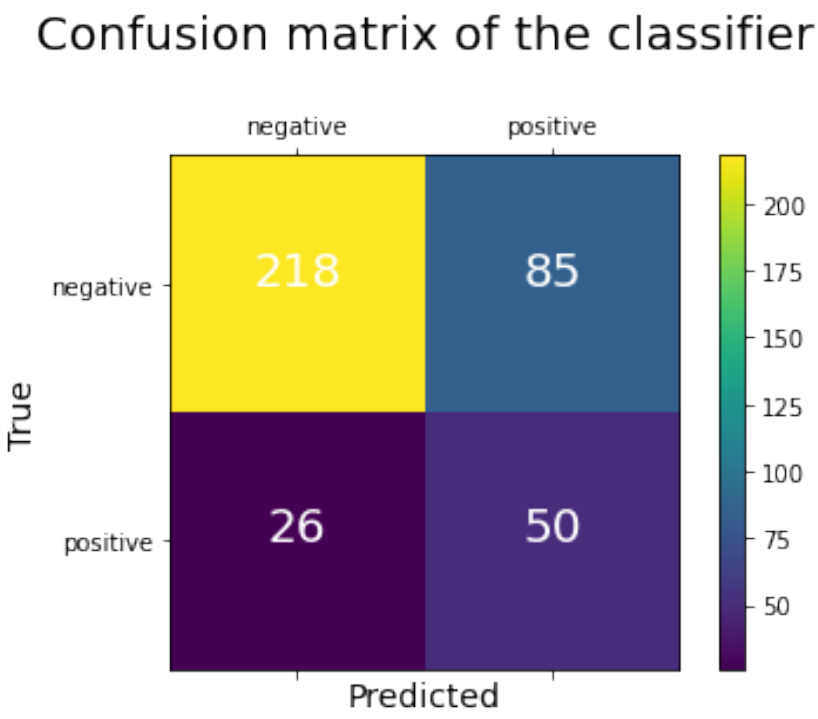
The final threshold is **0.5386**. This is the point where the F1 score graph reaches its maximum value.

Person / Algorithm	F1	95% CI min	95% CI max
Radiologist 1	0.383	0.309	0.453
Radiologist 2	0.356	0.282	0.428
Radiologist 3	0.365	0.291	0.435
Radiologist 4	0.442	0.390	0.492
Radiologist average	0.387	0.330	0.442
CheXNet	0.435	0.387	0.481
ChainRad	**0.476**		

The average F1 score of human radiologists is 0.387 according to paper of CheXNet, that is available here: https://arxiv.org/pdf/1711.05225.pdf

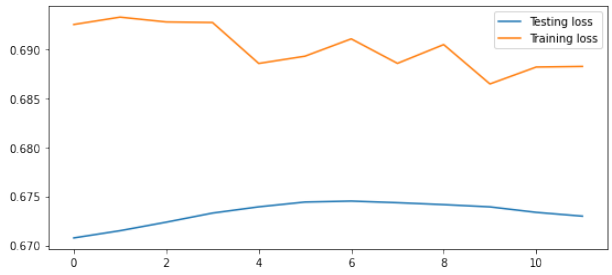
The main goal is for the model to reach the average F1 score of human radiologists. There are the major measures related to the final threshold.

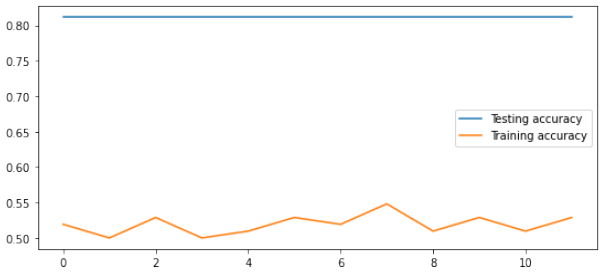
- F1 score max: **0.4762**
- Threshold: **0.5386**
- Precision: **0.3704**
- Recall: **0.6579**



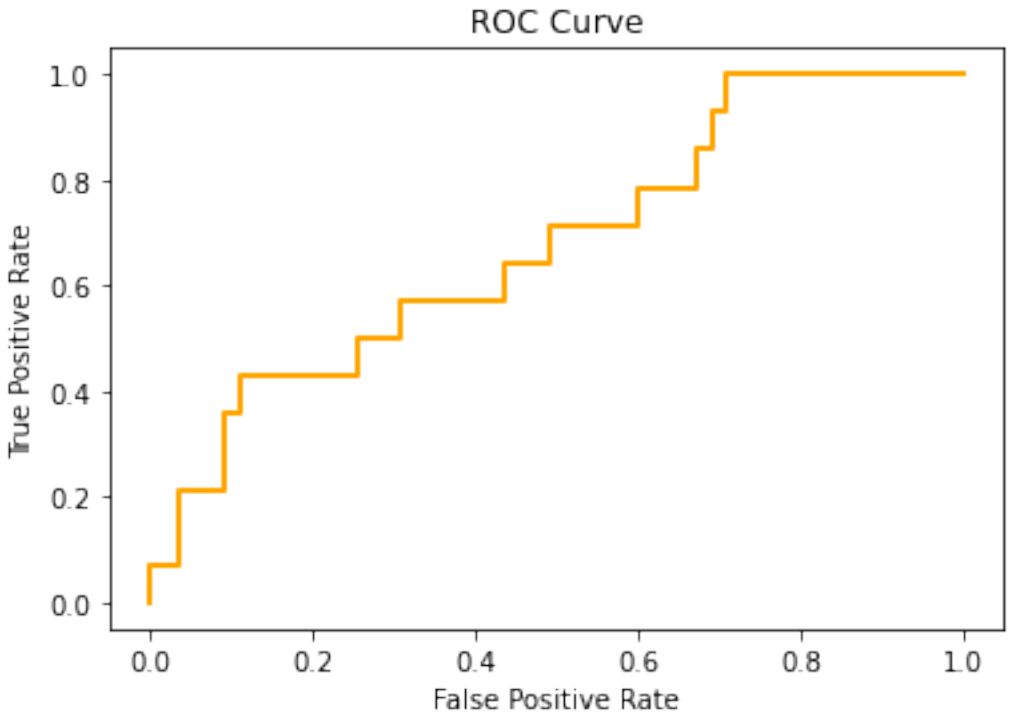
Edema

Algorithm training performance visualization

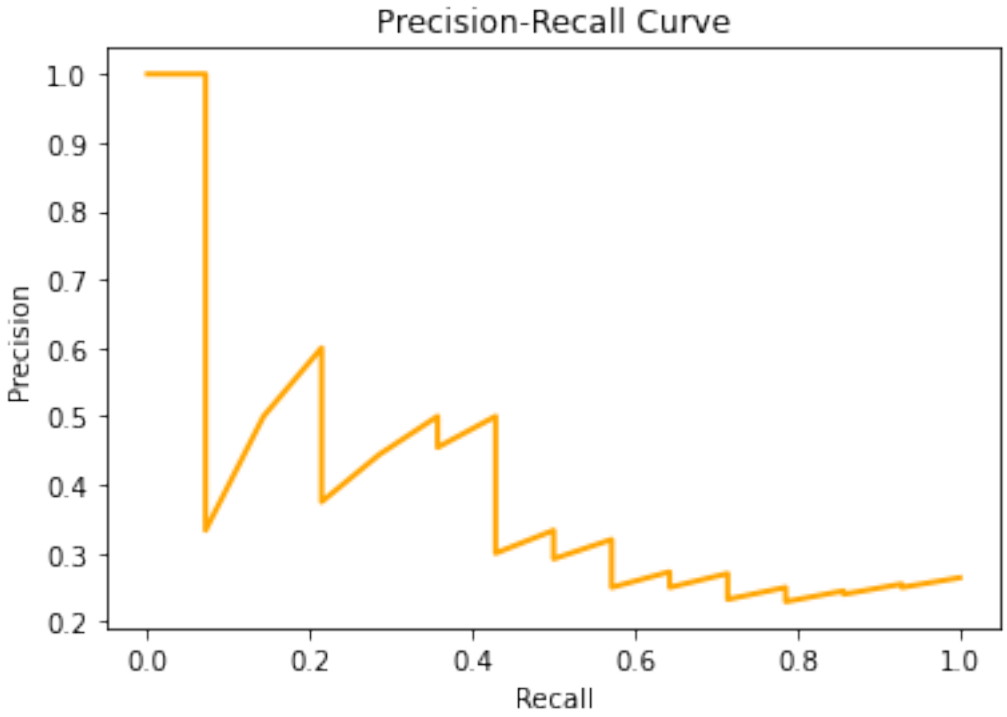


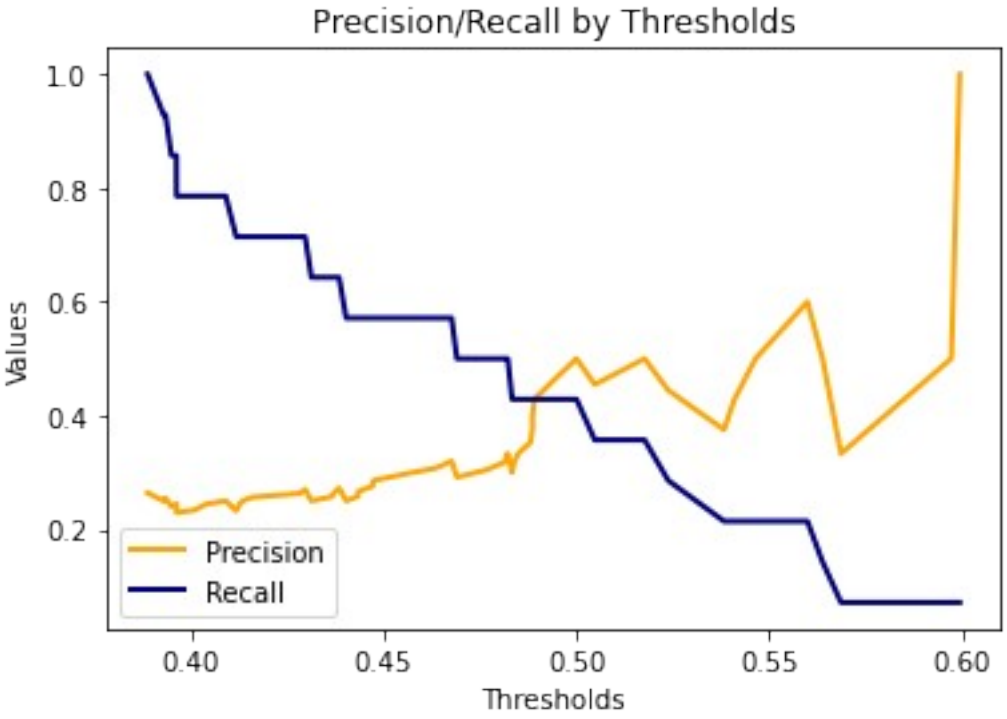


ROC curve shows the model learned something from training dataset. It will be better with smaller learning rate and longer training. This is the part of a future experiment.

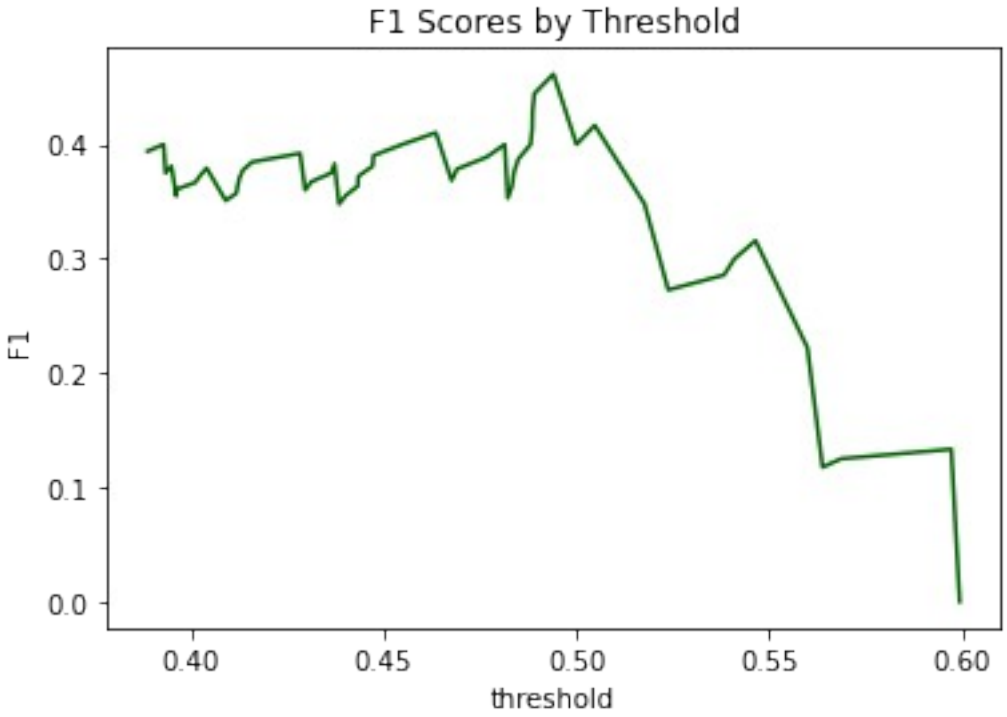


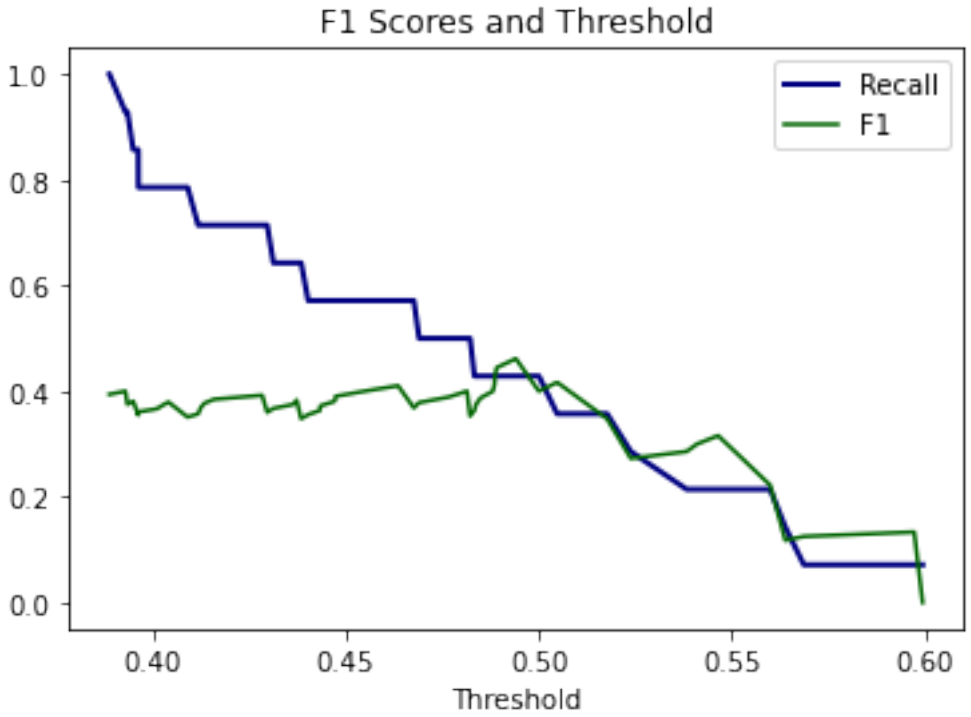
P-R curve


Precision-Recall Curve
Precision
Recall
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.0
0.2
0.4
0.6
0.8
1.0


Precision/Recall by Thresholds
Values
Thresholds
Precision
Recall
1.0
0.8
0.6
0.4
0.2
0.40
0.45
0.50
0.55
0.60

Final Threshold and Explanation:





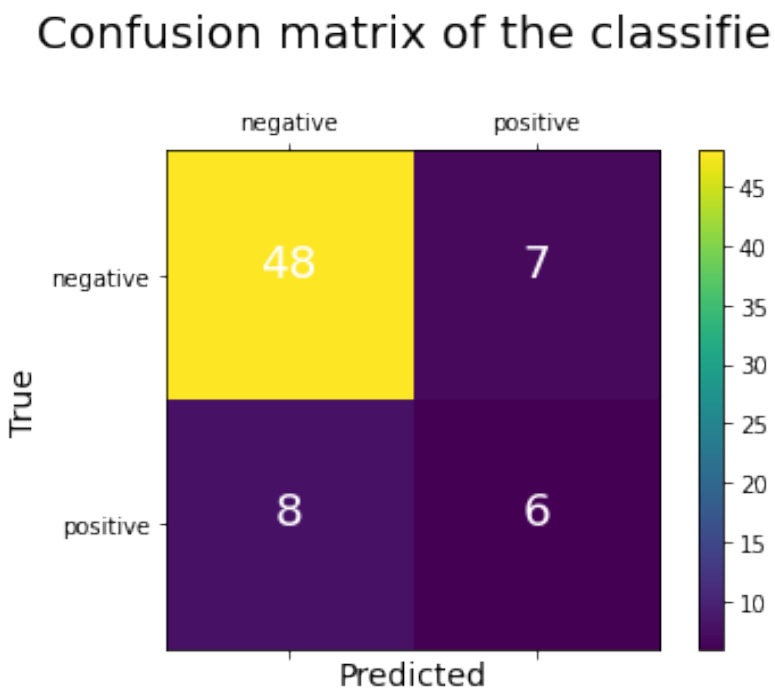
The final threshold is **0.4941**. This is the point where the F1 score graph reaches its maximum value.

Person / Algorithm	F1	95% CI min	95% CI max
Radiologist 1	0.383	0.309	0.453
Radiologist 2	0.356	0.282	0.428
Radiologist 3	0.365	0.291	0.435
Radiologist 4	0.442	0.390	0.492
Radiologist average	0.387	0.330	0.442
CheXNet	0.435	0.387	0.481
ChainRad	**0.462**		

The average F1 score of human radiologists is 0.387 according to paper of CheXNet, that is available here: https://arxiv.org/pdf/1711.05225.pdf

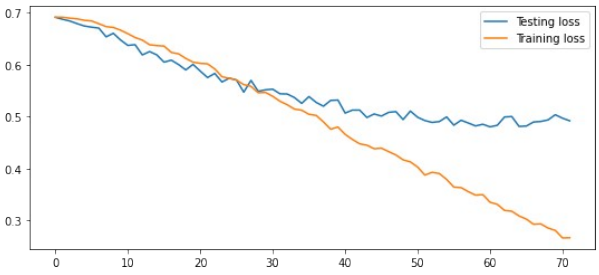
The main goal is for the model to reach the average F1 score of human radiologists. There are the major measures related to the final threshold.

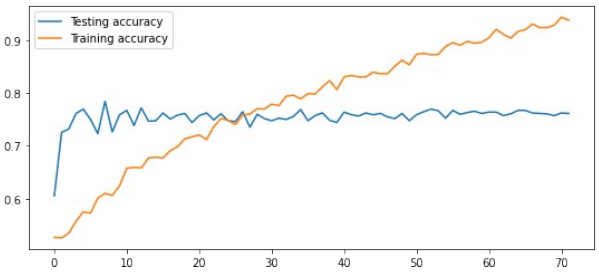
- F1 score max: **0.4615**
- Threshold: **0.4941**
- Precision: **0.4615**
- Recall: **0.4286**



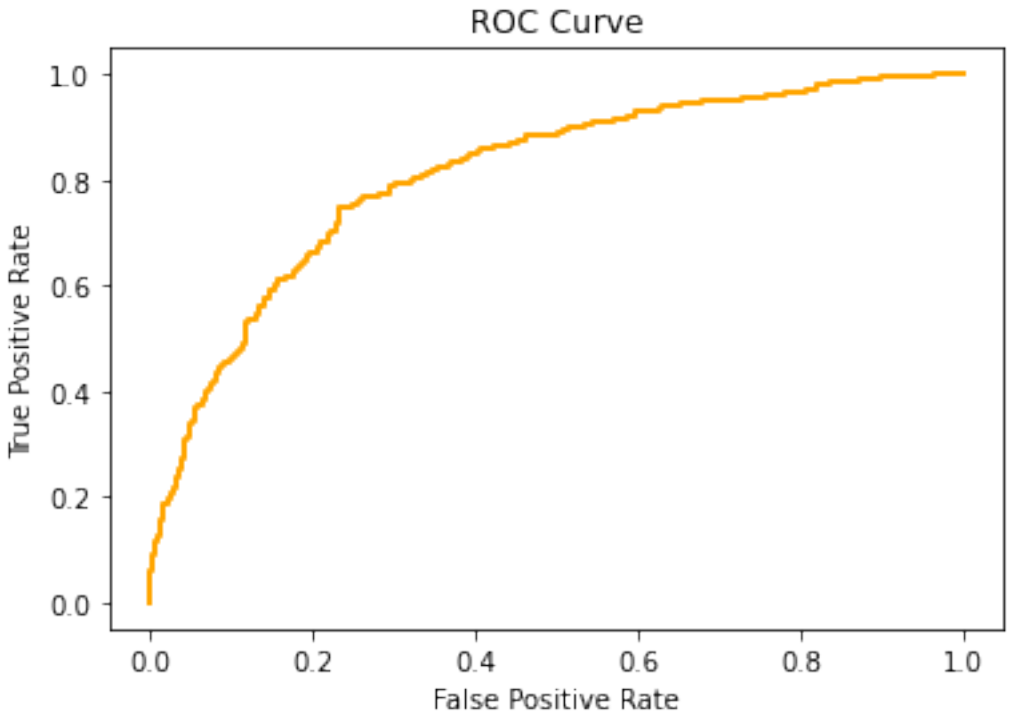
Effusion

Algorithm training performance visualization

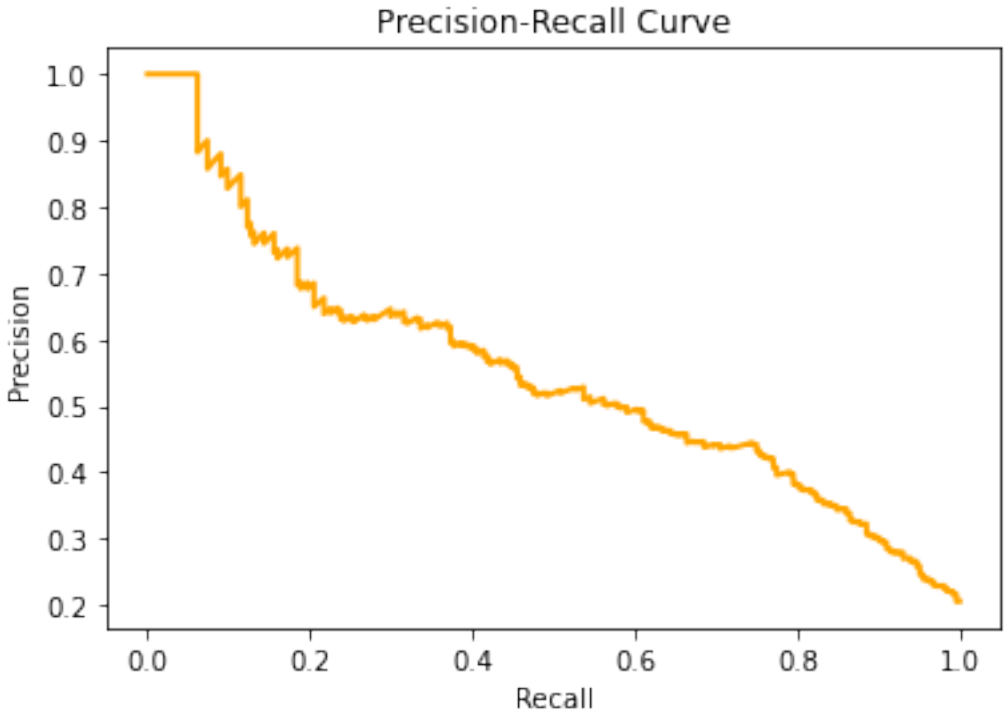


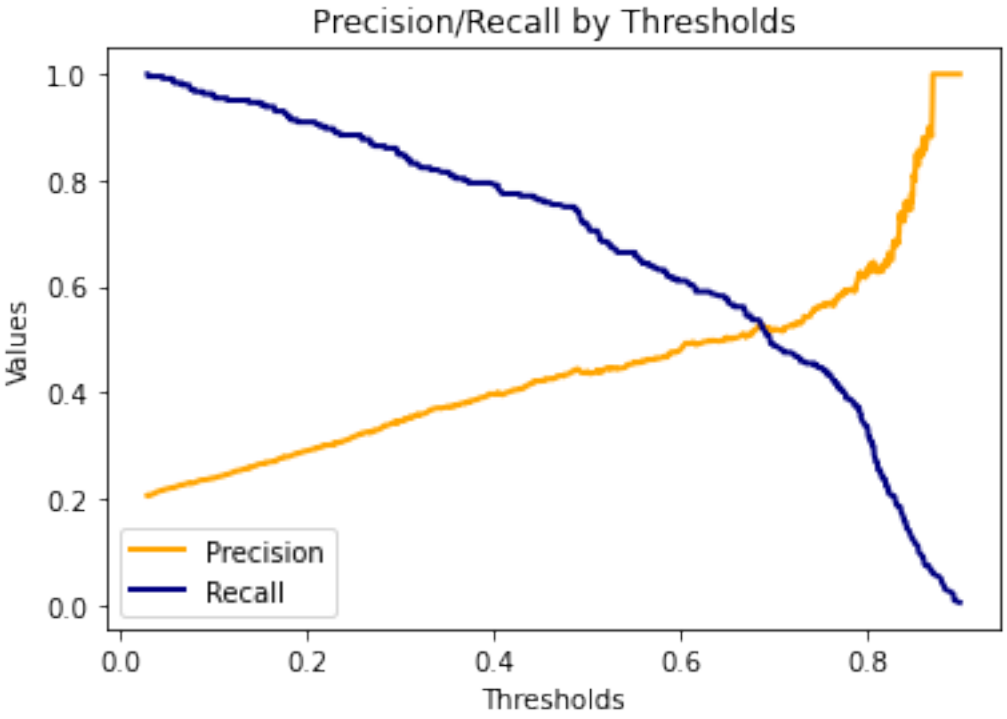


ROC curve shows the model learned something from training dataset. It will be better with smaller learning rate and longer training. This is the part of a future experiment.

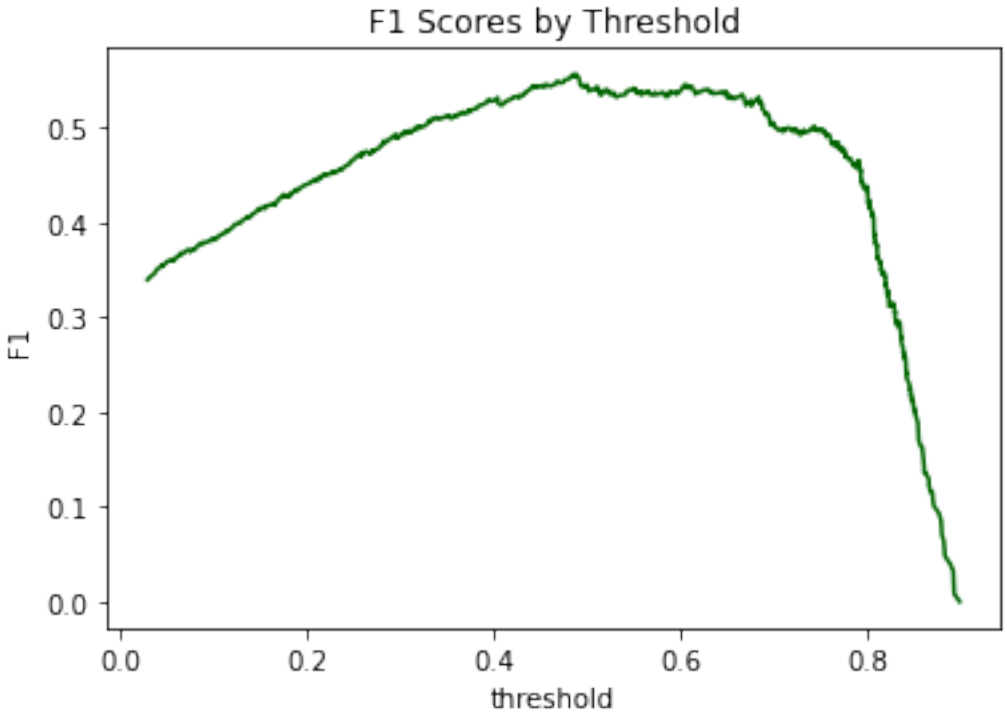


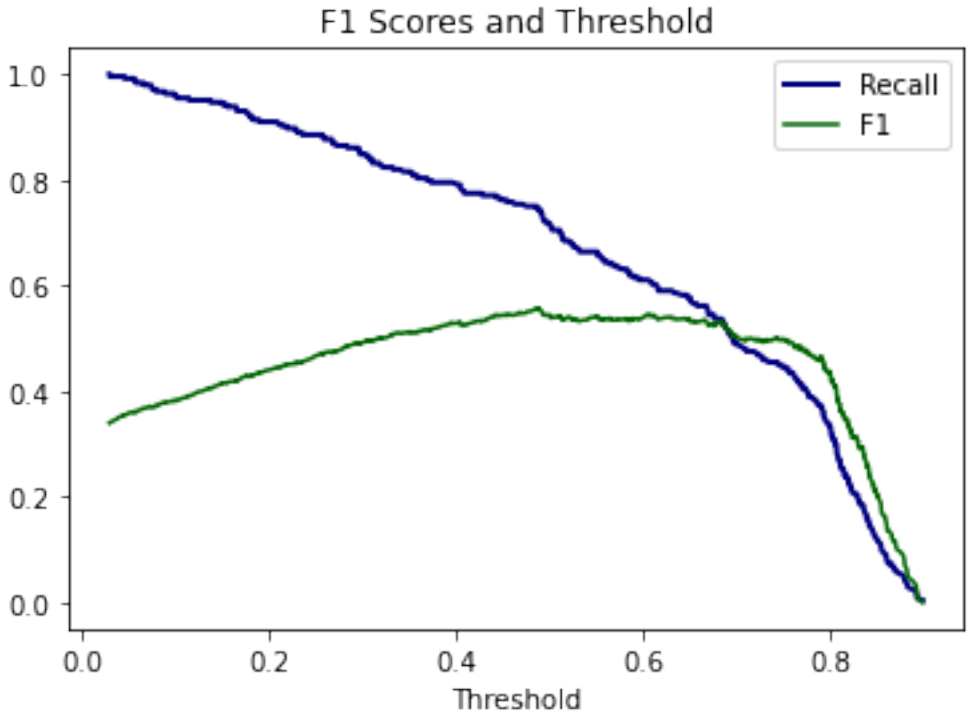
P-R curve


Precision-Recall Curve
Precision
Recall
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.0
0.2
0.4
0.6
0.8
1.0


Precision/Recall by Thresholds
Values
Thresholds
Precision
Recall
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8

Final Threshold and Explanation:





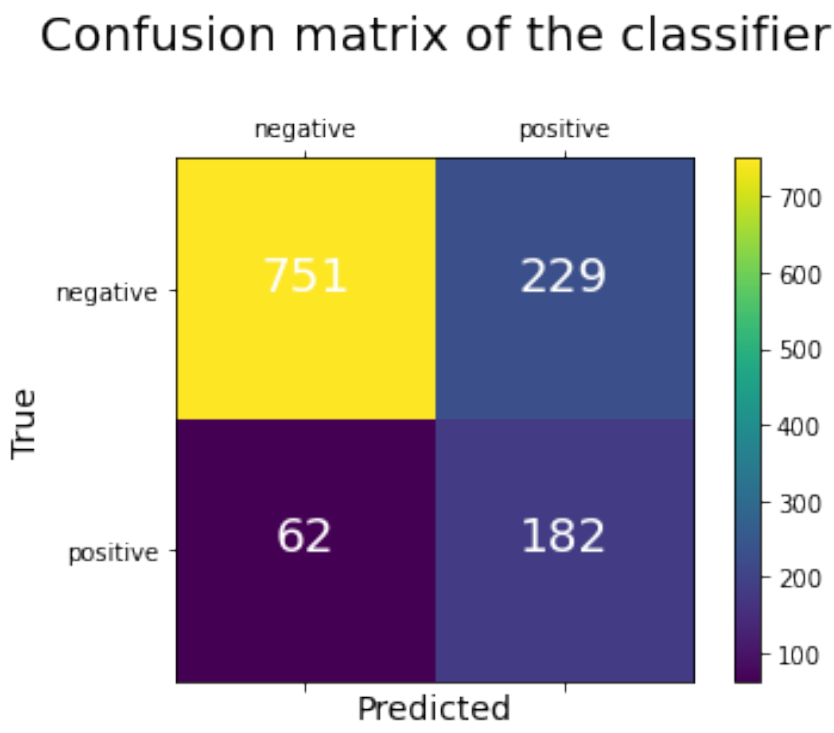
The final threshold is **0.4883**. This is the point where the F1 score graph reaches its maximum value.

Person / Algorithm	F1	95% CI min	95% CI max
Radiologist 1	0.383	0.309	0.453
Radiologist 2	0.356	0.282	0.428
Radiologist 3	0.365	0.291	0.435
Radiologist 4	0.442	0.390	0.492
Radiologist average	0.387	0.330	0.442
CheXNet	0.435	0.387	0.481
ChainRad	**0.557**		

The average F1 score of human radiologists is 0.387 according to paper of CheXNet, that is available here: https://arxiv.org/pdf/1711.05225.pdf

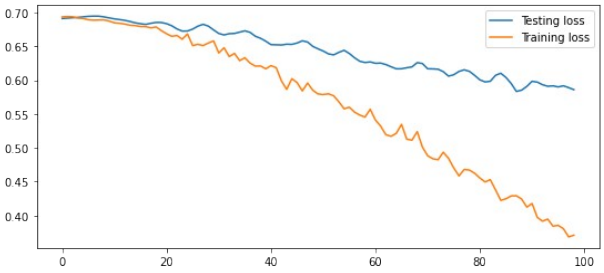
The main goal is for the model to reach the average F1 score of human radiologists. There are the major measures related to the final threshold.

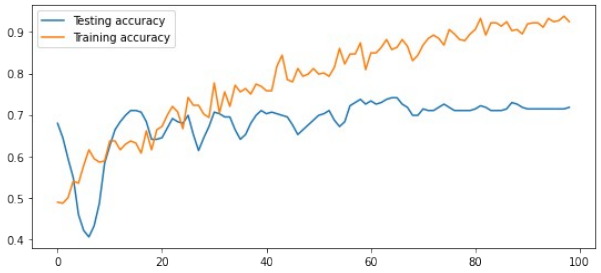
- F1 score max: **0.5566**
- Threshold: **0.4883**
- Precision: **0.4428**
- Recall: **0.7459**



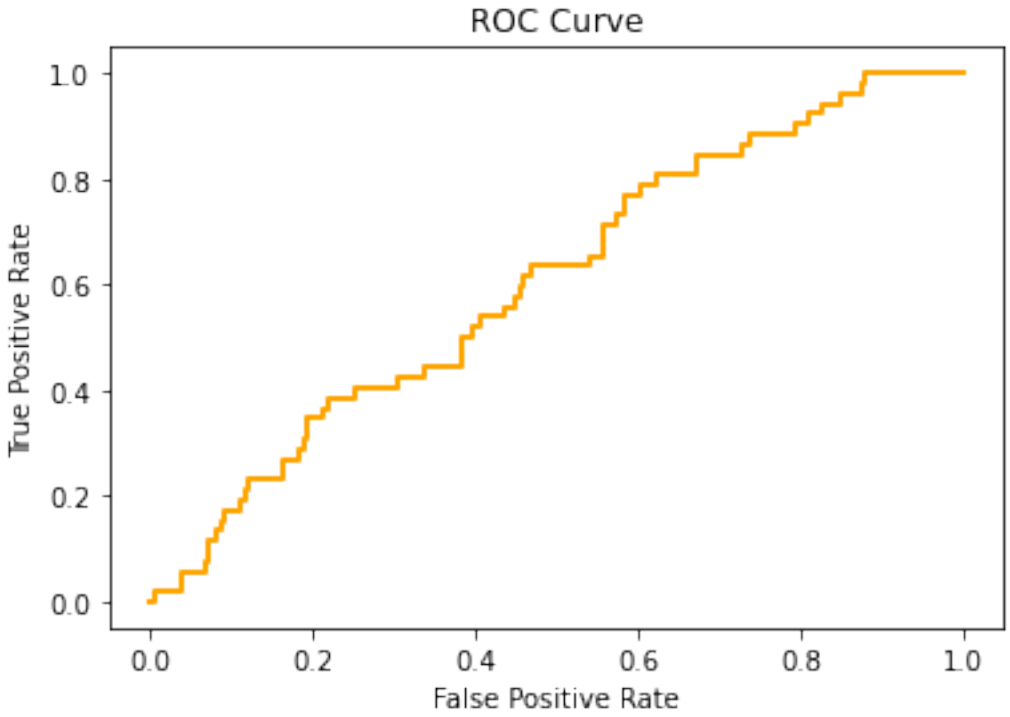
Emphysema

Algorithm training performance visualization

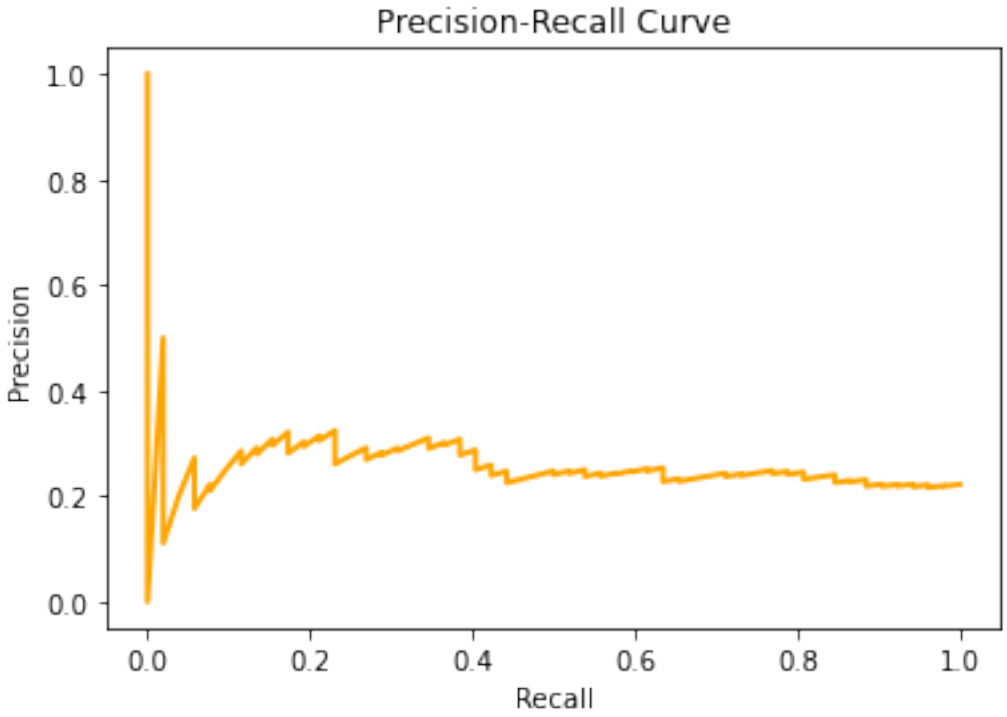


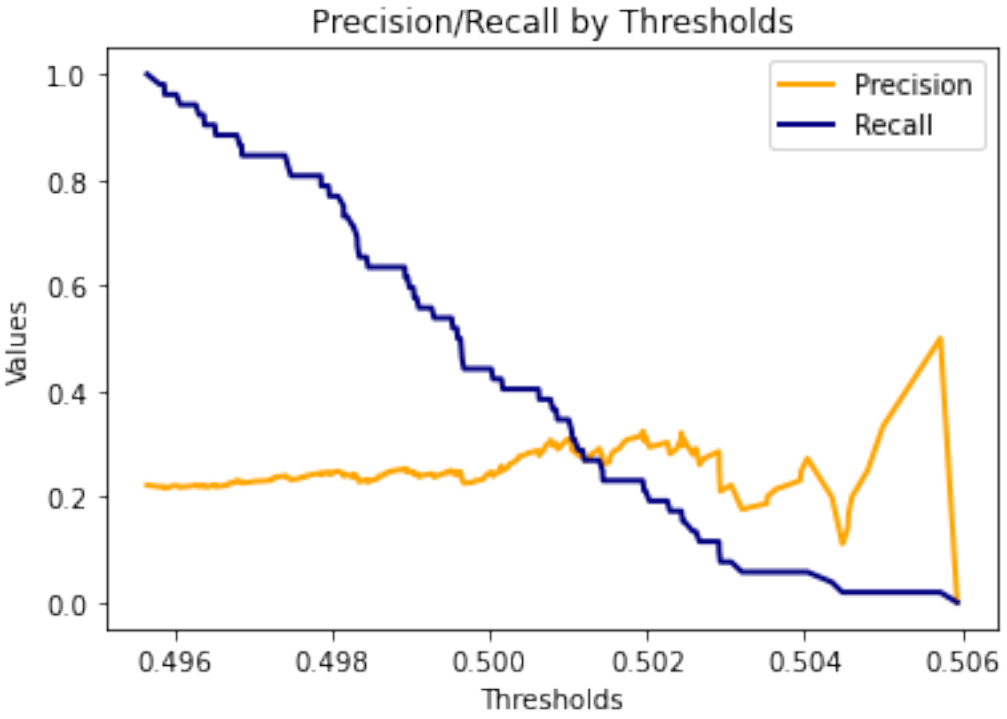


ROC curve shows the model learned something from training dataset. It will be better with smaller learning rate and longer training. This is the part of a future experiment.

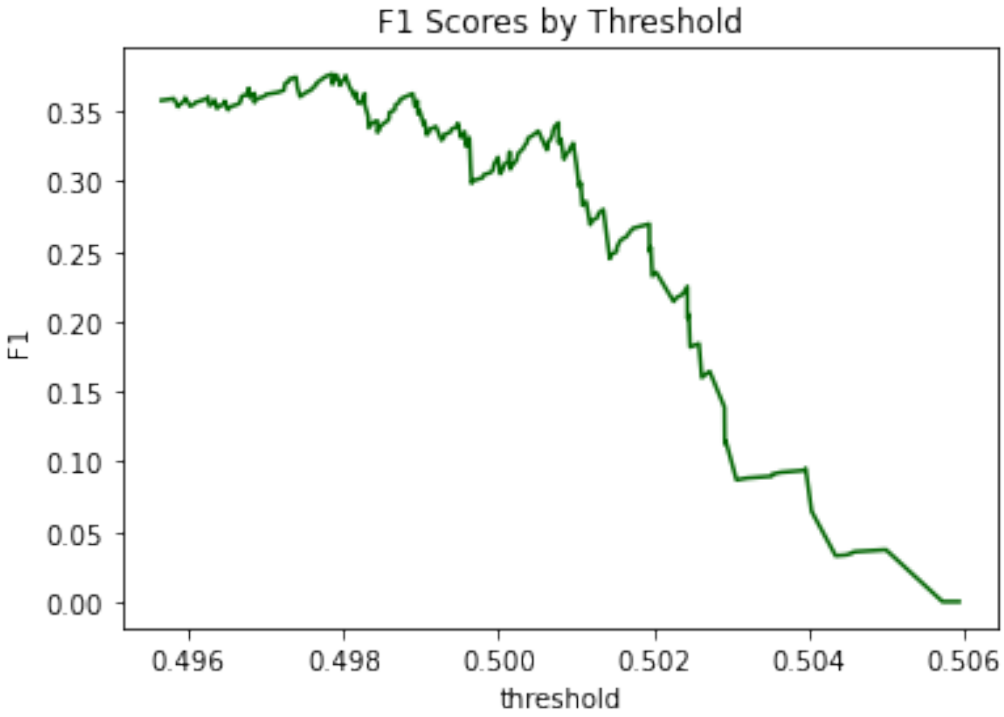


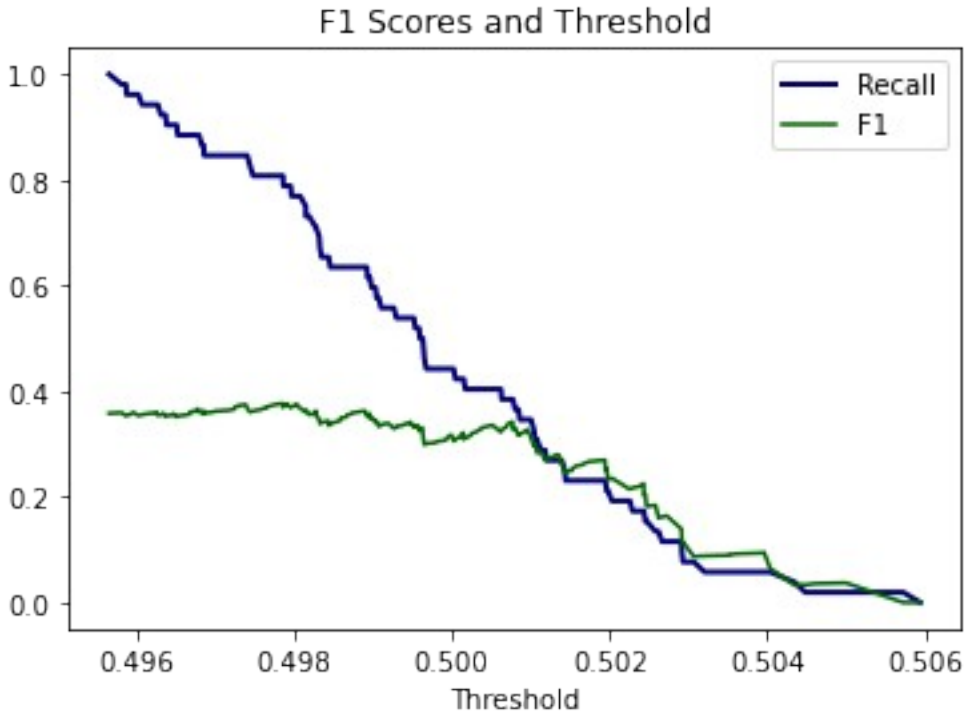
P-R curve


Precision-Recall Curve
Precision
Recall
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0


Precision/Recall by Thresholds
Precision
Recall
Values
Thresholds
1.0
0.8
0.6
0.4
0.2
0.0
0.496
0.498
0.500
0.502
0.504
0.506

Final Threshold and Explanation:





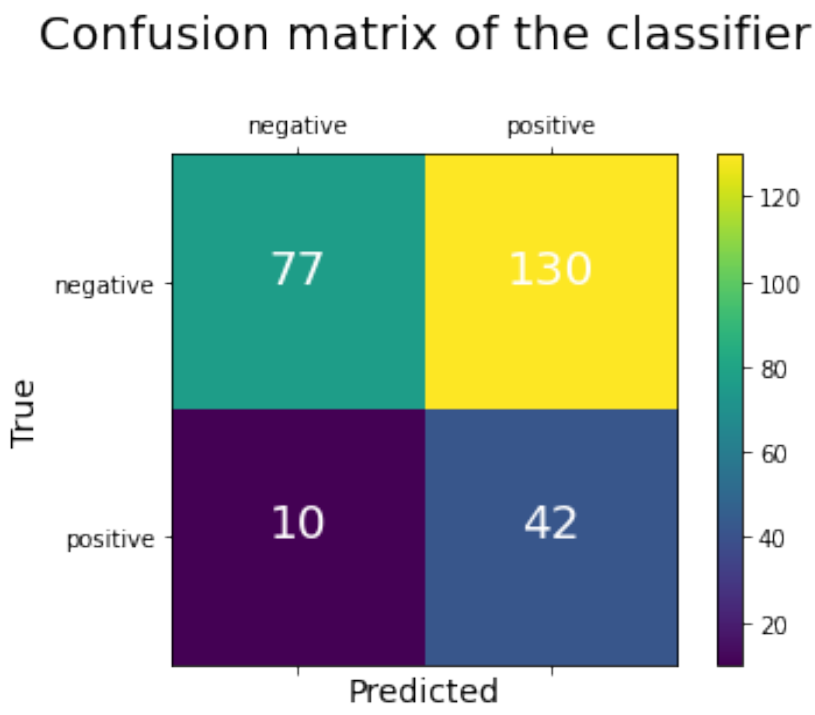
The final threshold is **0.4979**. This is the point where the F1 score graph reaches its maximum value.

Person / Algorithm	F1	95% CI min	95% CI max
Radiologist 1	0.383	0.309	0.453
Radiologist 2	0.356	0.282	0.428
Radiologist 3	0.365	0.291	0.435
Radiologist 4	0.442	0.390	0.492
Radiologist average	0.387	0.330	0.442
CheXNet	0.435	0.387	0.481
ChainRad	**0.377**		

The average F1 score of human radiologists is 0.387 according to paper of CheXNet, that is available here: https://arxiv.org/pdf/1711.05225.pdf

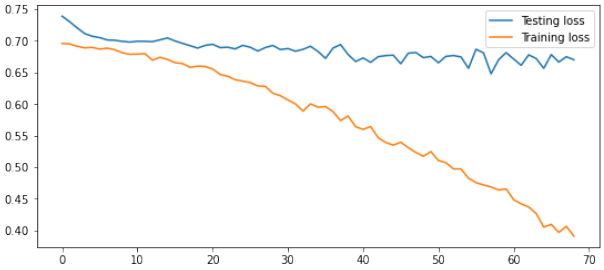
The main goal is for the model to reach the average F1 score of human radiologists. There are the major measures related to the final threshold.

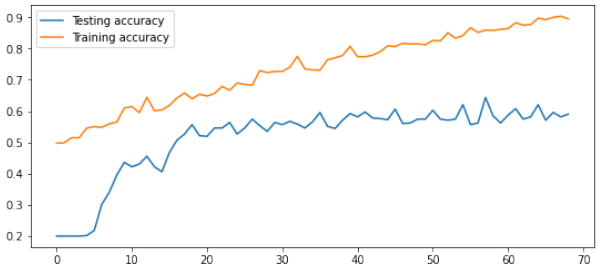
- F1 score max: **0.3767**
- Threshold: **0.4979**
- Precision: **0.2442**
- Recall: **0.8077**



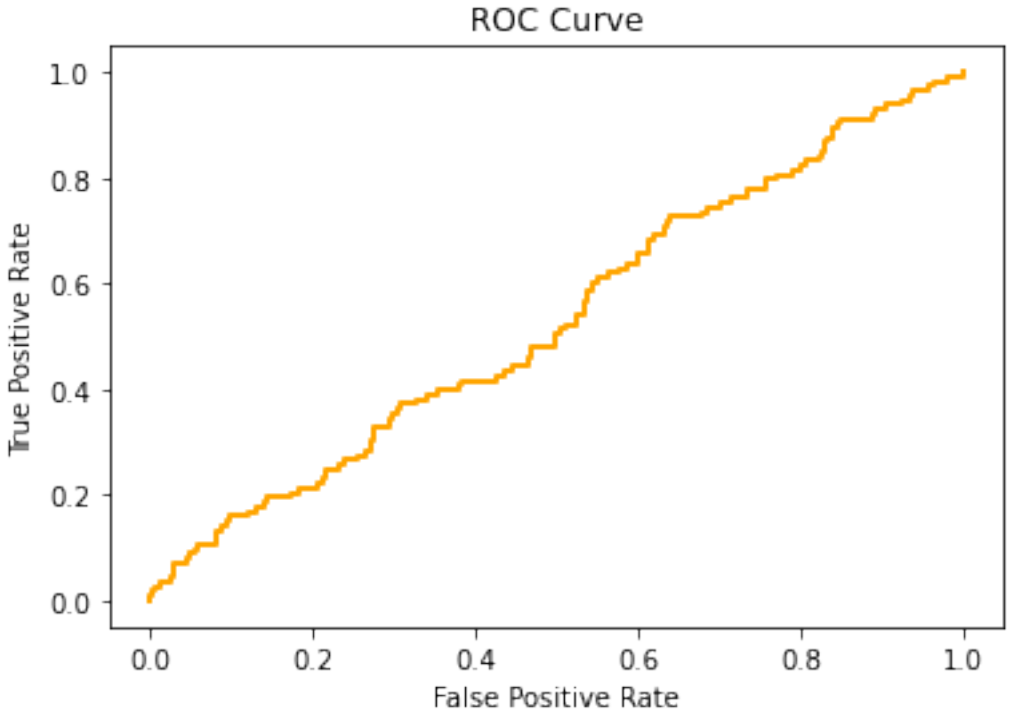
Fibrosis

Algorithm training performance visualization

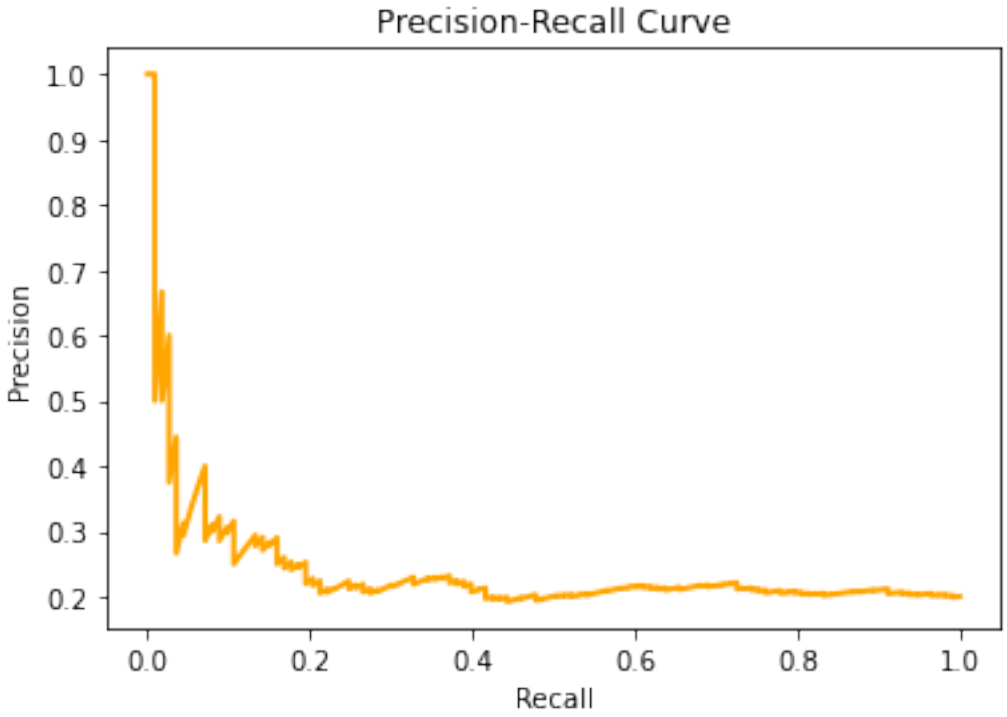


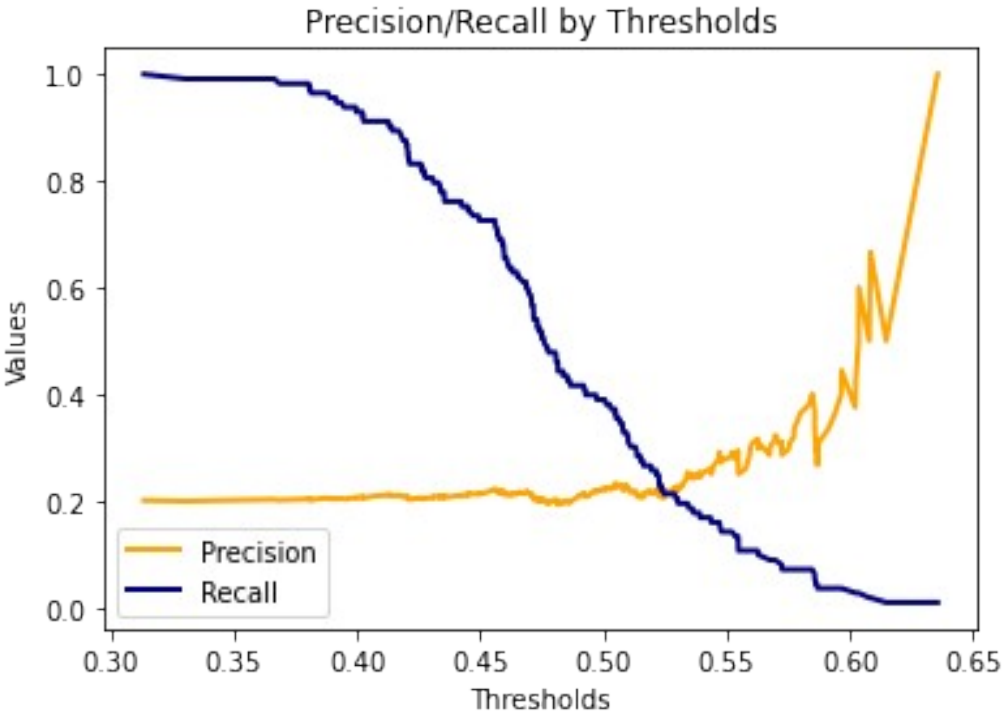


ROC curve shows the model learned something from training dataset. It will be better with smaller learning rate and longer training. This is the part of a future experiment.

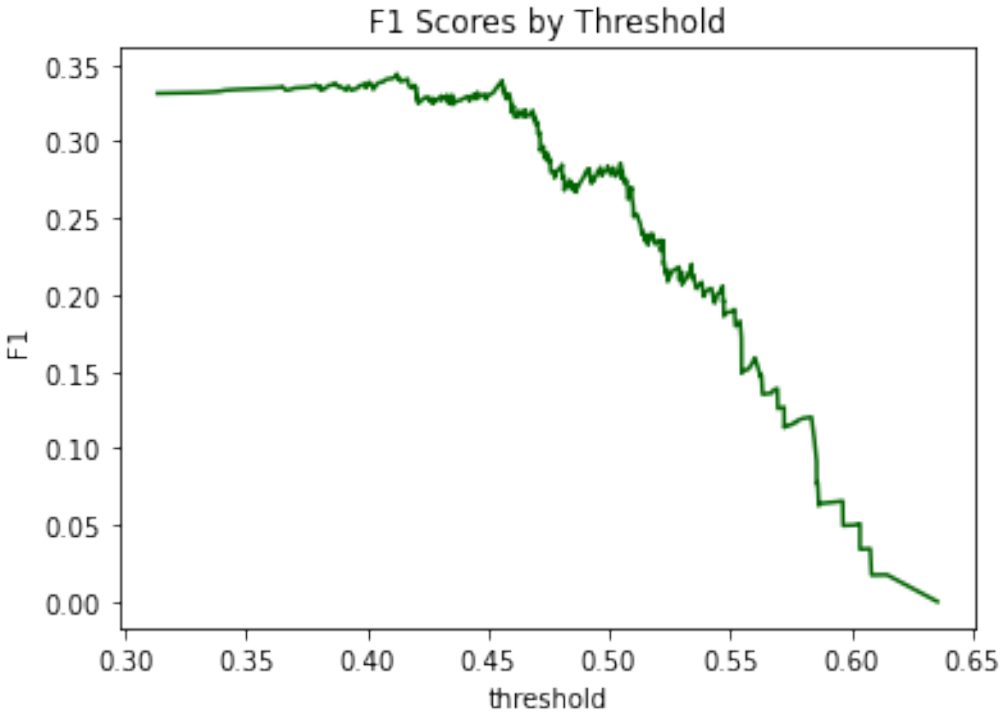


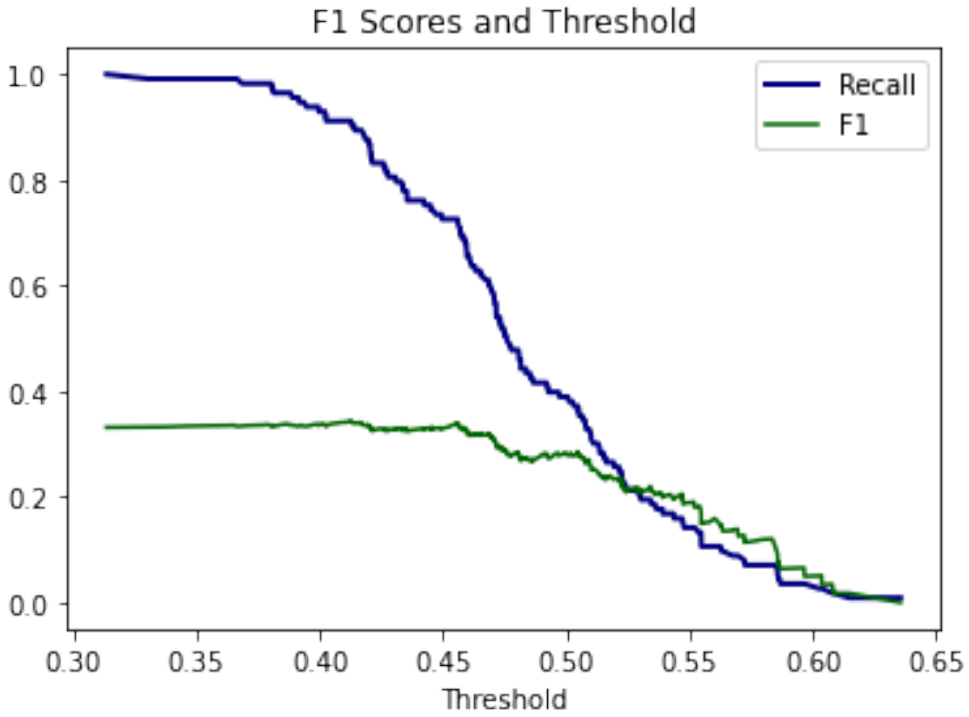
P-R curve


Precision-Recall Curve
Precision
Recall
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.0
0.2
0.4
0.6
0.8
1.0


Precision/Recall by Thresholds
Values
Thresholds
Precision
Recall
1.0
0.8
0.6
0.4
0.2
0.0
0.30
0.35
0.40
0.45
0.50
0.55
0.60
0.65

Final Threshold and Explanation:





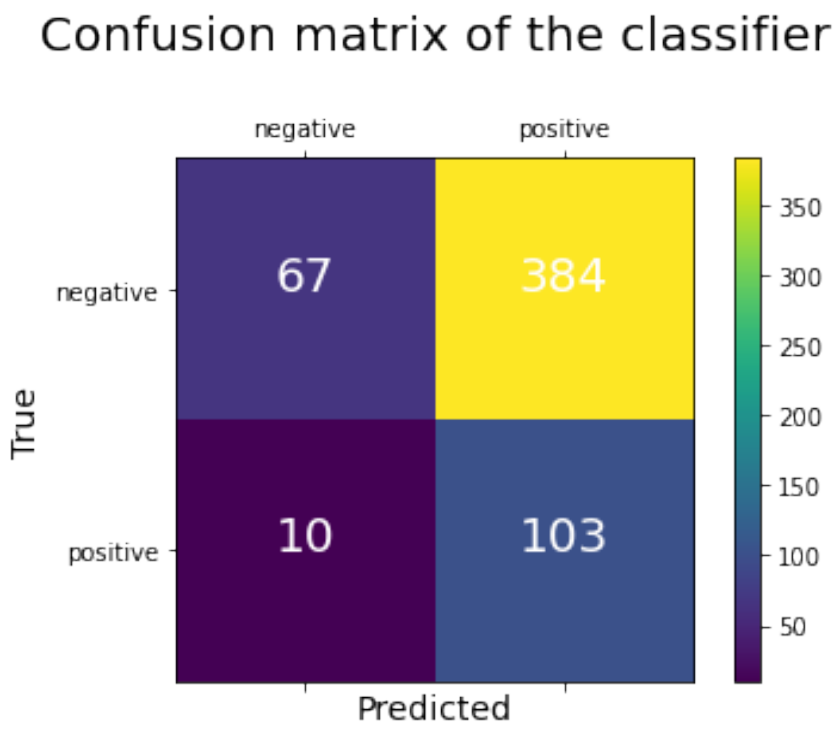
The final threshold is **0.4125**. This is the point where the F1 score graph reaches its maximum value.

Person / Algorithm	F1	95% CI min	95% CI max
Radiologist 1	0.383	0.309	0.453
Radiologist 2	0.356	0.282	0.428
Radiologist 3	0.365	0.291	0.435
Radiologist 4	0.442	0.390	0.492
Radiologist average	0.387	0.330	0.442
CheXNet	0.435	0.387	0.481
ChainRad	**0.344**		

The average F1 score of human radiologists is 0.387 according to paper of CheXNet, that is available here: https://arxiv.org/pdf/1711.05225.pdf

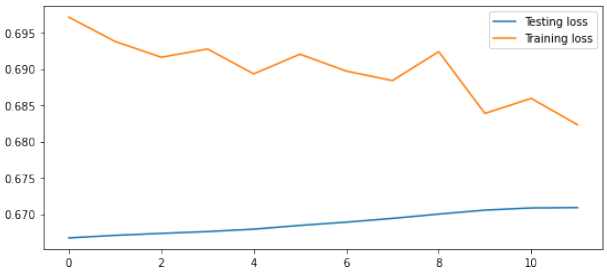
The main goal is for the model to reach the average F1 score of human radiologists. There are the major measures related to the final threshold.

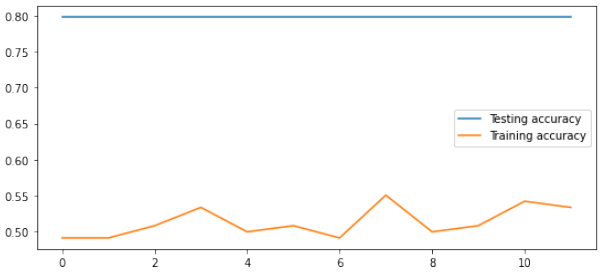
- F1 score max: **0.3439**
- Threshold: **0.4125**
- Precision: **0.2115**
- Recall: **0.9115**



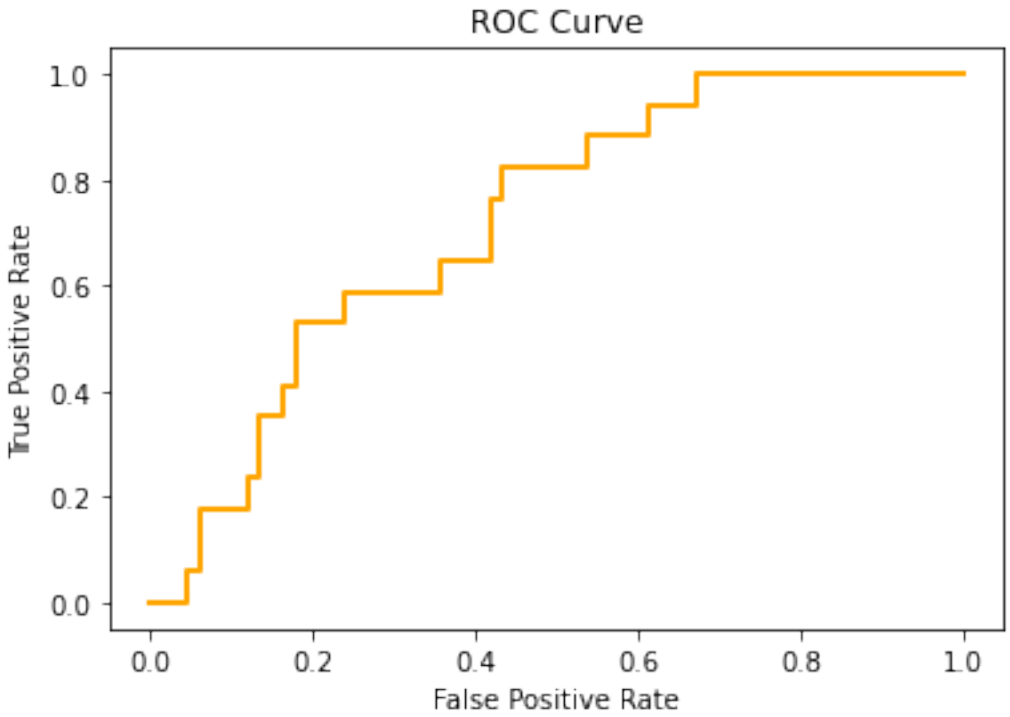
Hernia

Algorithm training performance visualization

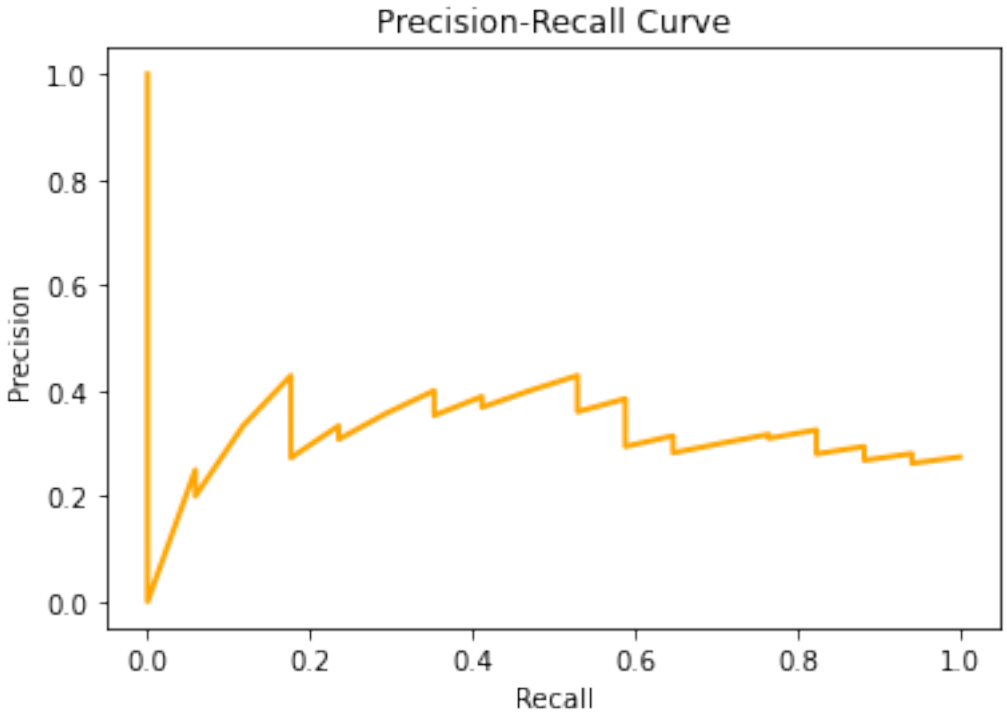


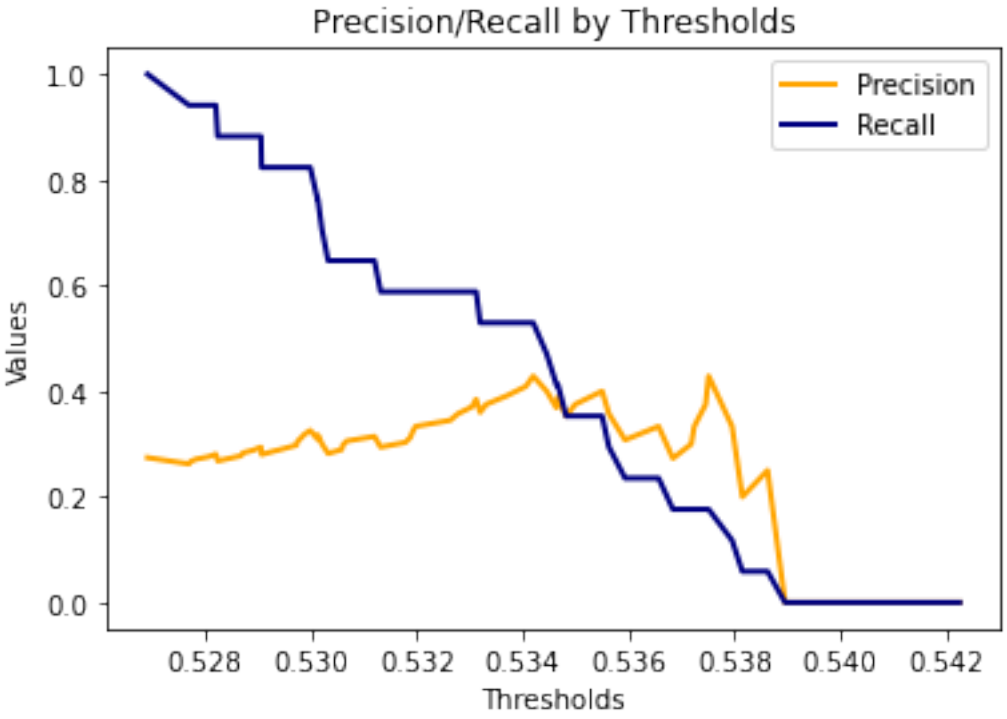


ROC curve shows the model learned something from training dataset. It will be better with smaller learning rate and longer training. This is the part of a future experiment.

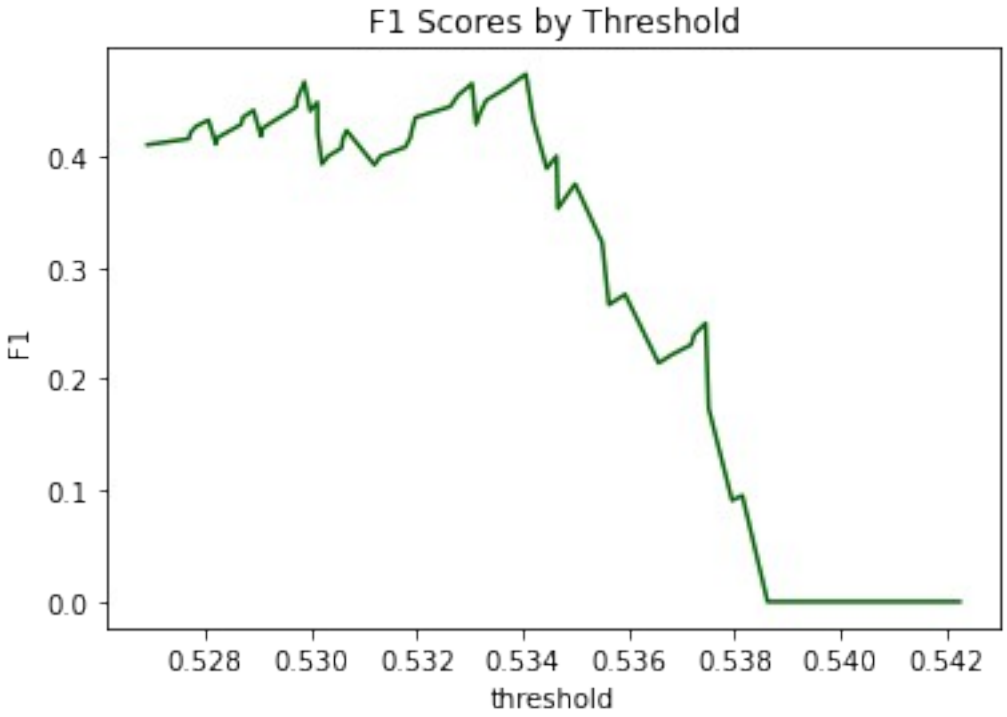


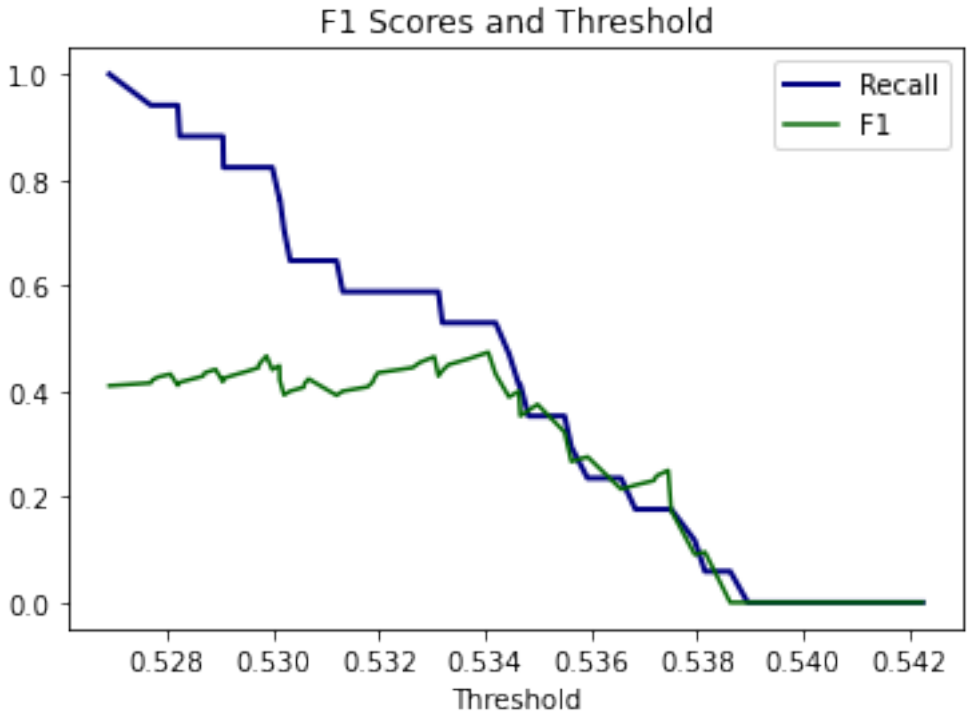
P-R curve


Precision-Recall Curve
Precision
Recall
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0


Precision/Recall by Thresholds
Precision
Recall
Values
Thresholds
1.0
0.8
0.6
0.4
0.2
0.0
0.528
0.530
0.532
0.534
0.536
0.538
0.540
0.542

Final Threshold and Explanation:





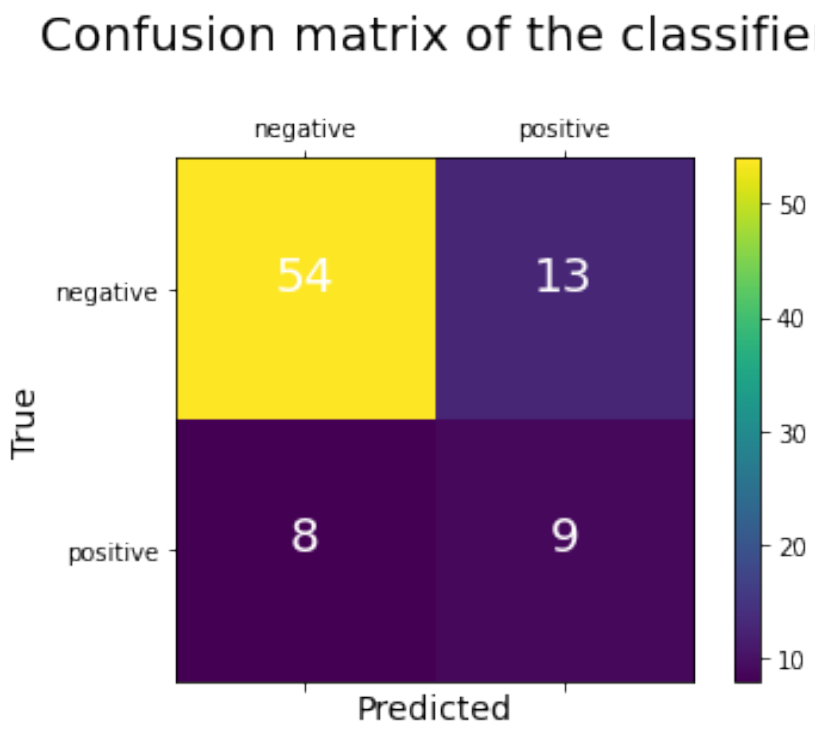
The final threshold is **0.5341**. This is the point where the F1 score graph reaches its maximum value.

Person / Algorithm	F1	95% CI min	95% CI max
Radiologist 1	0.383	0.309	0.453
Radiologist 2	0.356	0.282	0.428
Radiologist 3	0.365	0.291	0.435
Radiologist 4	0.442	0.390	0.492
Radiologist average	0.387	0.330	0.442
CheXNet	0.435	0.387	0.481
ChainRad	**0.474**		

The average F1 score of human radiologists is 0.387 according to paper of CheXNet, that is available here: https://arxiv.org/pdf/1711.05225.pdf

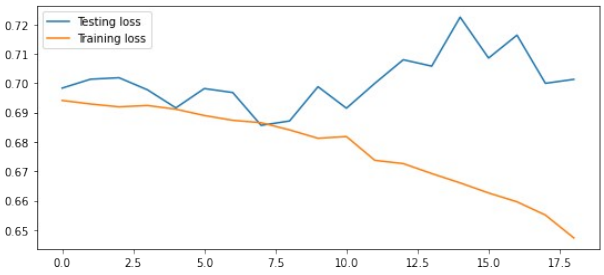
The main goal is for the model to reach the average F1 score of human radiologists. There are the major measures related to the final threshold.

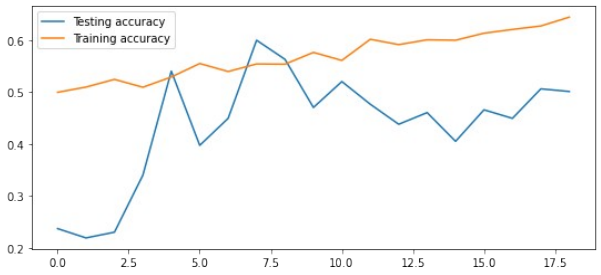
- F1 score max: **0.4737**
- Threshold: **0.5341**
- Precision: **0.4091**
- Recall: **0.5294**



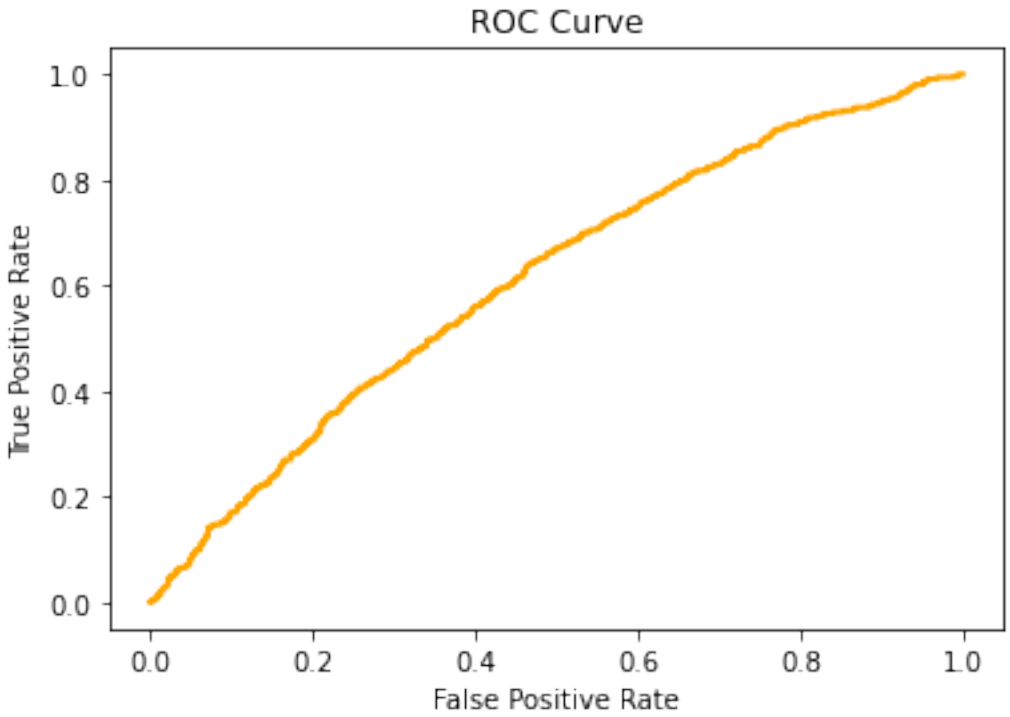
Infiltration

Algorithm training performance visualization

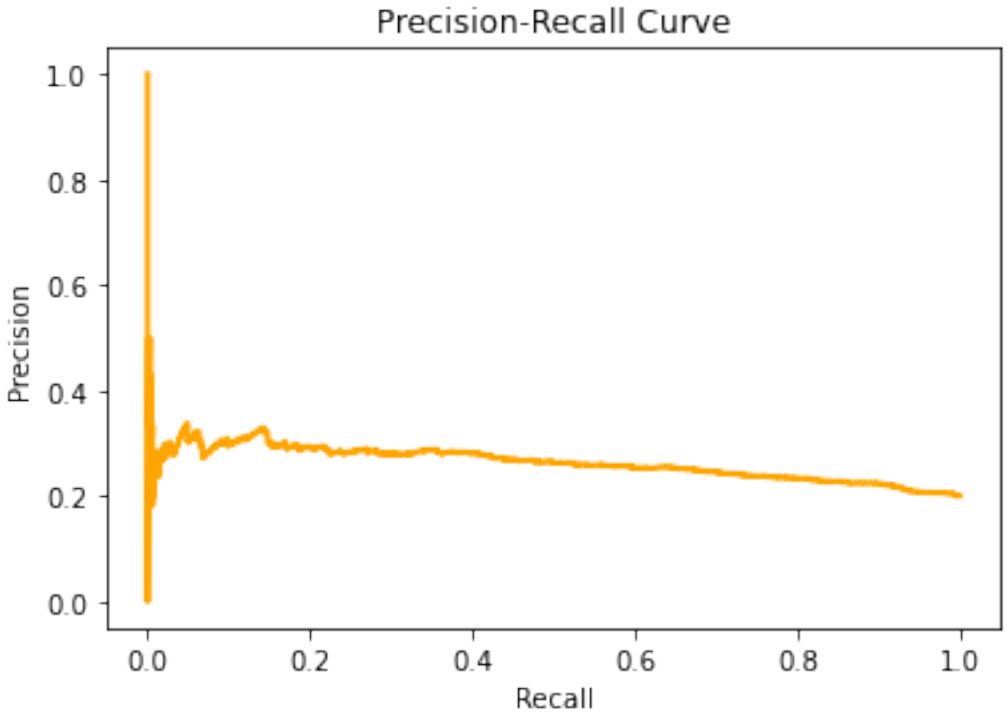


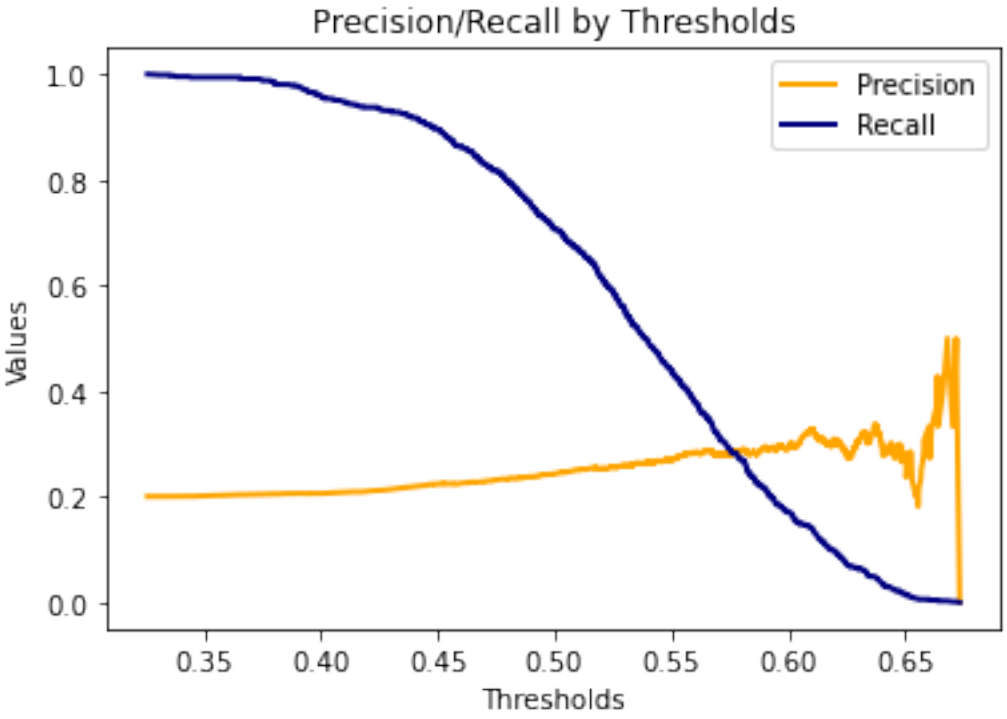


ROC curve shows the model learned something from training dataset. It will be better with smaller learning rate and longer training. This is the part of a future experiment.

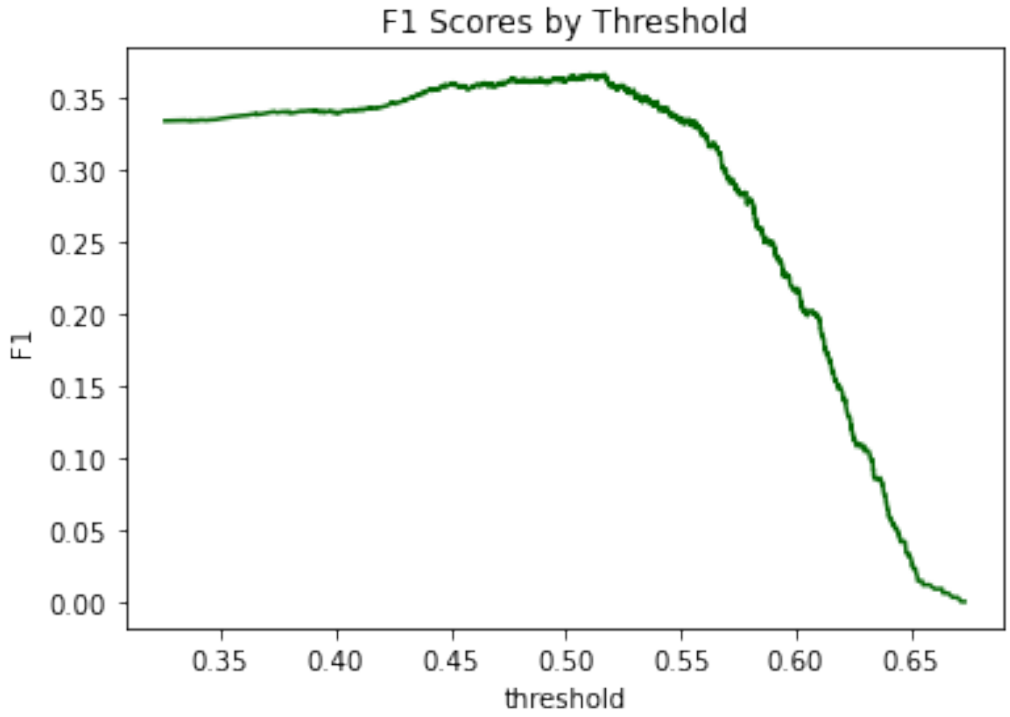


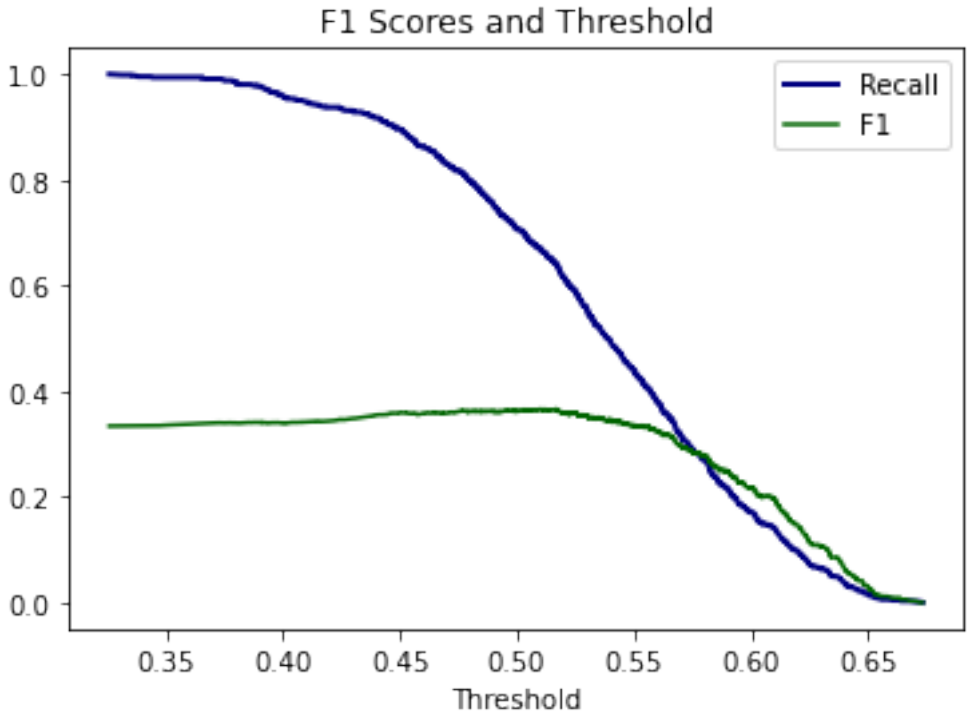
P-R curve


Precision-Recall Curve
Precision
Recall
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0


Precision/Recall by Thresholds
Precision
Recall
Values
Thresholds
1.0
0.8
0.6
0.4
0.2
0.0
0.35
0.40
0.45
0.50
0.55
0.60
0.65

Final Threshold and Explanation:





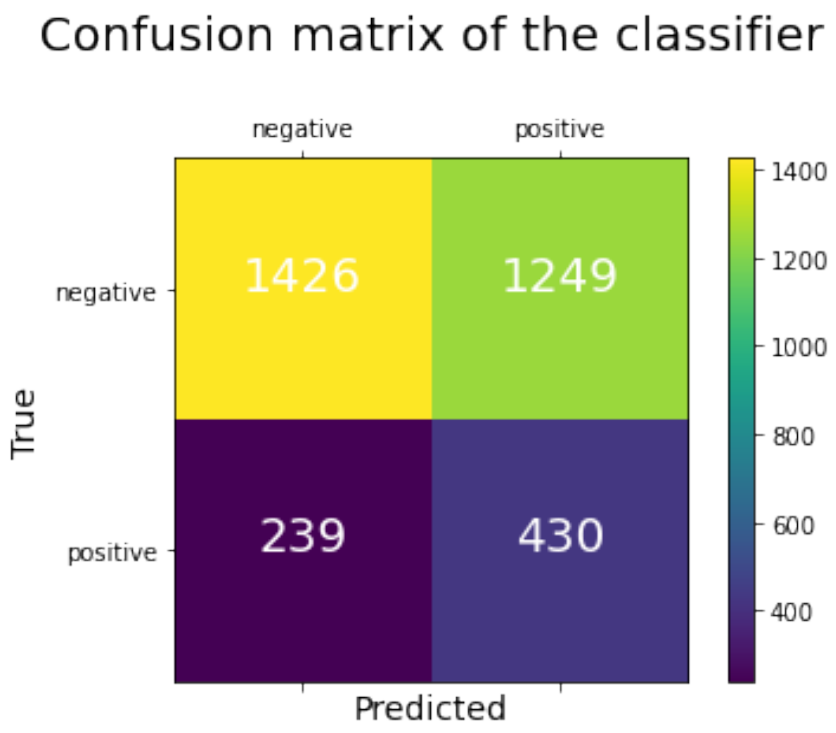
The final threshold is **0.5168**. This is the point where the F1 score graph reaches its maximum value.

Person / Algorithm	F1	95% CI min	95% CI max
Radiologist 1	0.383	0.309	0.453
Radiologist 2	0.356	0.282	0.428
Radiologist 3	0.365	0.291	0.435
Radiologist 4	0.442	0.390	0.492
Radiologist average	0.387	0.330	0.442
CheXNet	0.435	0.387	0.481
ChainRad	**0.366**		

The average F1 score of human radiologists is 0.387 according to paper of CheXNet, that is available here: https://arxiv.org/pdf/1711.05225.pdf

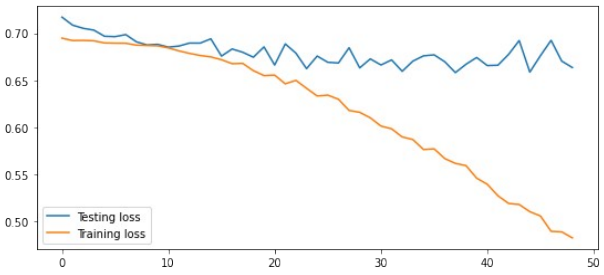
The main goal is for the model to reach the average F1 score of human radiologists. There are the major measures related to the final threshold.

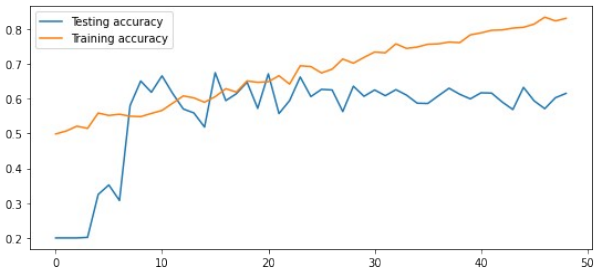
- F1 score max: **0.3664**
- Threshold: **0.5168**
- Precision: **0.2561**
- Recall: **0.6428**



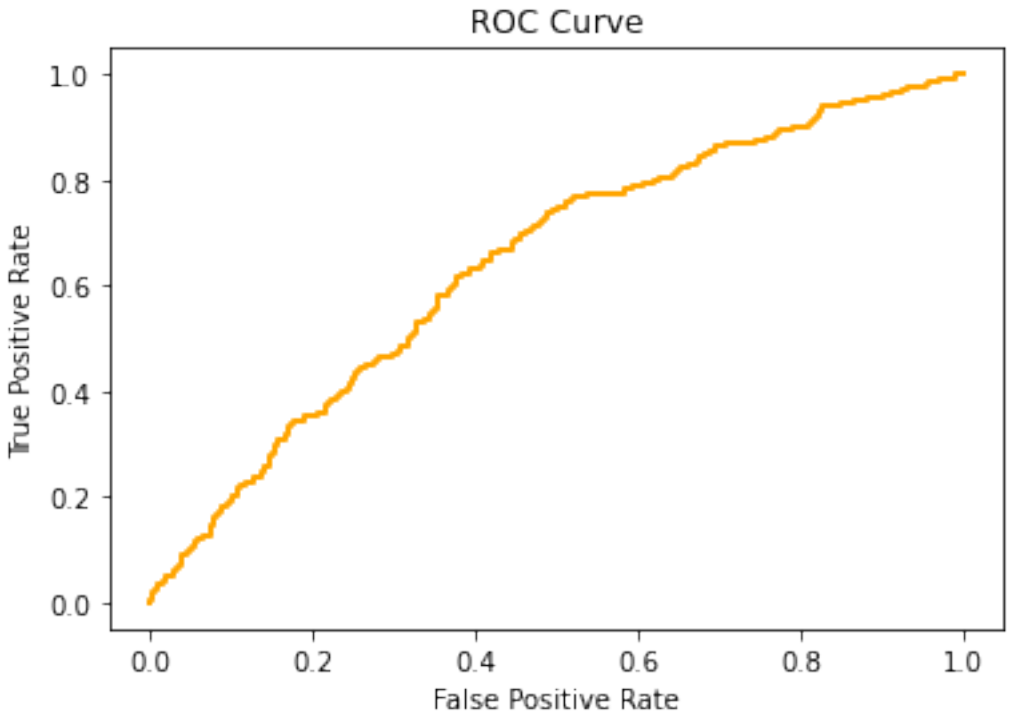
Mass

Algorithm training performance visualization

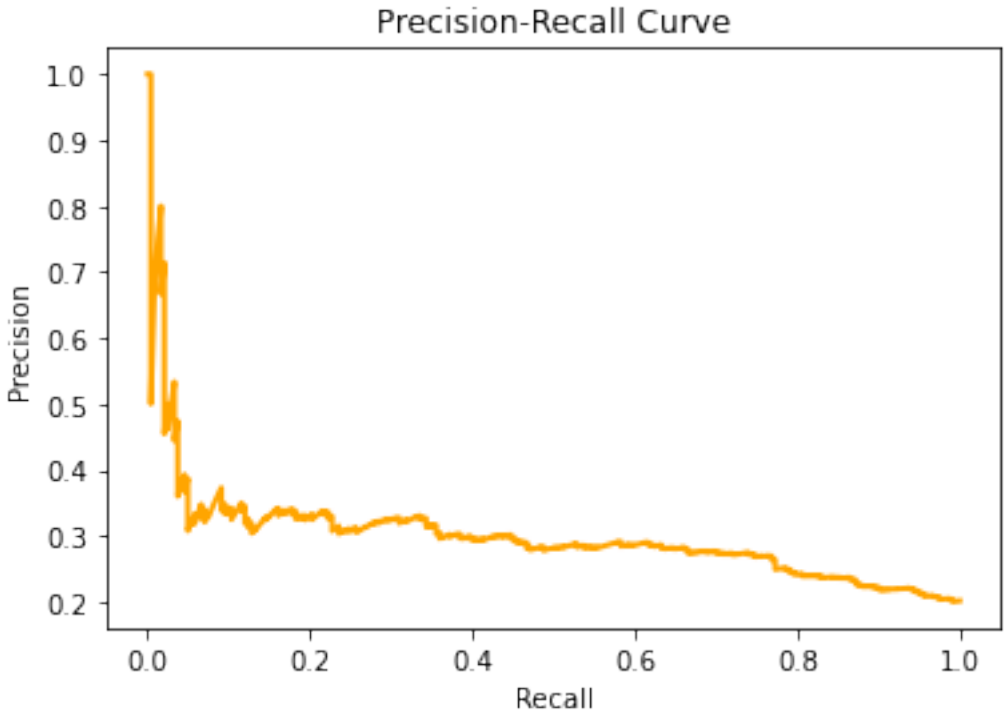


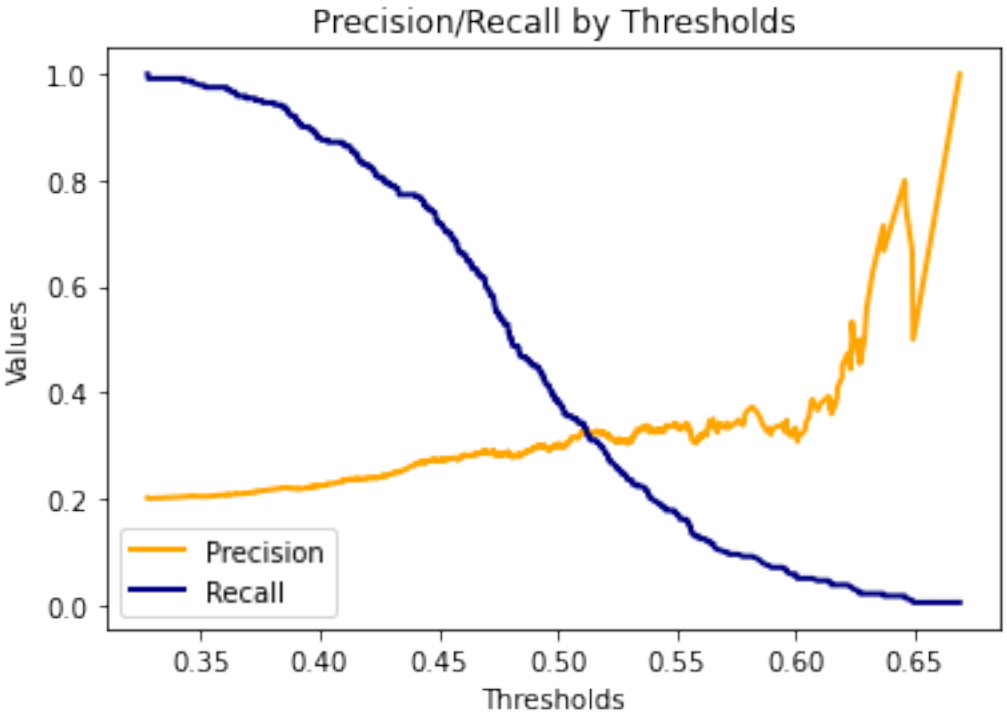


ROC curve shows the model learned something from training dataset. It will be better with smaller learning rate and longer training. This is the part of a future experiment.

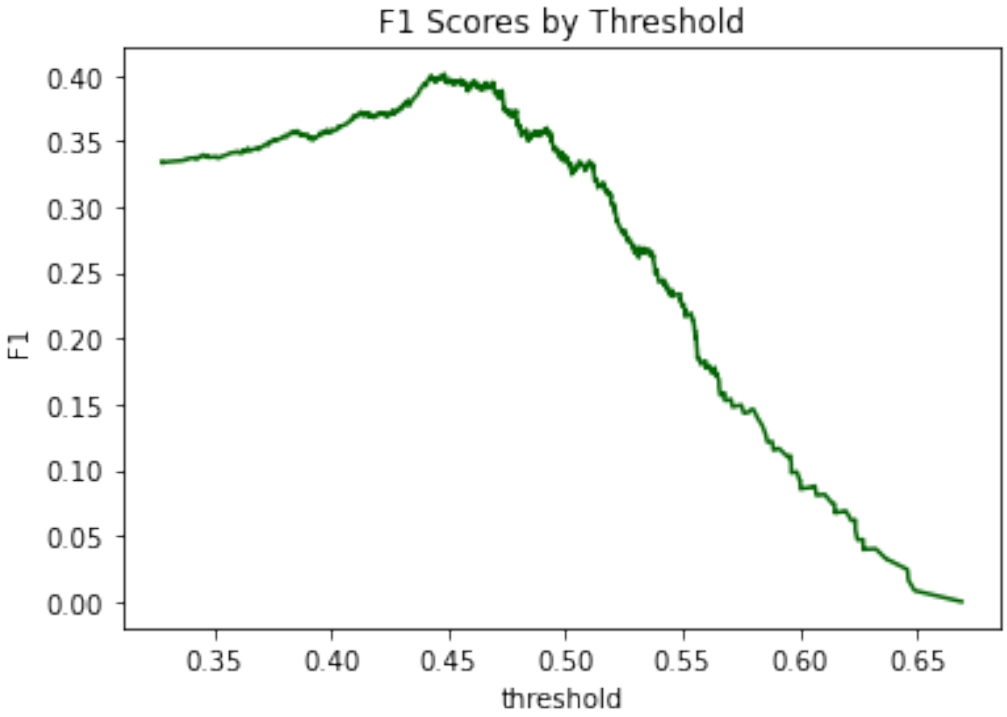


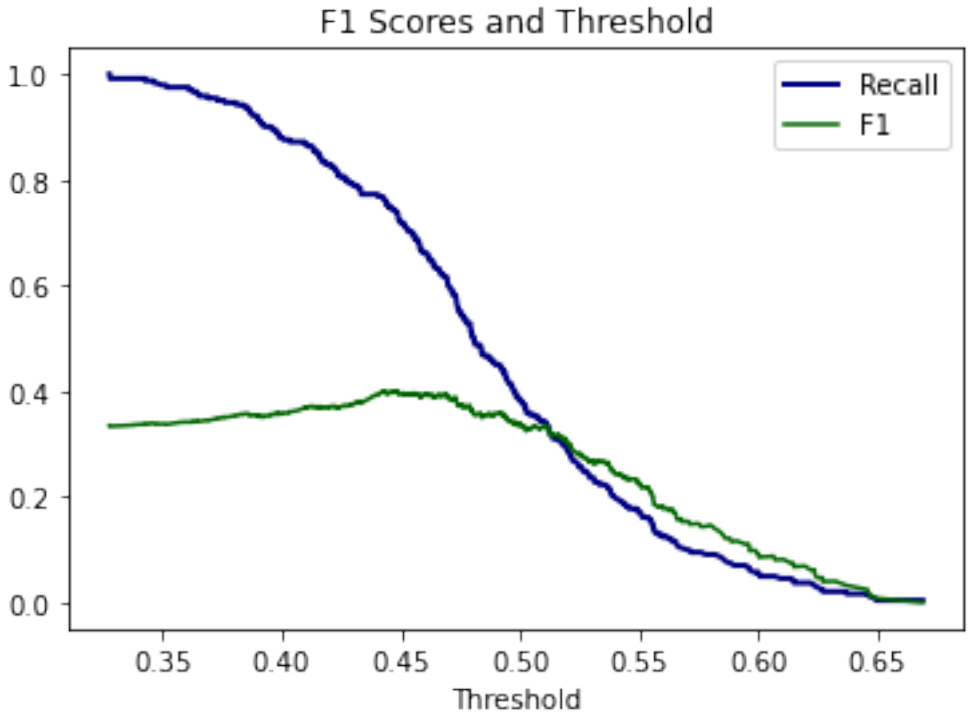
P-R curve


Precision-Recall Curve
Precision
Recall
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.0
0.2
0.4
0.6
0.8
1.0


Precision/Recall by Thresholds
Values
Thresholds
Precision
Recall
1.0
0.8
0.6
0.4
0.2
0.0
0.35
0.40
0.45
0.50
0.55
0.60
0.65

Final Threshold and Explanation:





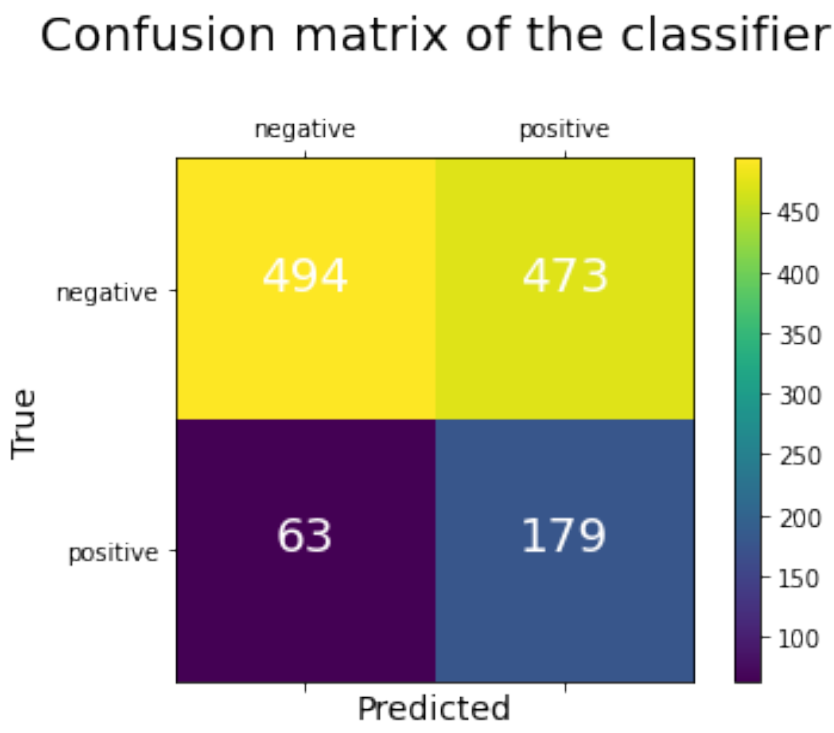
The final threshold is **0.4484**. This is the point where the F1 score graph reaches its maximum value.

Person / Algorithm	F1	95% CI min	95% CI max
Radiologist 1	0.383	0.309	0.453
Radiologist 2	0.356	0.282	0.428
Radiologist 3	0.365	0.291	0.435
Radiologist 4	0.442	0.390	0.492
Radiologist average	0.387	0.330	0.442
CheXNet	0.435	0.387	0.481
ChainRad	**0.401**		

The average F1 score of human radiologists is 0.387 according to paper of CheXNet, that is available here: https://arxiv.org/pdf/1711.05225.pdf

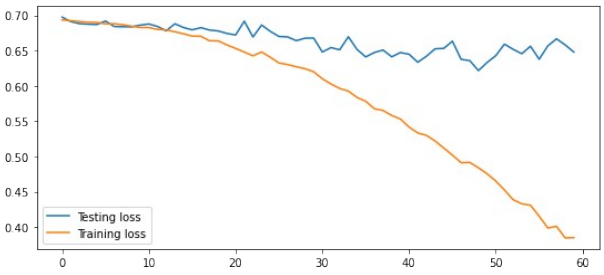
The main goal is for the model to reach the average F1 score of human radiologists. There are the major measures related to the final threshold.

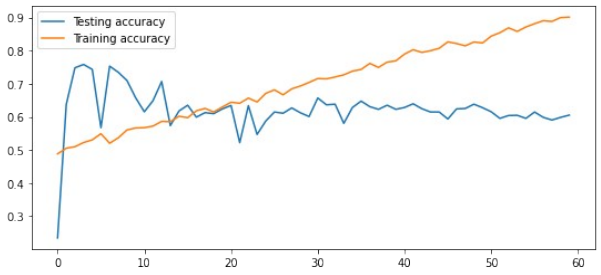
- F1 score max: **0.4009**
- Threshold: **0.4484**
- Precision: **0.2745**
- Recall: **0.7397**



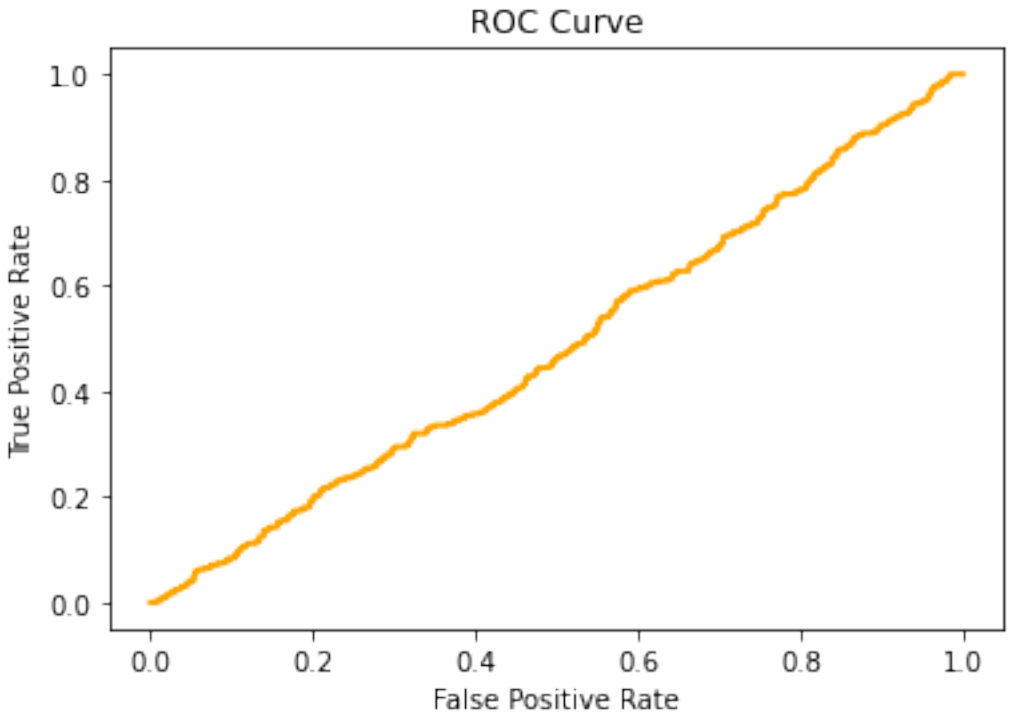
Nodule

Algorithm training performance visualization

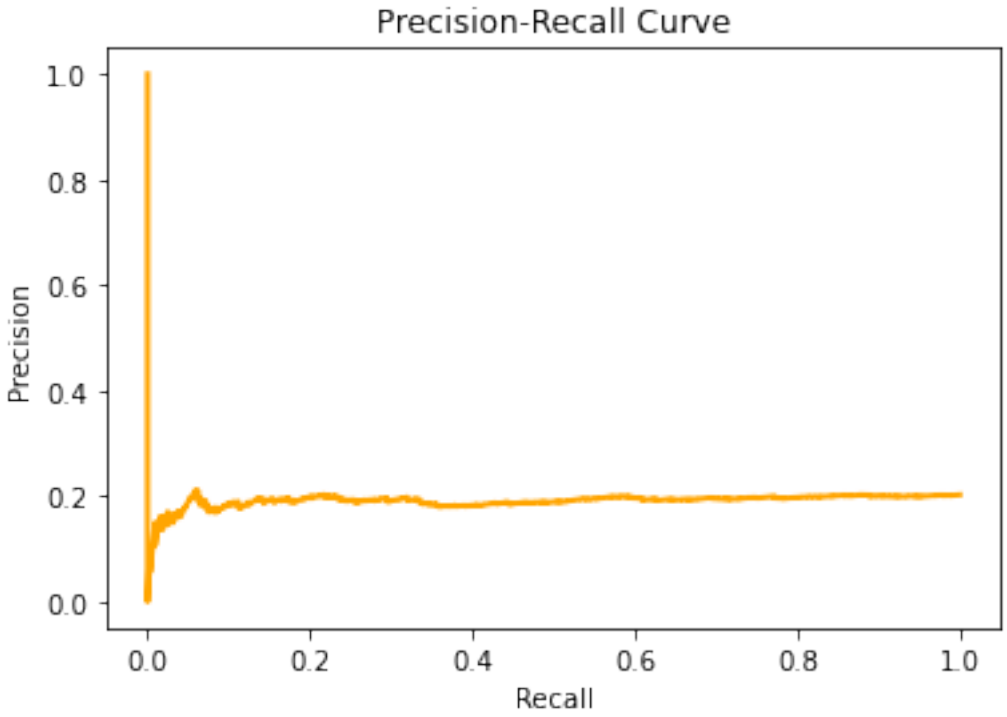


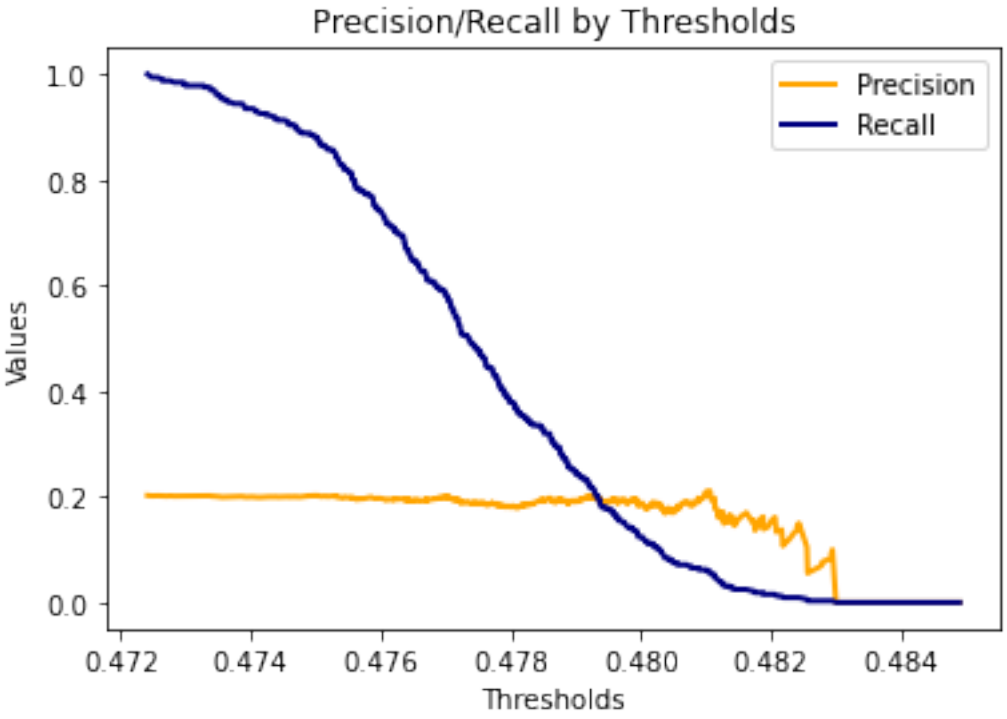


ROC curve shows the model learned nearly nothing from training dataset. It will be better with smaller learning rate and longer training. This is the part of a future experiment.

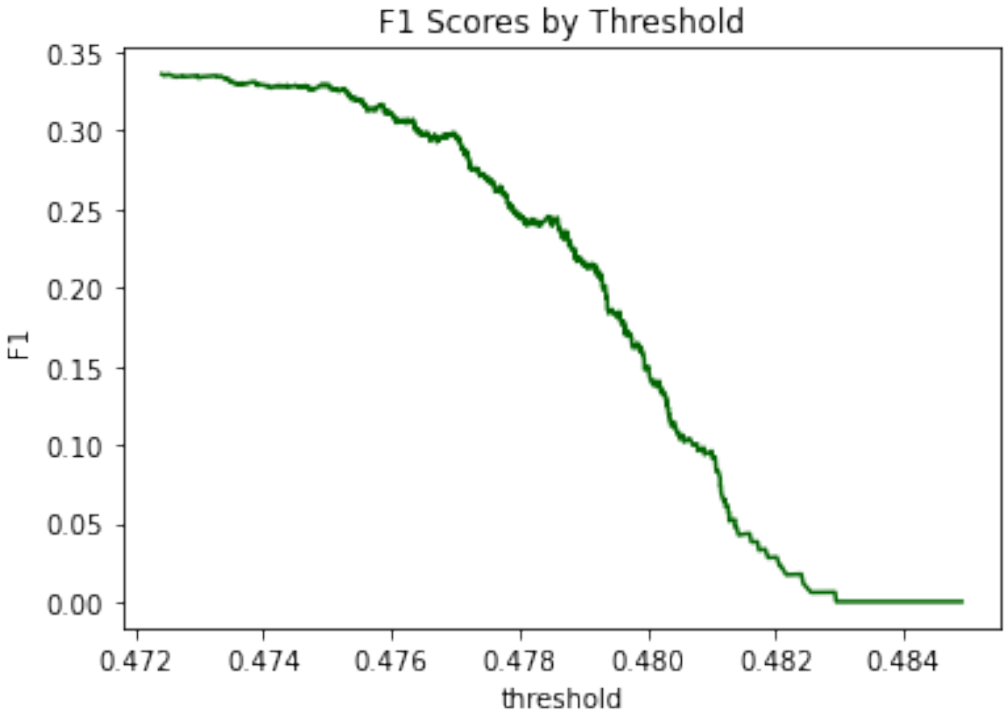


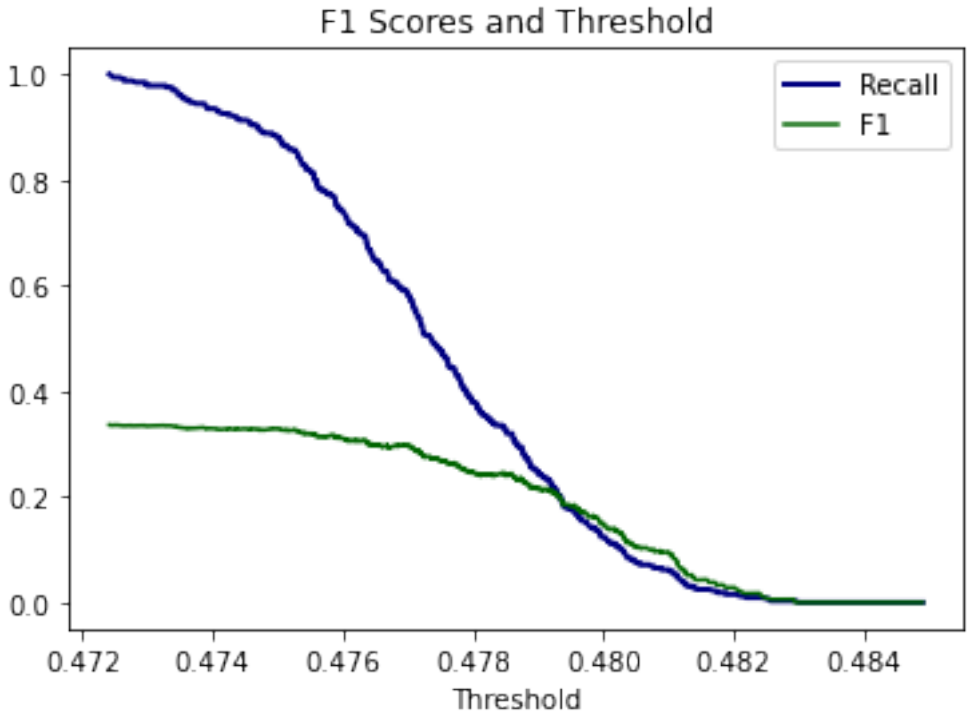
P-R curve


Precision-Recall Curve
Precision
Recall
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0


Precision/Recall by Thresholds
Precision
Recall
Values
Thresholds
1.0
0.8
0.6
0.4
0.2
0.0
0.472
0.474
0.476
0.478
0.480
0.482
0.484

Final Threshold and Explanation:





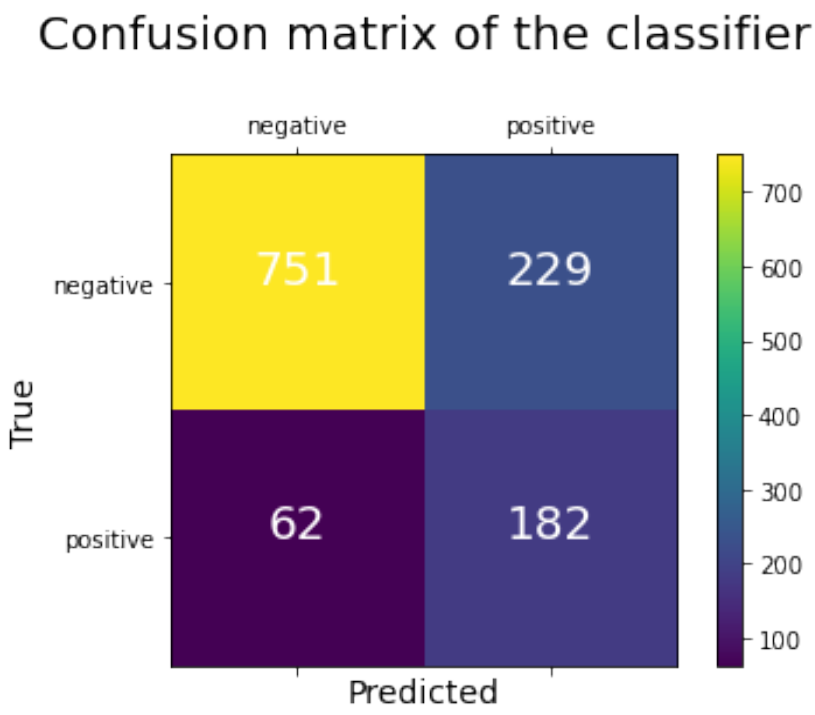
The final threshold is **0.4724**. This is the point where the F1 score graph reaches its maximum value.

Person / Algorithm	F1	95% CI min	95% CI max
Radiologist 1	0.383	0.309	0.453
Radiologist 2	0.356	0.282	0.428
Radiologist 3	0.365	0.291	0.435
Radiologist 4	0.442	0.390	0.492
Radiologist average	0.387	0.330	0.442
CheXNet	0.435	0.387	0.481
ChainRad	**0.336**		

The average F1 score of human radiologists is 0.387 according to paper of CheXNet, that is available here: https://arxiv.org/pdf/1711.05225.pdf

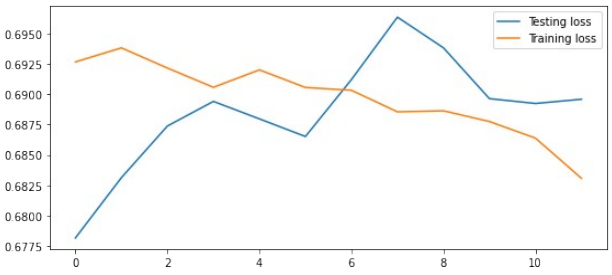
The main goal is for the model to reach the average F1 score of human radiologists. There are the major measures related to the final threshold.

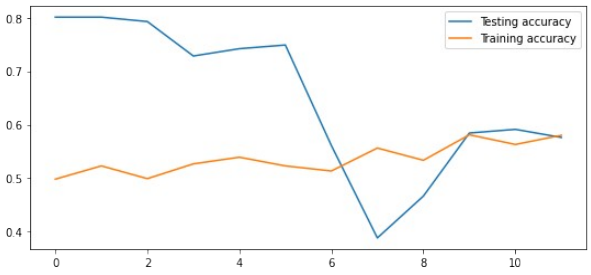
- F1 score max: **0.3361**
- Threshold: **0.4724**
- Precision: **0.2026**
- Recall: **1.0000**



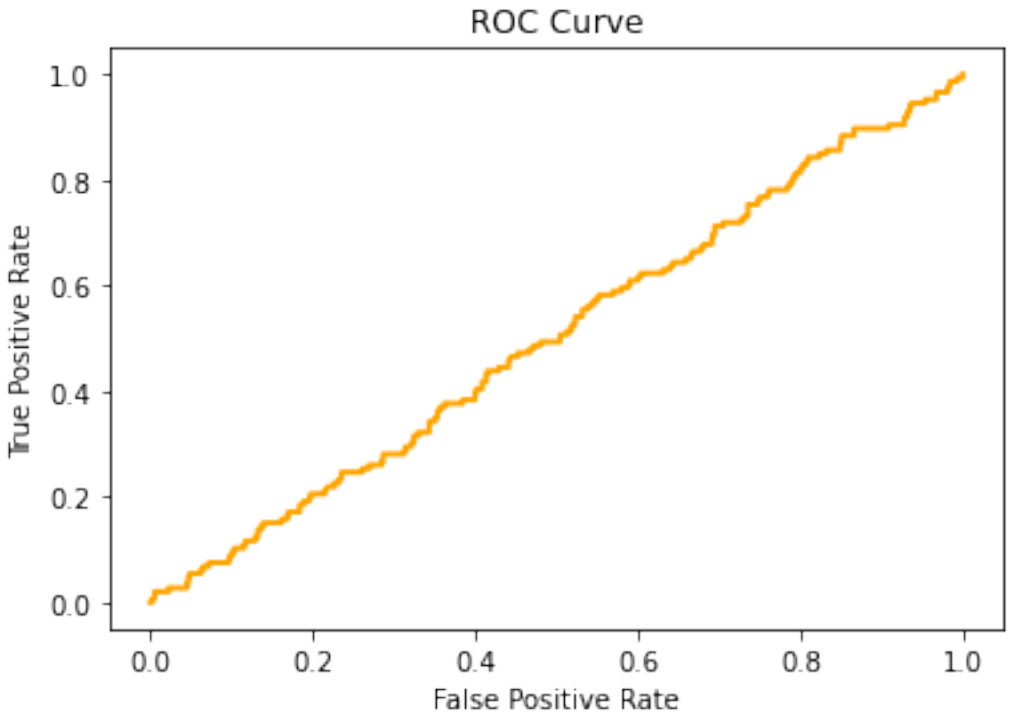
Pleural Thickening

Algorithm training performance visualization

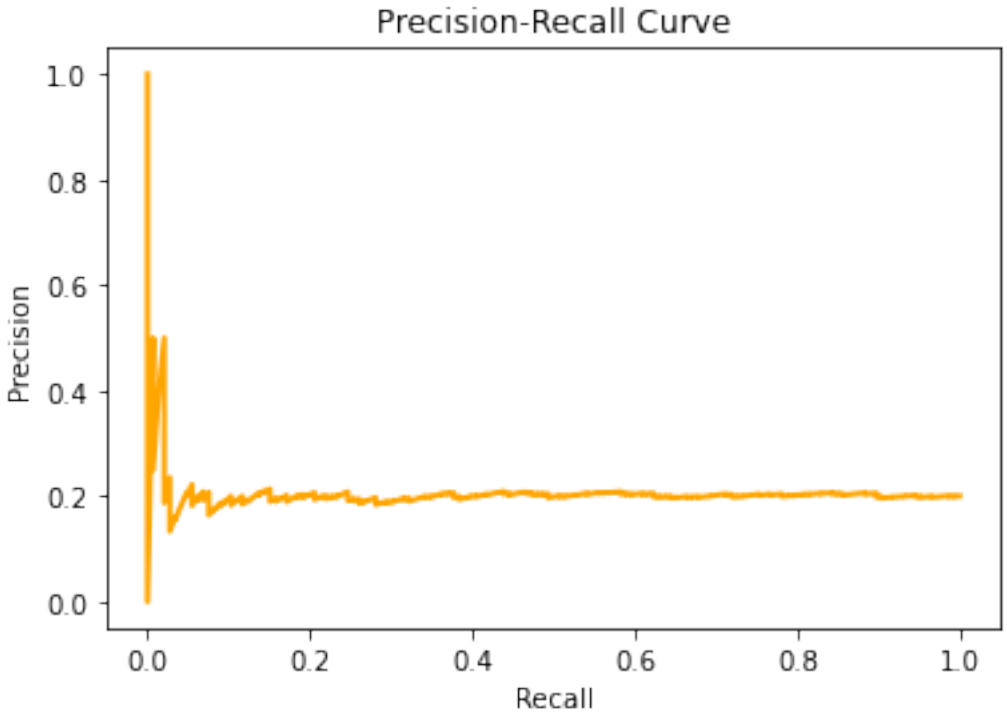


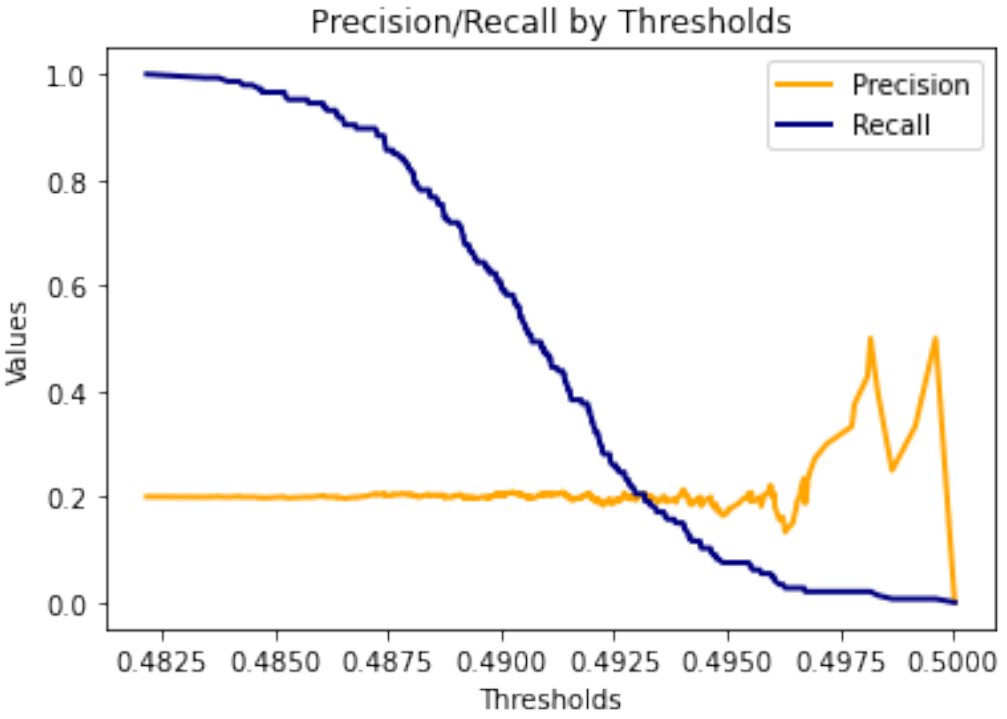


ROC curve shows the model learned nearly nothing from training dataset. It will be better with smaller learning rate and longer training. This is the part of a future experiment.

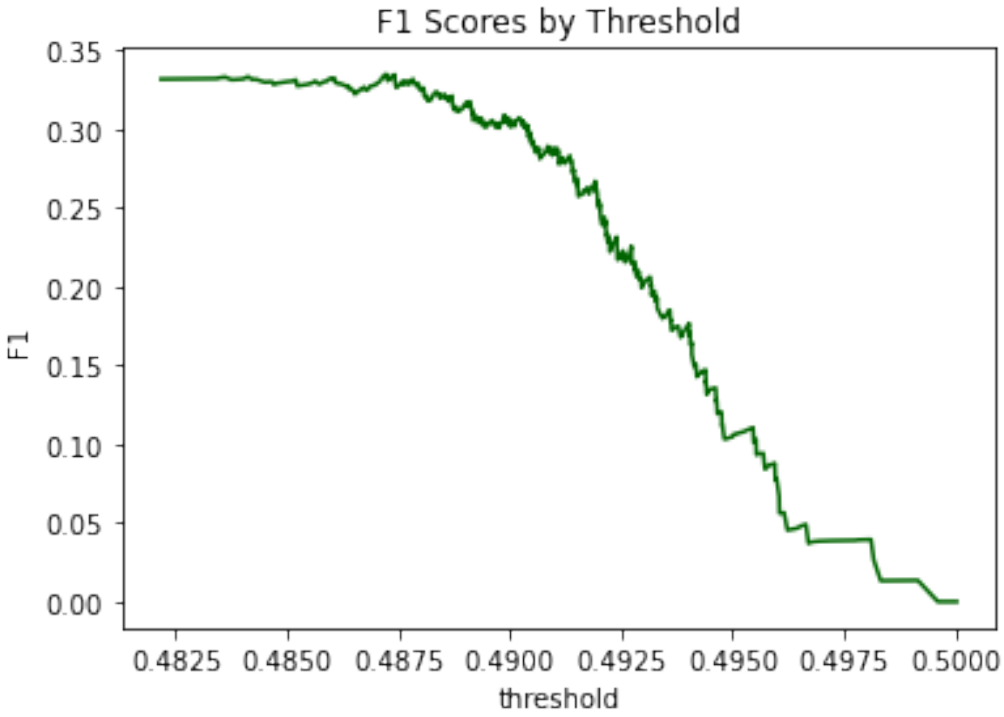


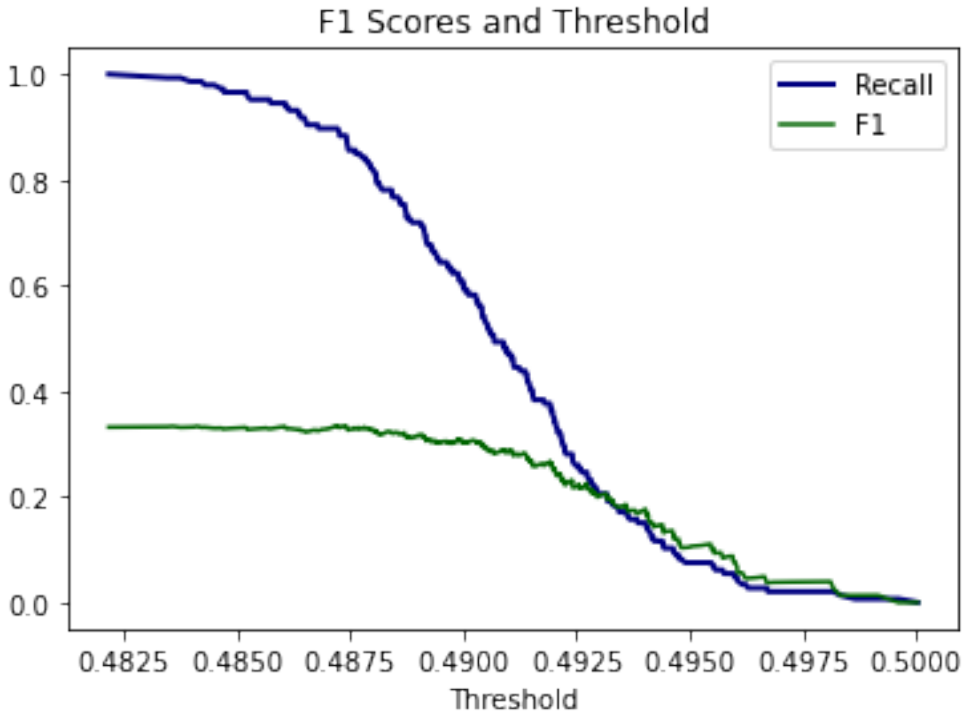
P-R curve


Precision-Recall Curve
Precision
Recall
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0


Precision/Recall by Thresholds
Precision
Recall
Values
Thresholds
1.0
0.8
0.6
0.4
0.2
0.0
0.4825
0.4850
0.4875
0.4900
0.4925
0.4950
0.4975
0.5000

Final Threshold and Explanation:





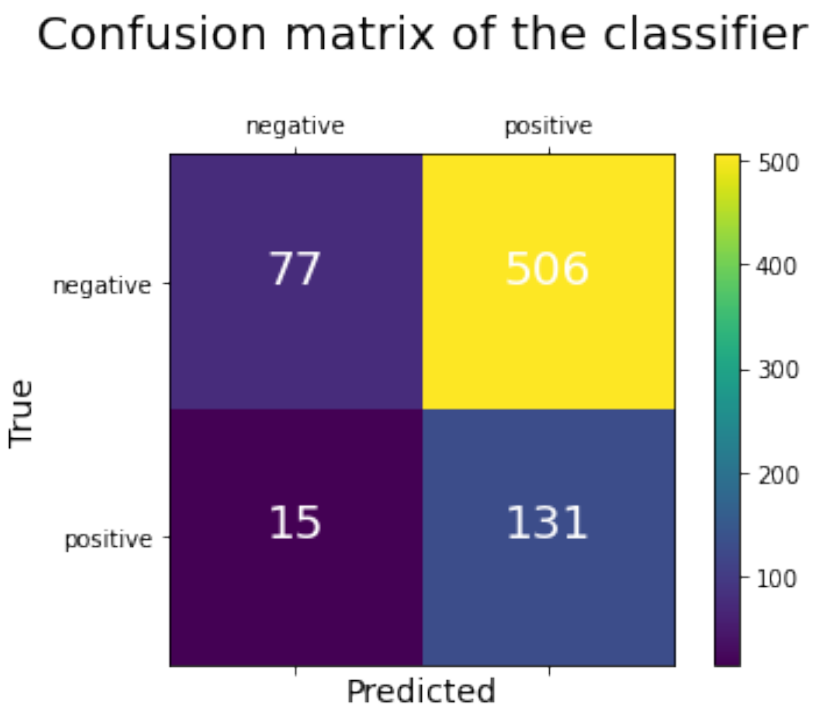
The final threshold is **0.4872**. This is the point where the F1 score graph reaches its maximum value.

Person / Algorithm	F1	95% CI min	95% CI max
Radiologist 1	0.383	0.309	0.453
Radiologist 2	0.356	0.282	0.428
Radiologist 3	0.365	0.291	0.435
Radiologist 4	0.442	0.390	0.492
Radiologist average	0.387	0.330	0.442
CheXNet	0.435	0.387	0.481
ChainRad	**0.335**		

The average F1 score of human radiologists is 0.387 according to paper of CheXNet, that is available here: https://arxiv.org/pdf/1711.05225.pdf

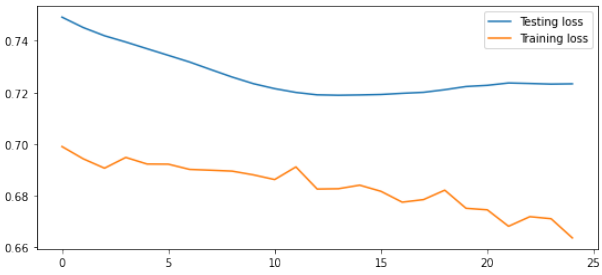
The main goal is for the model to reach the average F1 score of human radiologists. There are the major measures related to the final threshold.

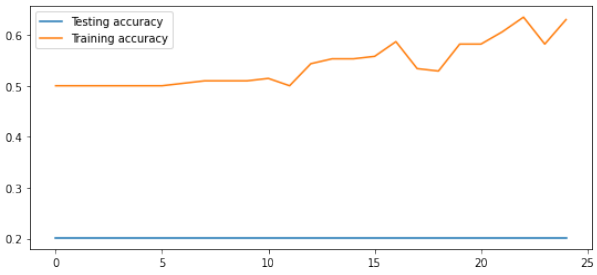
- F1 score max: **0.3350**
- Threshold: **0.4872**
- Precision: **0.2057**
- Recall: **0.8973**



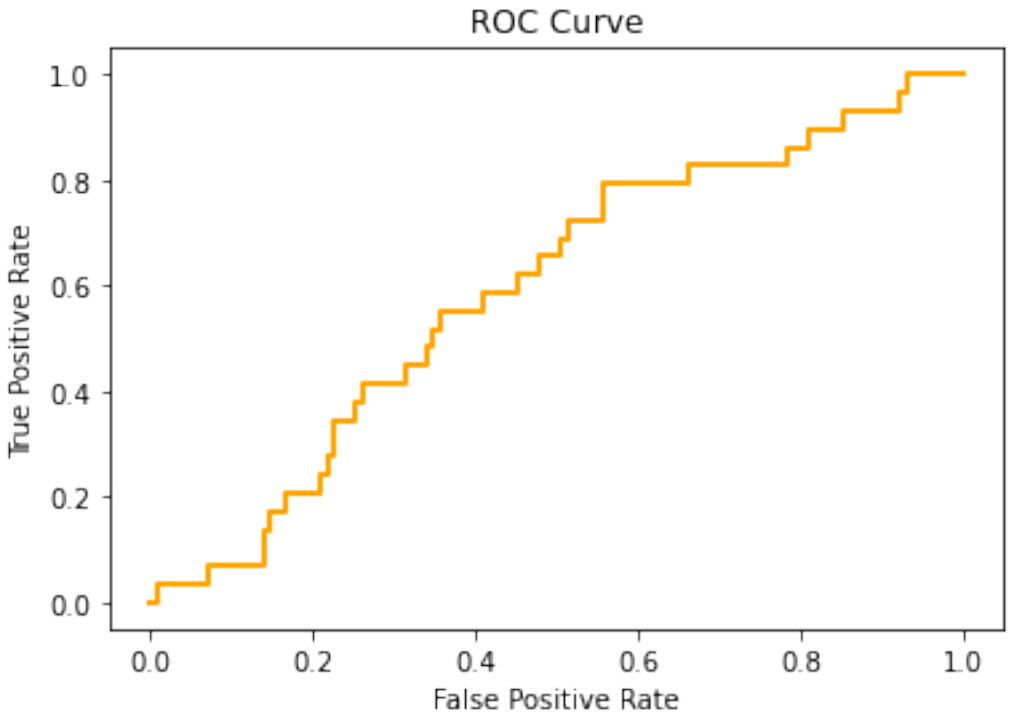
Pneumonia

Algorithm training performance visualization

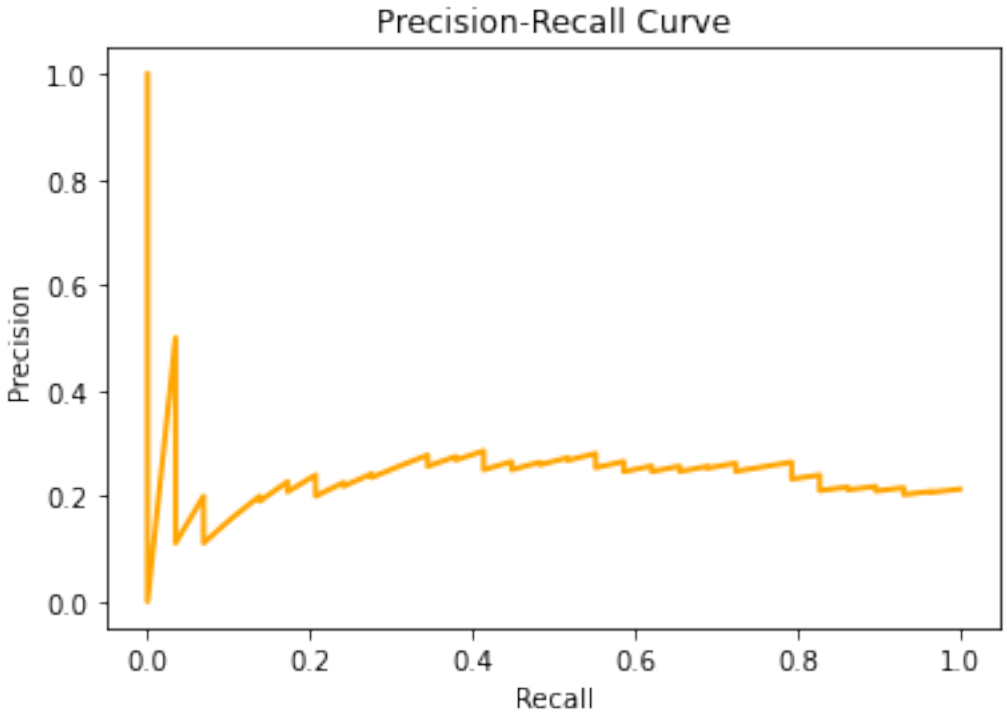


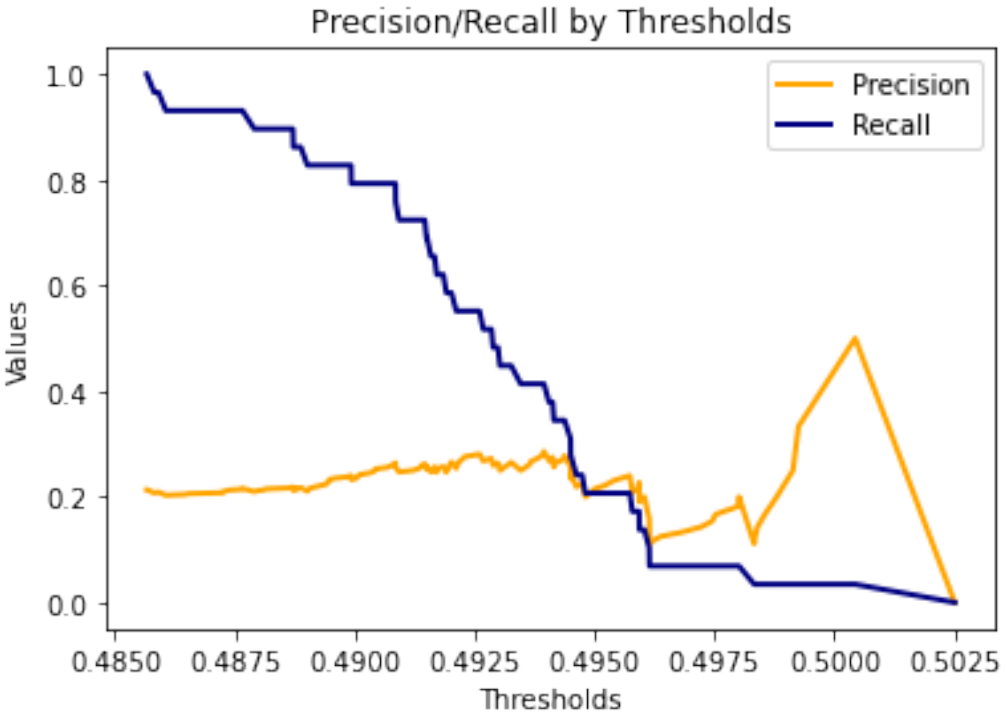


ROC curve shows the model learned something from training dataset. It will be better with smaller learning rate and longer training. This is the part of a future experiment.

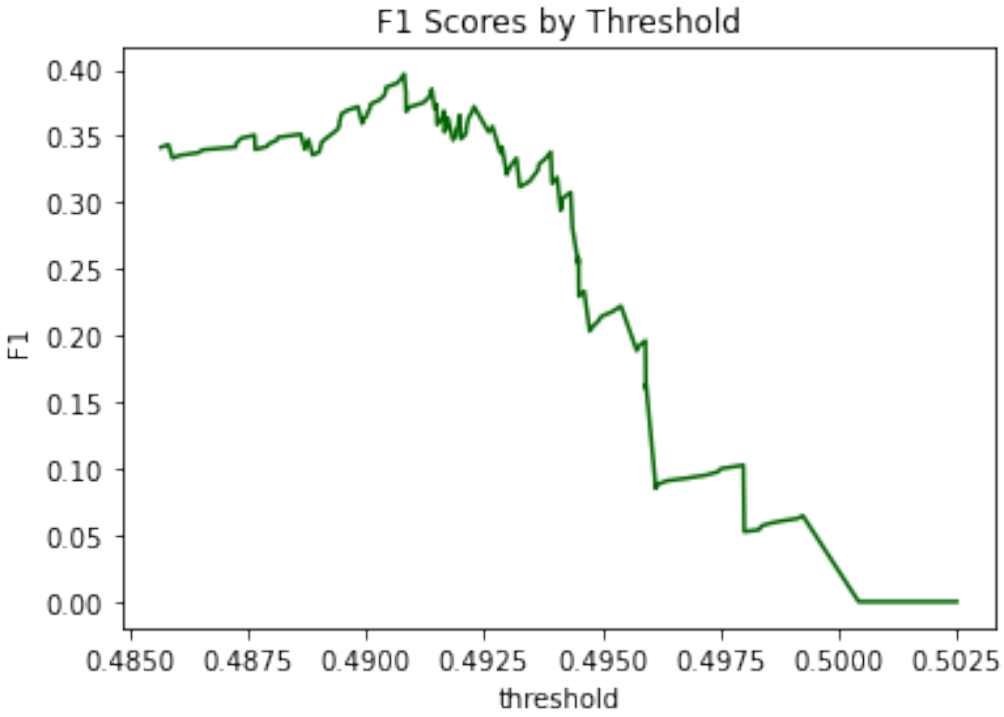


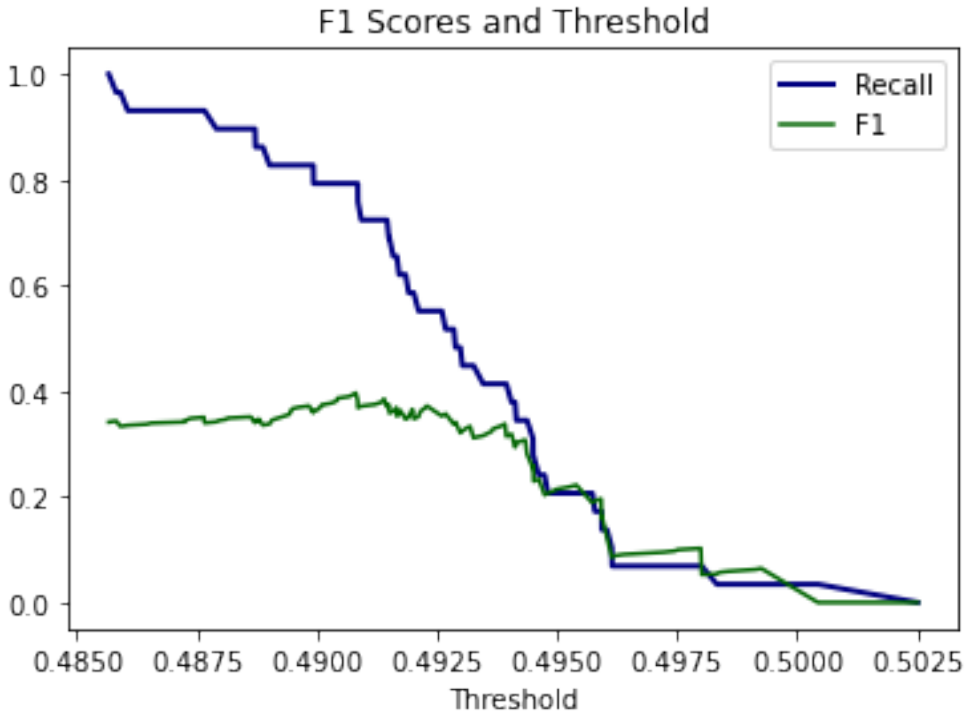
P-R curve


Precision-Recall Curve
Precision
Recall
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0


Precision/Recall by Thresholds
Precision
Recall
Values
Thresholds
1.0
0.8
0.6
0.4
0.2
0.0
0.4850
0.4875
0.4900
0.4925
0.4950
0.4975
0.5000
0.5025

Final Threshold and Explanation:





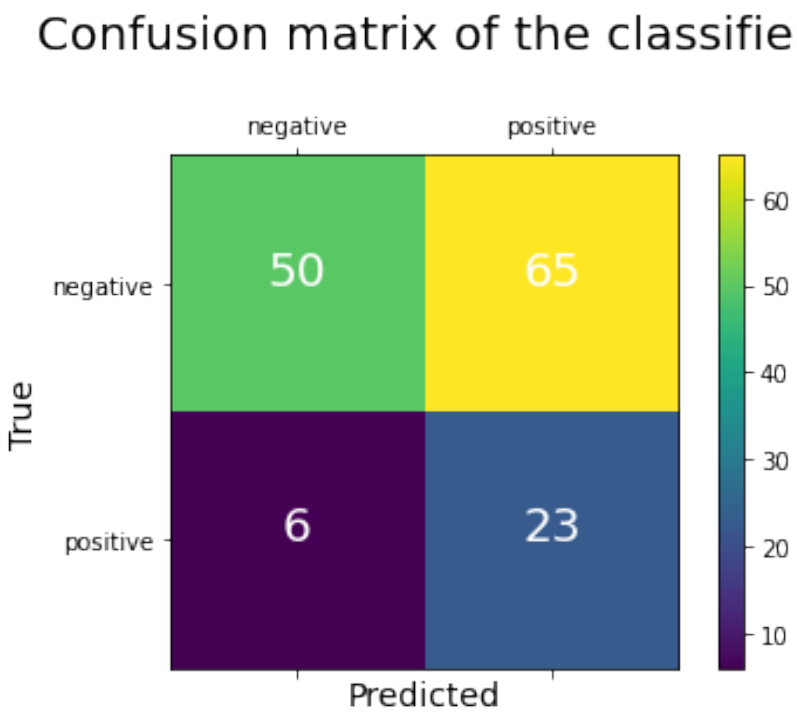
The final threshold is **0.4908**. This is the point where the F1 score graph reaches its maximum value.

Person / Algorithm	F1	95% CI min	95% CI max
Radiologist 1	0.383	0.309	0.453
Radiologist 2	0.356	0.282	0.428
Radiologist 3	0.365	0.291	0.435
Radiologist 4	0.442	0.390	0.492
Radiologist average	0.387	0.330	0.442
CheXNet	0.435	0.387	0.481
ChainRad	**0.397**		

The average F1 score of human radiologists is 0.387 according to paper of CheXNet, that is available here: https://arxiv.org/pdf/1711.05225.pdf

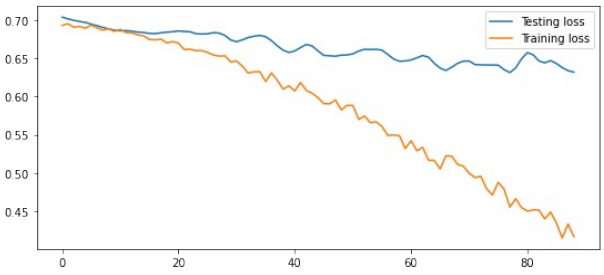
The main goal is for the model to reach the average F1 score of human radiologists. There are the major measures related to the final threshold.

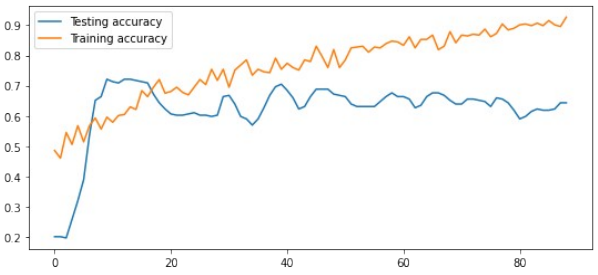
- F1 score max: **0.3966**
- Threshold: **0.4908**
- Precision: **0.2614**
- Recall: **0.7931**



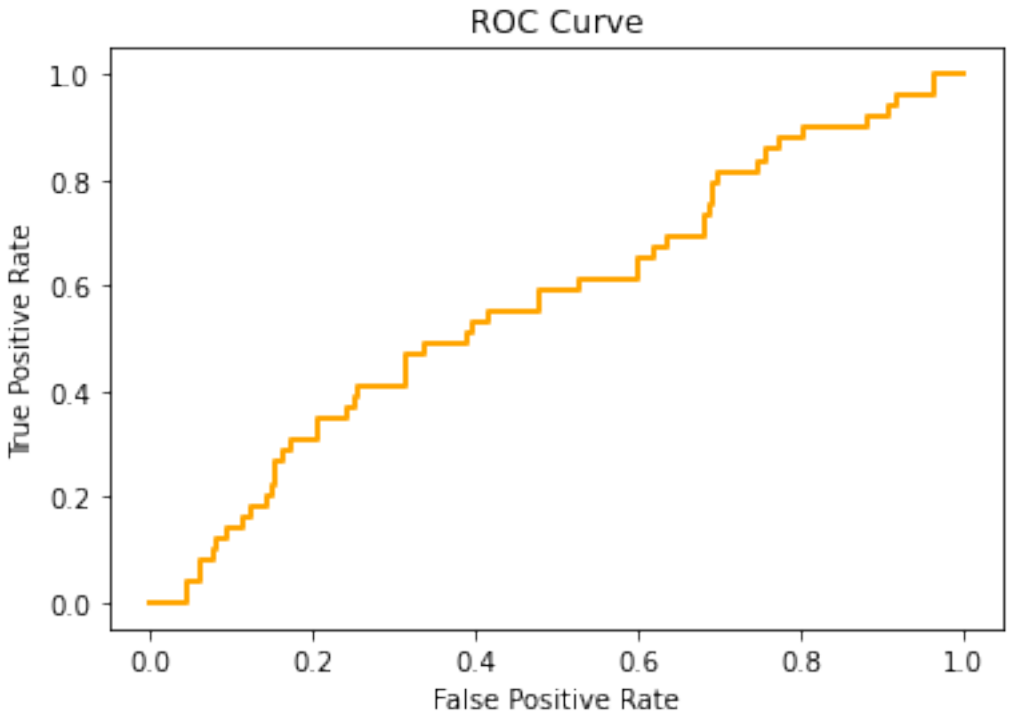
Pneumothorax

Algorithm training performance visualization

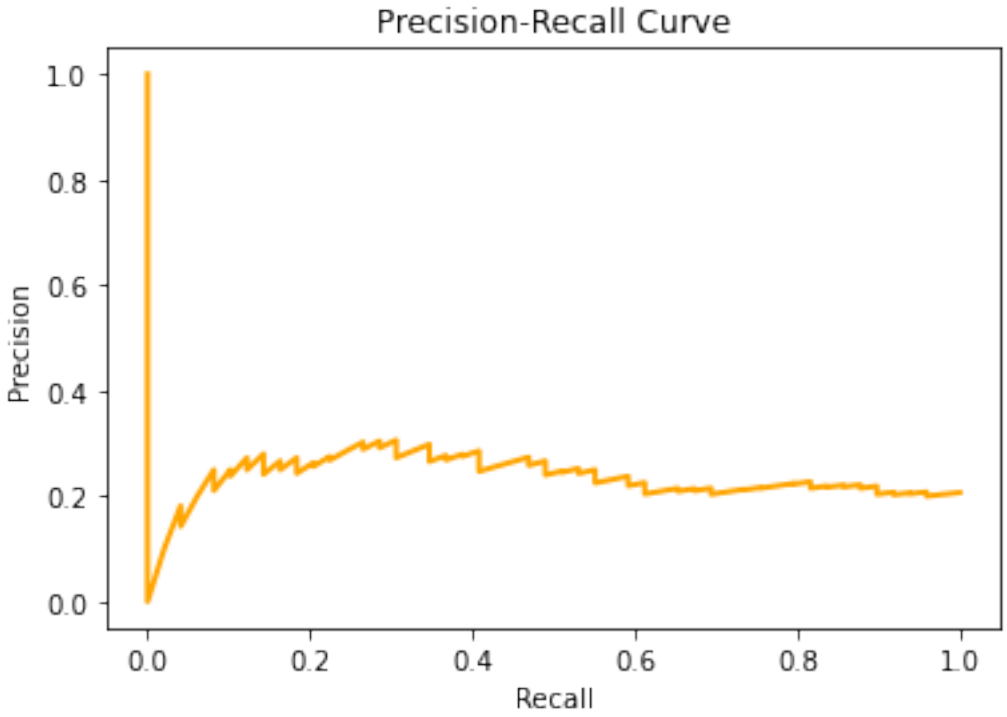


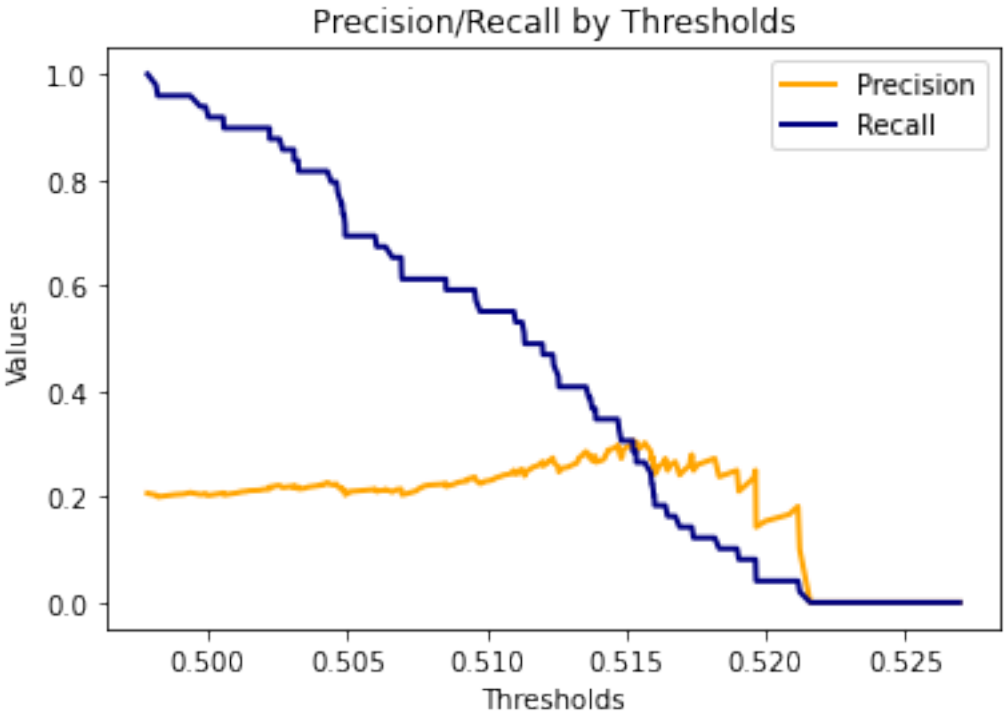


ROC curve shows the model learned something from training dataset. It will be better with smaller learning rate and longer training. This is the part of a future experiment.

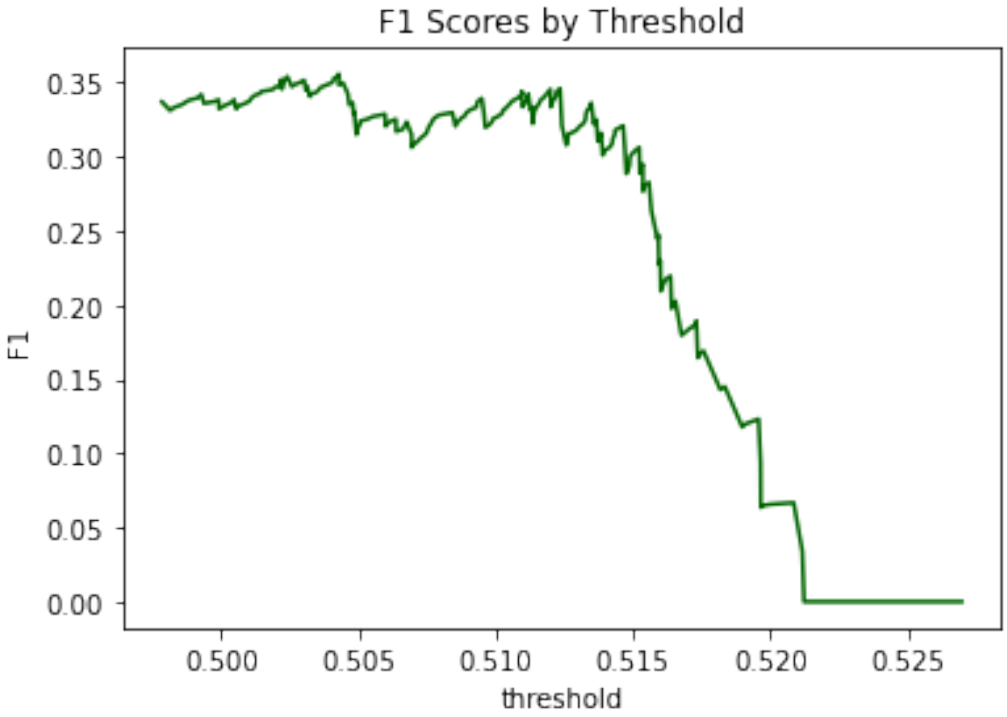


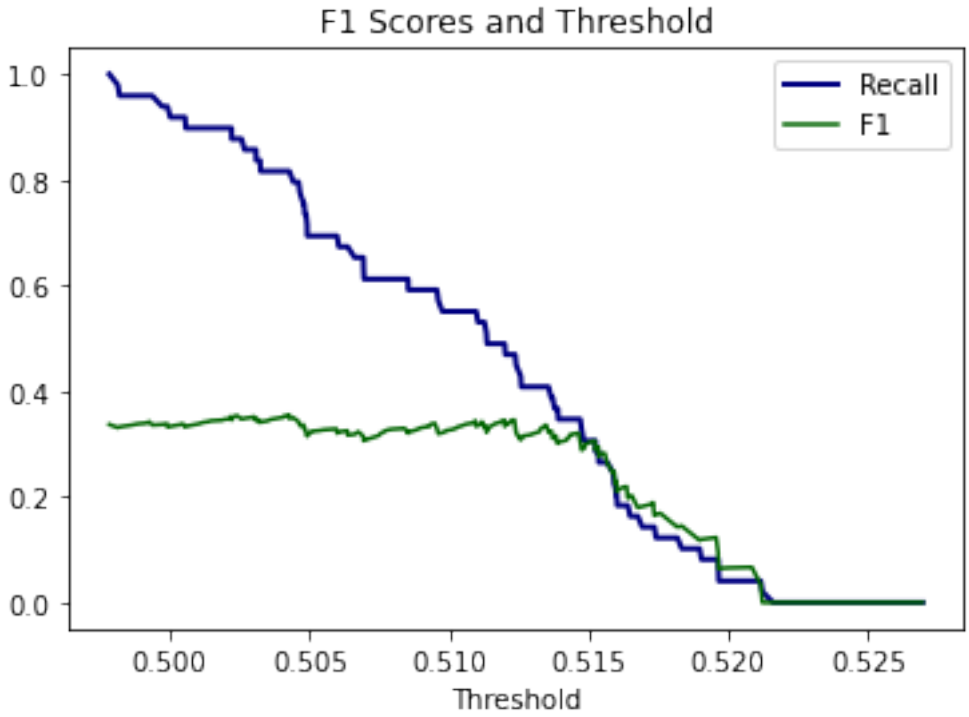
P-R curve


Precision-Recall Curve
Precision
Recall
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0


Precision/Recall by Thresholds
Precision
Recall
Values
Thresholds
1.0
0.8
0.6
0.4
0.2
0.0
0.500
0.505
0.510
0.515
0.520
0.525

Final Threshold and Explanation:





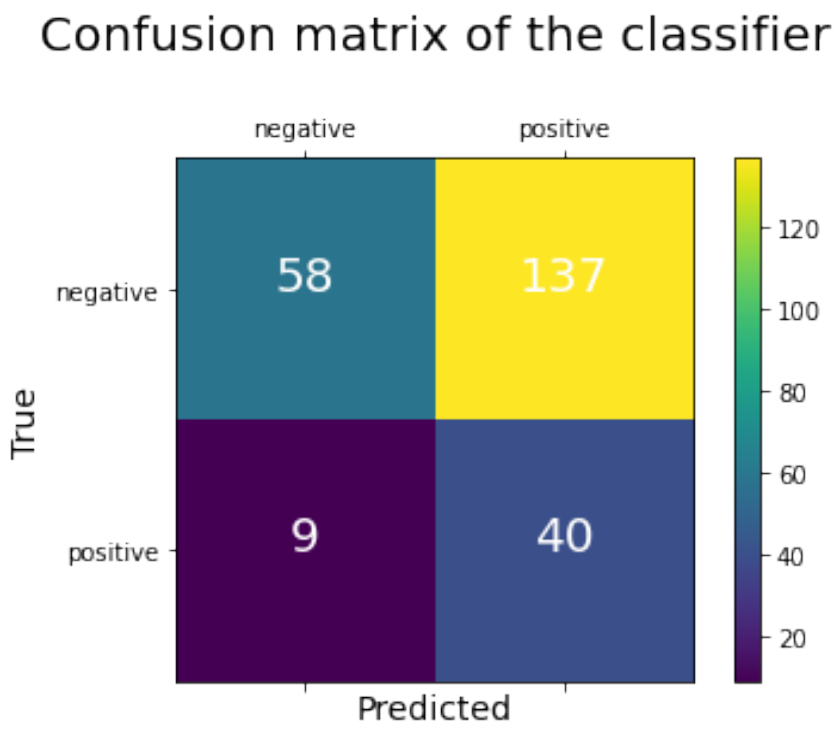
The final threshold is **0.5043**. This is the point where the F1 score graph reaches its maximum value.

Person / Algorithm	F1	95% CI min	95% CI max
Radiologist 1	0.383	0.309	0.453
Radiologist 2	0.356	0.282	0.428
Radiologist 3	0.365	0.291	0.435
Radiologist 4	0.442	0.390	0.492
Radiologist average	0.387	0.330	0.442
CheXNet	0.435	0.387	0.481
ChainRad	**0.356**		

The average F1 score of human radiologists is 0.387 according to paper of CheXNet, that is available here: https://arxiv.org/pdf/1711.05225.pdf

The main goal is for the model to reach the average F1 score of human radiologists. There are the major measures related to the final threshold.

- F1 score max: **0.3556**
- Threshold: **0.5043**
- Precision: **0.2260**
- Recall: **0.8163**



4. Databases

The dataset is publicly available from Kaggle or from AcademicTorrents:
https://www.kaggle.com/nih-chest-xrays/data
https://academictorrents.com/details/557481faacd824c83fbf57dcf7b6da9383b3235a

This dataset is know as NIH Chest X-ray Dataset made by National Institutes of Health (US). Details about dataset are available here: https://www.nih.gov/news-events/news-releases/nih-clinical-center-provides-one-largest-publicly-available-chest-x-ray-datasets-scientific-community

It contains 112120 different images with disease labels from 30805 unique patients. The disease labels system build up from from 15 different element:

1. No Finding 60361
2. Infiltration 19894
3. Effusion 13317
4. Atelectasis 11559
5. Nodule 6331
6. Mass 5782
7. Pneumothorax 5302
8. Consolidation 4667
9. Pleural Thickening 3385
10. Cardiomegaly 2776
11. Emphysema 2516
12. Edema 2303
13. Fibrosis 1686
14. **Pneumonia** 1431
15. Hernia 227

Images can be labeled with multiple labels if the patient suffers from different diseases from categories above. For ChainRad all of them are relevant labels. The dataset was cleaned before creating the training, testing and validation subsets, since it contained some bad records about patients' age, such as: '148, 149, 150, 151, 152, 153, 154, 155, 411, 412, 413 and 414'. These records were removed. Some patients have multiple images. We removed all patient duplications to avoid the data-leak between train, test and validation subsets. This process really decreased the number of available data. After this cleaning process, in the dataset there are data from 28712 unique patients. The creators use Natural Language Processing to text-mine disease classifications from the associated radiological reports. The labels are expected to be >90% accurate.

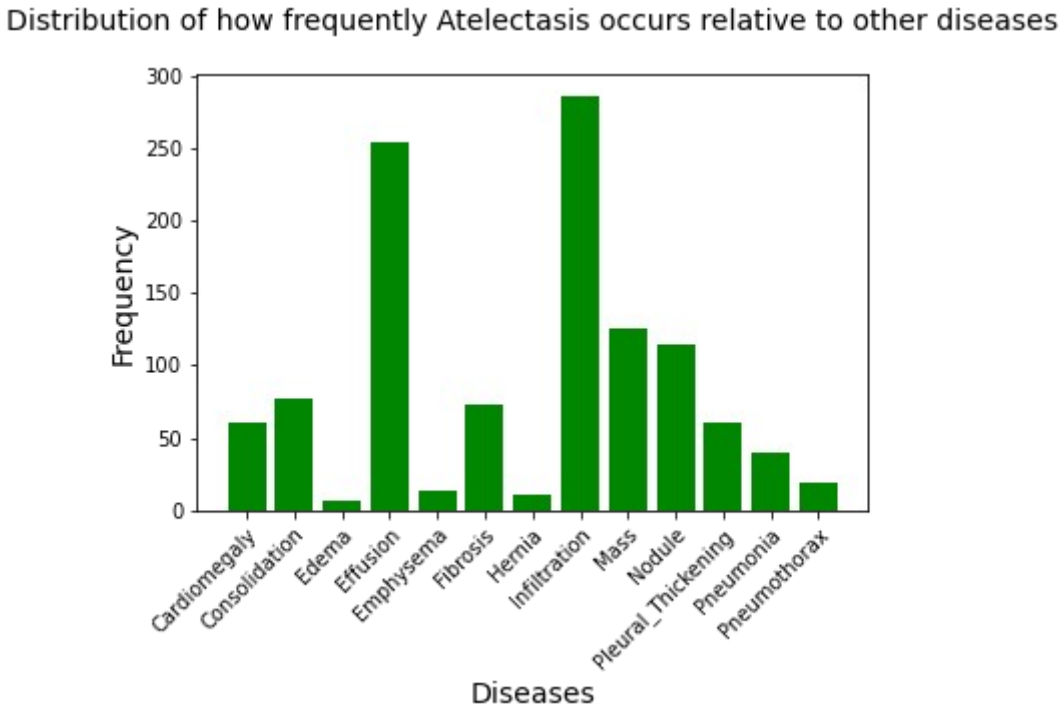
Paper are available here: http://openaccess.thecvf.com/content_cvpr_2017/papers/Wang_ChestX-ray8_Hospital-Scale_Chest_CVPR_2017_paper.pdf

Description of Cleaned Dataset:

The cleaned dataset contains images from 28712 unique patients between 21 and 90 years. The images contain unique diseases and diseases with other co-morbidities. This is a really small dataset that is not suitable for proper training. Co-morbities are really important, since they can make noise during the training. Learning is possible only when there is available a huge number of images about co-morbitidites with a similar histogram spectrum.

Co-diseases with Atelectasis

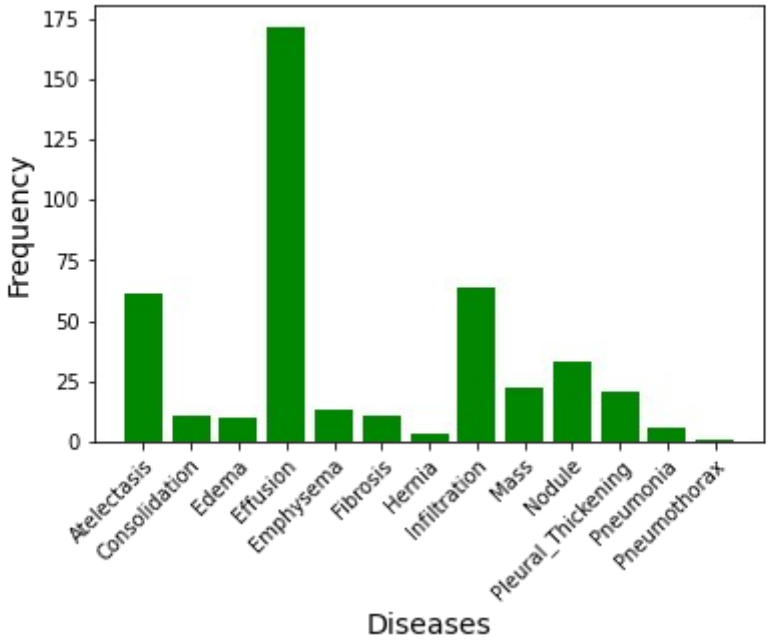
Atelectasis:	864	(43.07%)
Cardiomegaly:	61	(3.04%)
Consolidation:	77	(3.84%)
Edema:	7	(0.35%)
Effusion:	254	(12.66%)
Emphysema:	14	(0.70%)
Fibrosis:	73	(3.64%)
Hernia:	11	(0.55%)
Infiltration:	286	(14.26%)
Mass:	125	(6.23%)
NoFinding:	0	(0.00%)
Nodule:	114	(5.68%)
Pleural_Thickening:	61	(3.04%)
Pneumonia:	40	(1.99%)
Pneumothorax:	19	(0.95%)



Co-diseases with Cardiomegaly

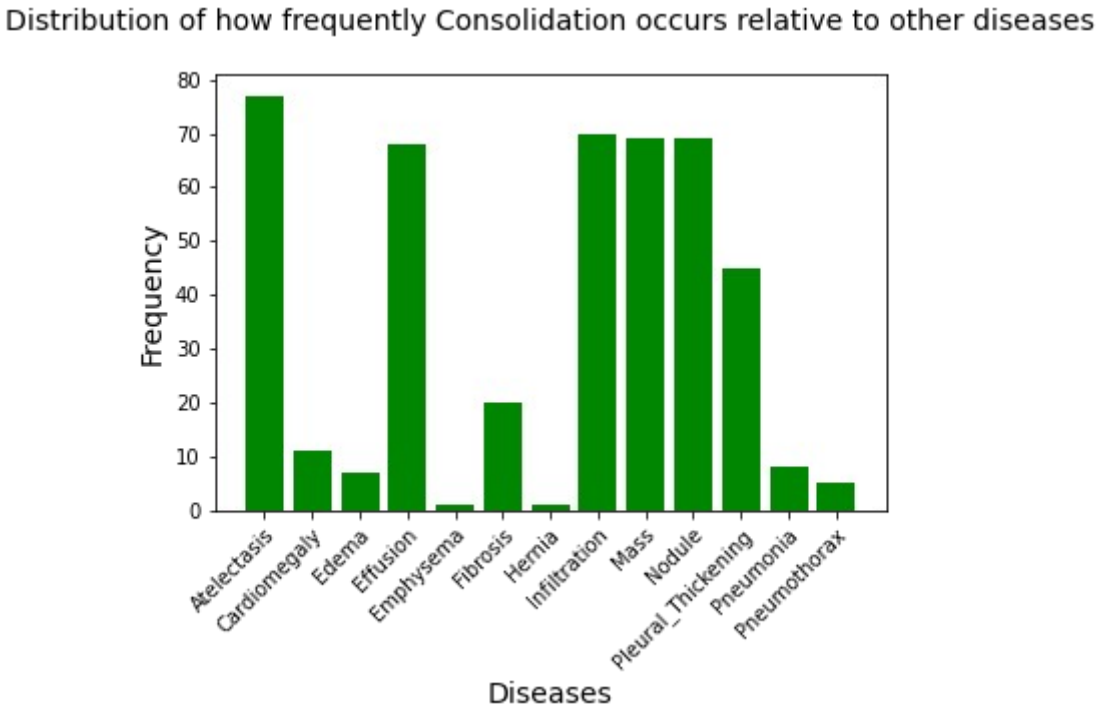
Atelectasis:	61	(7.12%)
Cardiomegaly:	429	(50.06%)
Consolidation:	11	(1.28%)
Edema:	10	(1.17%)
Effusion:	172	(20.07%)
Emphysema:	13	(1.52%)
Fibrosis:	11	(1.28%)
Hernia:	3	(0.35%)
Infiltration:	64	(7.47%)
Mass:	22	(2.57%)
NoFinding:	0	(0.00%)
Nodule:	33	(3.85%)
Pleural_Thickening:	21	(2.45%)
Pneumonia:	6	(0.70%)
Pneumothorax:	1	(0.12%)

Distribution of how frequently Cardiomegaly occurs relative to other diseases



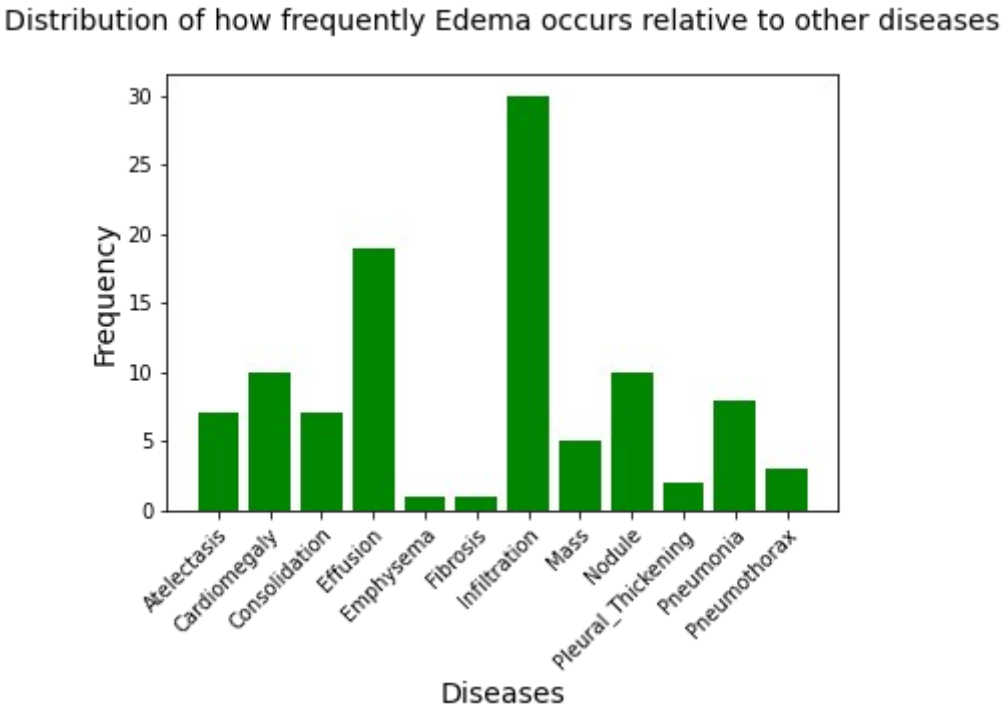
Co-diseases with Consolidation

Atelectasis:	77	(13.51%)
Cardiomegaly:	11	(1.93%)
Consolidation:	119	(20.88%)
Edema:	7	(1.23%)
Effusion:	68	(11.93%)
Emphysema:	1	(0.18%)
Fibrosis:	20	(3.51%)
Hernia:	1	(0.18%)
Infiltration:	70	(12.28%)
Mass:	69	(12.11%)
NoFinding:	0	(0.00%)
Nodule:	69	(12.11%)
Pleural_Thickening:	45	(7.89%)
Pneumonia:	8	(1.40%)
Pneumothorax:	5	(0.88%)



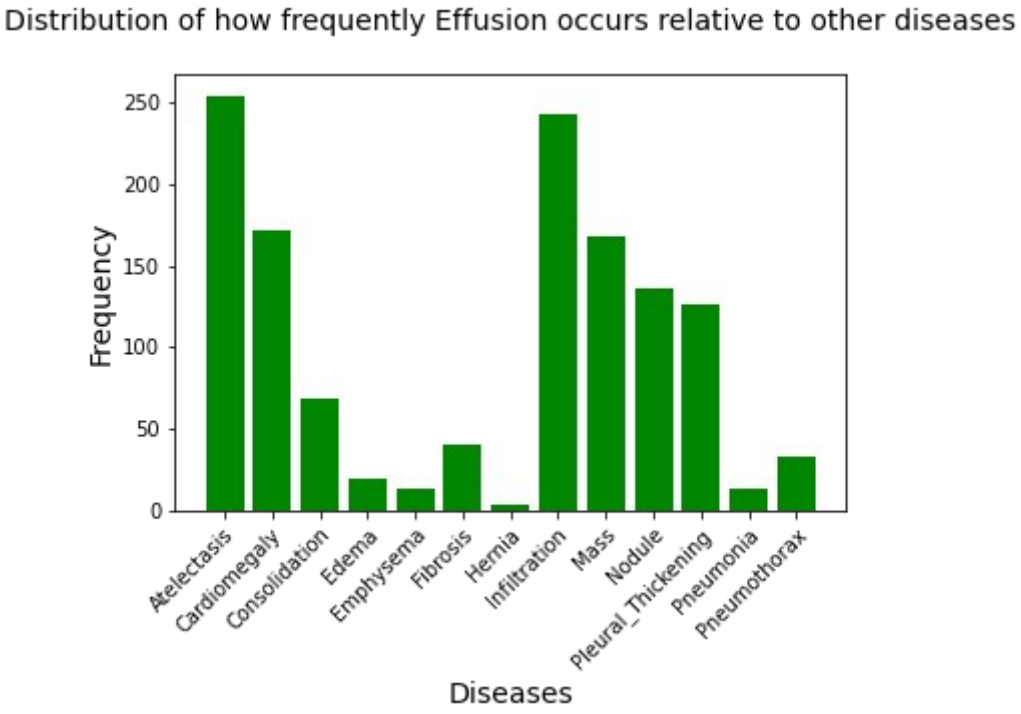
Co-diseases with Edema

Atelectasis:	7	(5.69%)
Cardiomegaly:	10	(8.13%)
Consolidation:	7	(5.69%)
Edema:	20	(16.26%)
Effusion:	19	(15.45%)
Emphysema:	1	(0.81%)
Fibrosis:	1	(0.81%)
Hernia:	0	(0.00%)
Infiltration:	30	(24.39%)
Mass:	5	(4.07%)
NoFinding:	0	(0.00%)
Nodule:	10	(8.13%)
Pleural_Thickening:	2	(1.63%)
Pneumonia:	8	(6.50%)
Pneumothorax:	3	(2.44%)



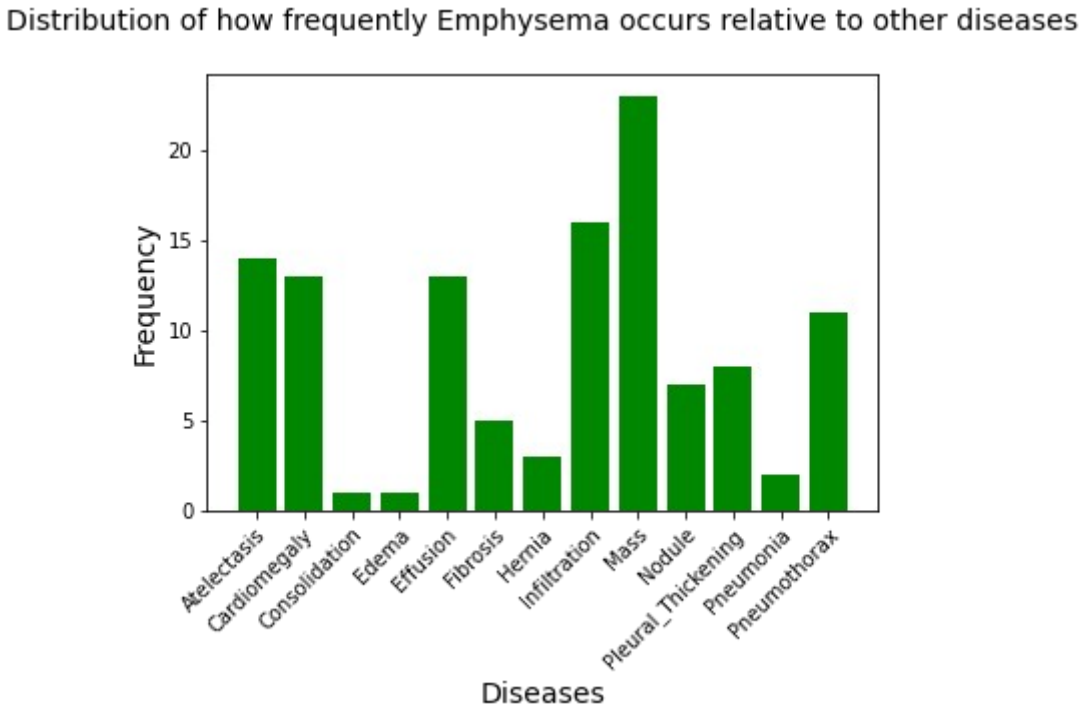
Co-diseases with Effusion

Atelectasis:	254	(15.18%)
Cardiomegaly:	172	(10.28%)
Consolidation:	68	(4.06%)
Edema:	19	(1.14%)
Effusion:	385	(23.01%)
Emphysema:	13	(0.78%)
Fibrosis:	40	(2.39%)
Hernia:	3	(0.18%)
Infiltration:	243	(14.52%)
Mass:	168	(10.04%)
NoFinding:	0	(0.00%)
Nodule:	136	(8.13%)
Pleural_Thickening:	126	(7.53%)
Pneumonia:	13	(0.78%)
Pneumothorax:	33	(1.97%)



Co-diseases with Emphysema

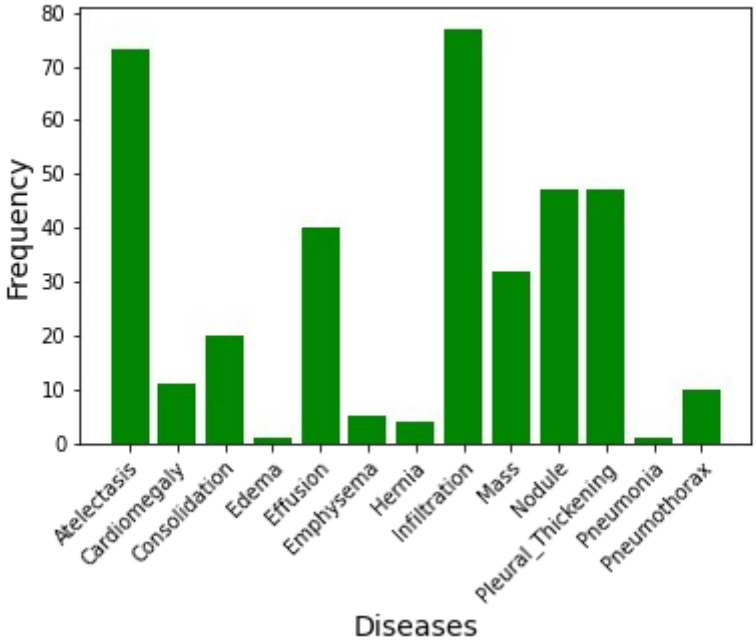
Atelectasis:	14	(4.78%)
Cardiomegaly:	13	(4.44%)
Consolidation:	1	(0.34%)
Edema:	1	(0.34%)
Effusion:	13	(4.44%)
Emphysema:	176	(60.07%)
Fibrosis:	5	(1.71%)
Hernia:	3	(1.02%)
Infiltration:	16	(5.46%)
Mass:	23	(7.85%)
NoFinding:	0	(0.00%)
Nodule:	7	(2.39%)
Pleural_Thickening:	8	(2.73%)
Pneumonia:	2	(0.68%)
Pneumothorax:	11	(3.75%)



Co-diseases with Fibrosis

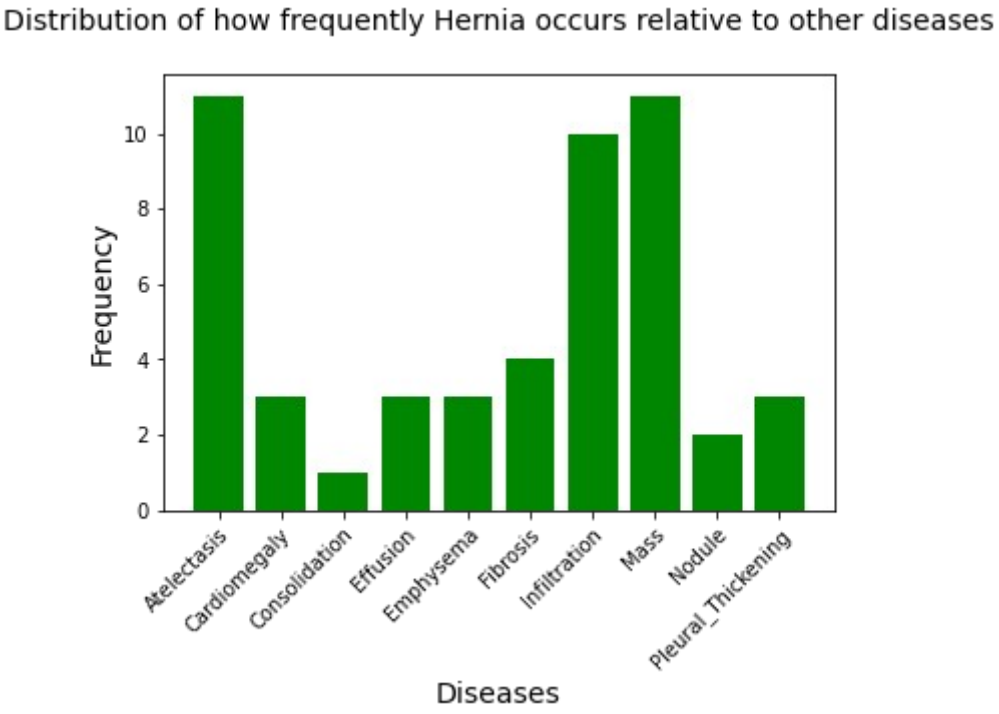
Atelectasis:	73	(10.93%)
Cardiomegaly:	11	(1.65%)
Consolidation:	20	(2.99%)
Edema:	1	(0.15%)
Effusion:	40	(5.99%)
Emphysema:	5	(0.75%)
Fibrosis:	300	(44.91%)
Hernia:	4	(0.60%)
Infiltration:	77	(11.53%)
Mass:	32	(4.79%)
NoFinding:	0	(0.00%)
Nodule:	47	(7.04%)
Pleural_Thickening:	47	(7.04%)
Pneumonia:	1	(0.15%)
Pneumothorax:	10	(1.50%)

Distribution of how frequently Fibrosis occurs relative to other diseases



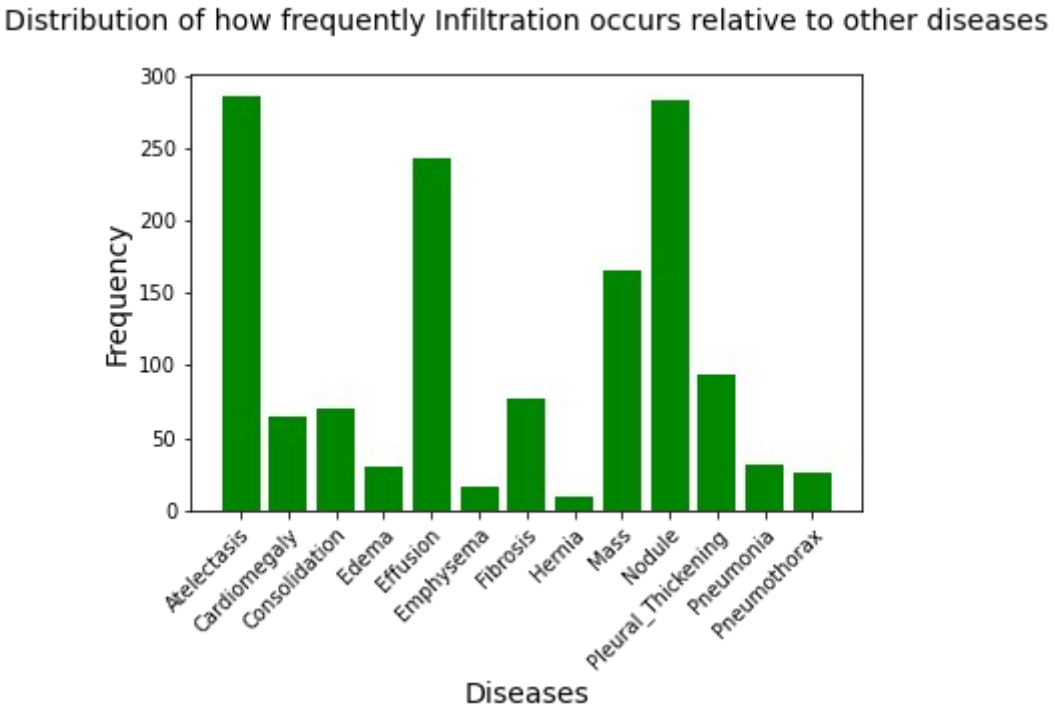
Co-diseases with Hernia

Atelectasis:	11	(11.58%)
Cardiomegaly:	3	(3.16%)
Consolidation:	1	(1.05%)
Edema:	0	(0.00%)
Effusion:	3	(3.16%)
Emphysema:	3	(3.16%)
Fibrosis:	4	(4.21%)
Hernia:	44	(46.32%)
Infiltration:	10	(10.53%)
Mass:	11	(11.58%)
NoFinding:	0	(0.00%)
Nodule:	2	(2.11%)
Pleural_Thickening:	3	(3.16%)
Pneumonia:	0	(0.00%)
Pneumothorax:	0	(0.00%)



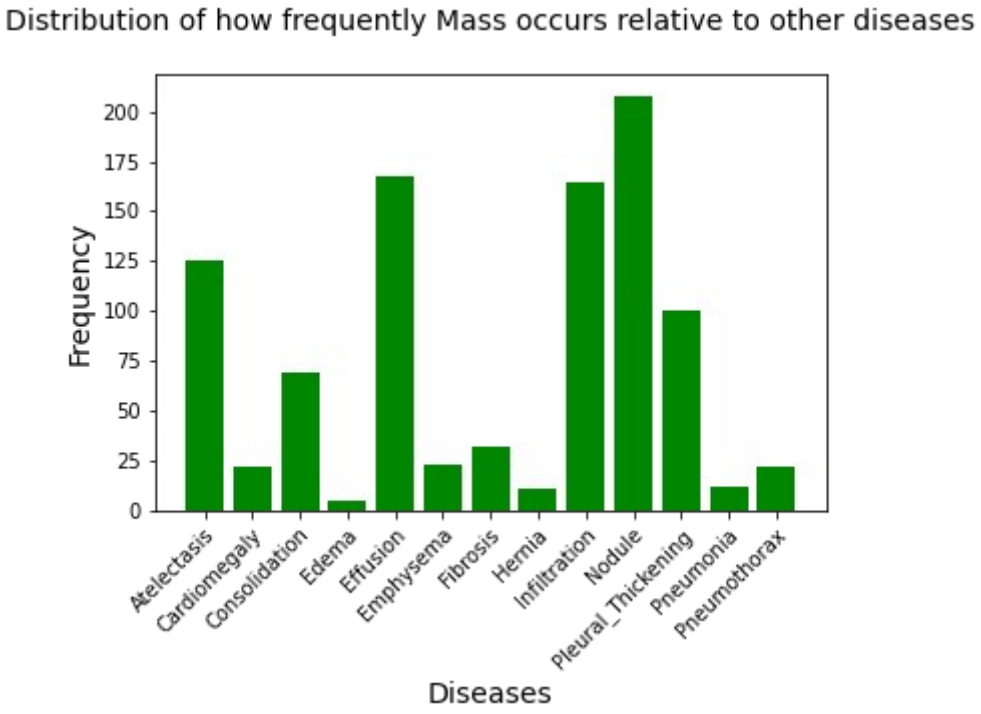
Co-diseases with Infiltration

Atelectasis:	286	(7.59%)
Cardiomegaly:	64	(1.70%)
Consolidation:	70	(1.86%)
Edema:	30	(0.80%)
Effusion:	243	(6.45%)
Emphysema:	16	(0.42%)
Fibrosis:	77	(2.04%)
Hernia:	10	(0.27%)
Infiltration:	2372	(62.98%)
Mass:	165	(4.38%)
NoFinding:	0	(0.00%)
Nodule:	283	(7.51%)
Pleural_Thickening:	93	(2.47%)
Pneumonia:	31	(0.82%)
Pneumothorax:	26	(0.69%)



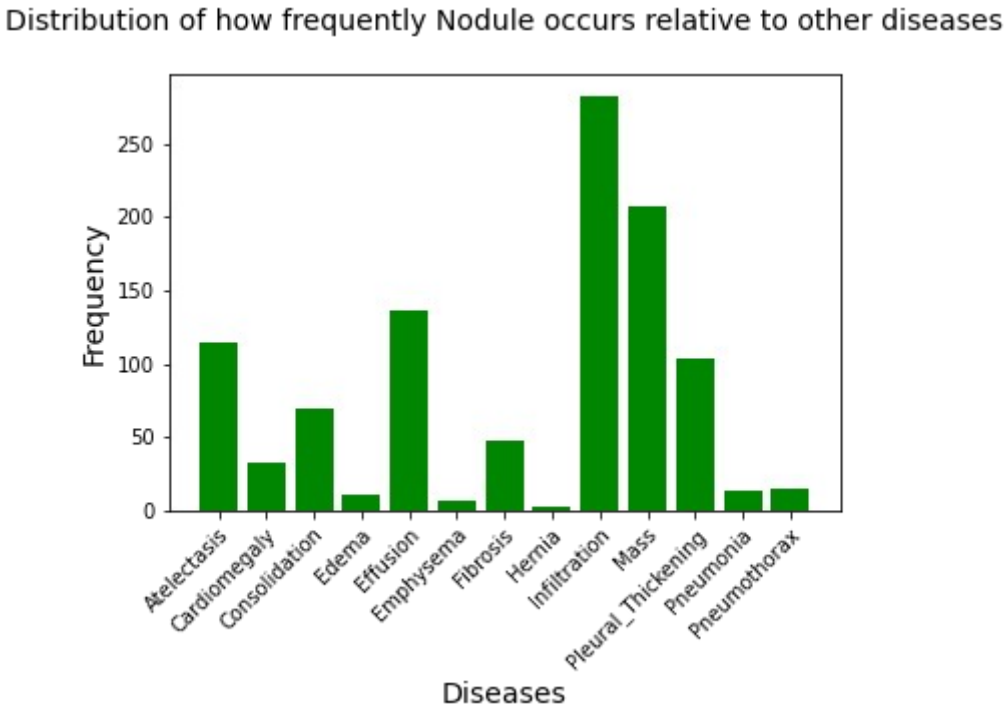
Co-diseases with Mass

Atelectasis:	125	(8.06%)
Cardiomegaly:	22	(1.42%)
Consolidation:	69	(4.45%)
Edema:	5	(0.32%)
Effusion:	168	(10.83%)
Emphysema:	23	(1.48%)
Fibrosis:	32	(2.06%)
Hernia:	11	(0.71%)
Infiltration:	165	(10.64%)
Mass:	589	(37.98%)
NoFinding:	0	(0.00%)
Nodule:	208	(13.41%)
Pleural_Thickening:	100	(6.45%)
Pneumonia:	12	(0.77%)
Pneumothorax:	22	(1.42%)



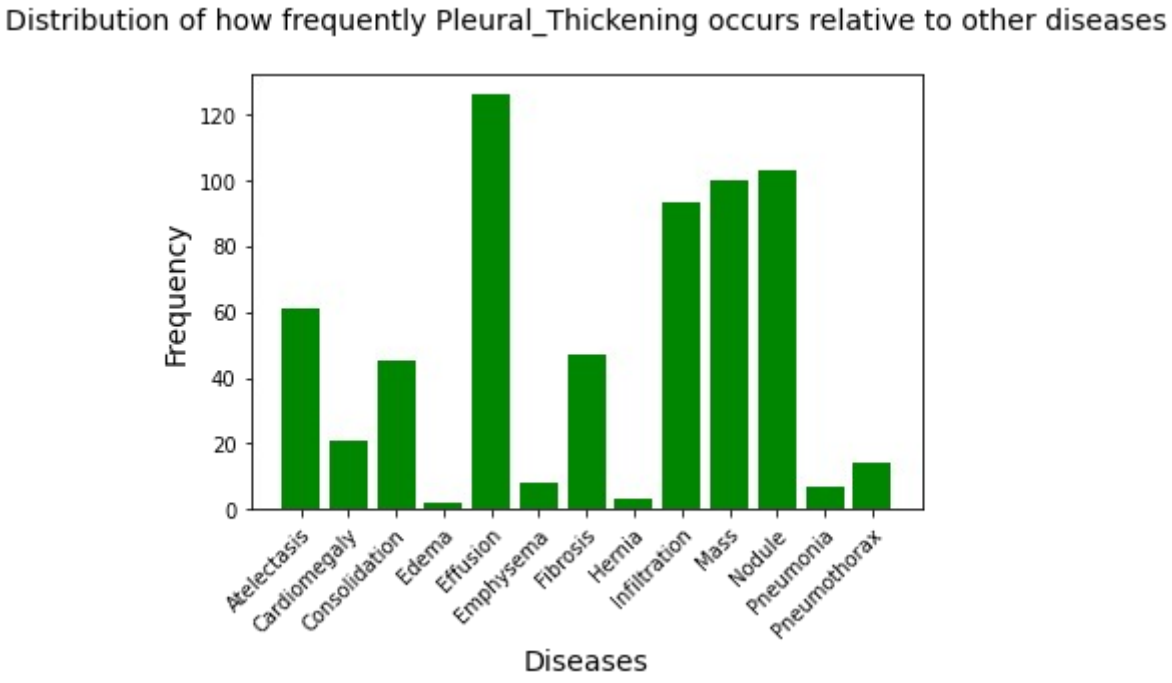
Co-diseases with Nodule

Atelectasis:	114	(6.01%)
Cardiomegaly:	33	(1.74%)
Consolidation:	69	(3.64%)
Edema:	10	(0.53%)
Effusion:	136	(7.17%)
Emphysema:	7	(0.37%)
Fibrosis:	47	(2.48%)
Hernia:	2	(0.11%)
Infiltration:	283	(14.91%)
Mass:	208	(10.96%)
NoFinding:	0	(0.00%)
Nodule:	857	(45.15%)
Pleural_Thickening:	103	(5.43%)
Pneumonia:	14	(0.74%)
Pneumothorax:	15	(0.79%)



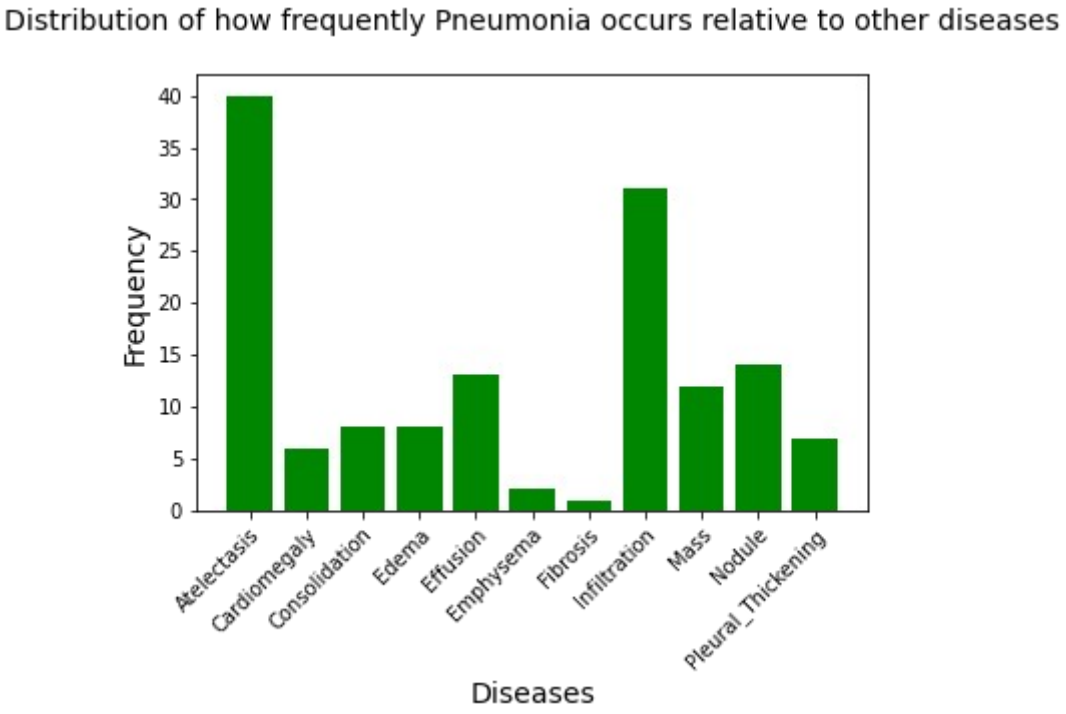
Co-diseases with Pleural Thickening

Atelectasis:	61	(6.26%)
Cardiomegaly:	21	(2.16%)
Consolidation:	45	(4.62%)
Edema:	2	(0.21%)
Effusion:	126	(12.94%)
Emphysema:	8	(0.82%)
Fibrosis:	47	(4.83%)
Hernia:	3	(0.31%)
Infiltration:	93	(9.55%)
Mass:	100	(10.27%)
NoFinding:	0	(0.00%)
Nodule:	103	(10.57%)
Pleural_Thickening:	344	(35.32%)
Pneumonia:	7	(0.72%)
Pneumothorax:	14	(1.44%)



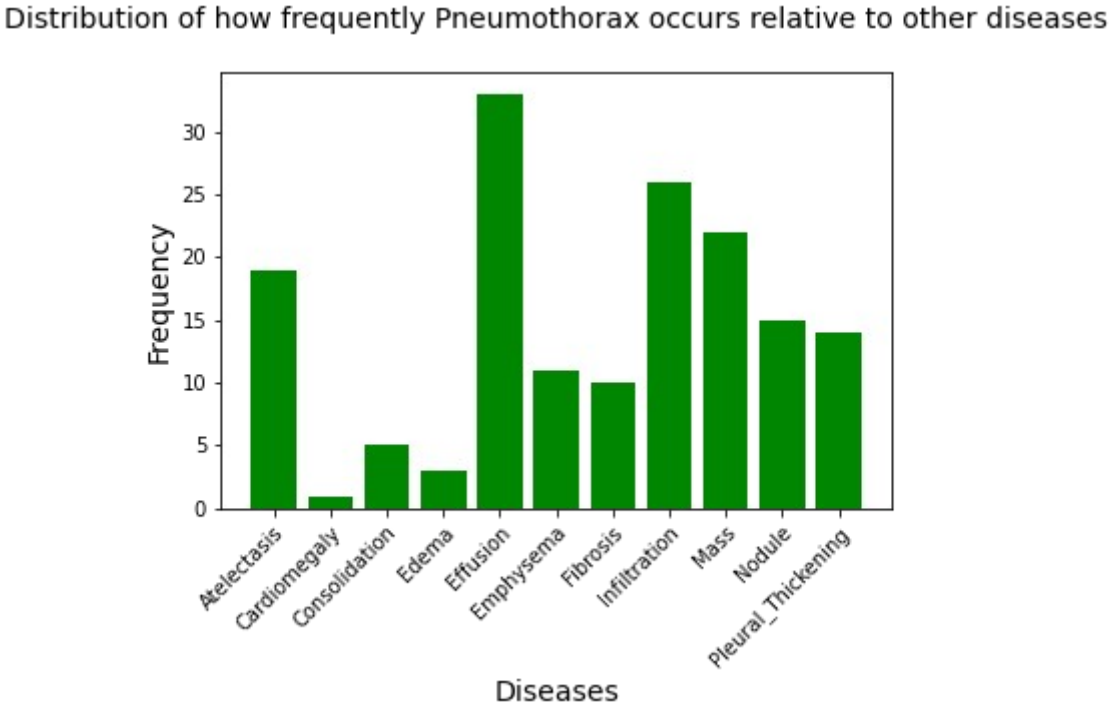
Co-diseases with Pneumonia

Atelectasis:	40	(21.05%)
Cardiomegaly:	6	(3.16%)
Consolidation:	8	(4.21%)
Edema:	8	(4.21%)
Effusion:	13	(6.84%)
Emphysema:	2	(1.05%)
Fibrosis:	1	(0.53%)
Hernia:	0	(0.00%)
Infiltration:	31	(16.32%)
Mass:	12	(6.32%)
NoFinding:	0	(0.00%)
Nodule:	14	(7.37%)
Pleural_Thickening:	7	(3.68%)
Pneumonia:	48	(25.26%)
Pneumothorax:	0	(0.00%)

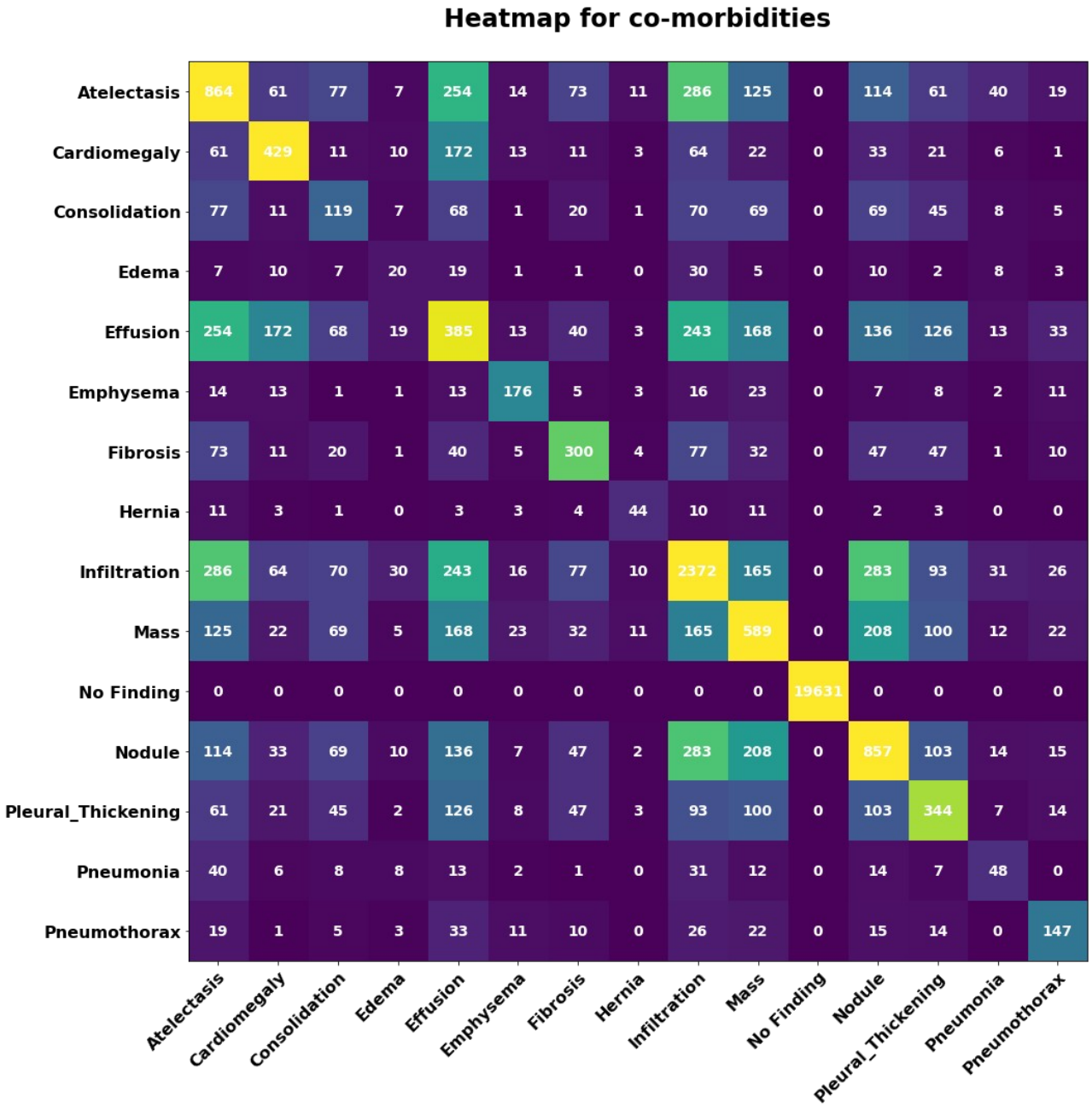


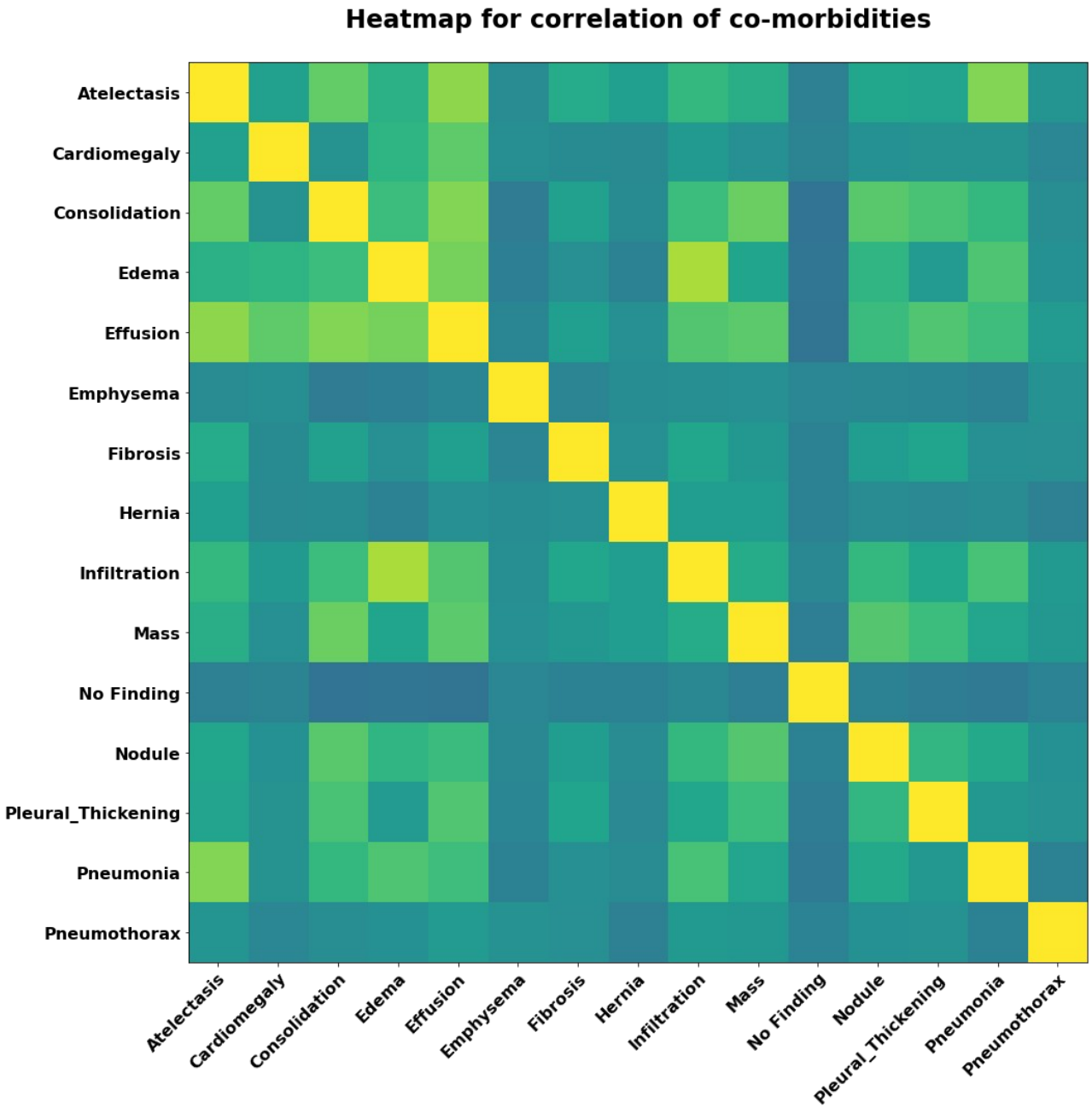
Co-diseases with Pneumothorax

Atelectasis:	19	(6.21%)
Cardiomegaly:	1	(0.33%)
Consolidation:	5	(1.63%)
Edema:	3	(0.98%)
Effusion:	33	(10.78%)
Emphysema:	11	(3.59%)
Fibrosis:	10	(3.27%)
Hernia:	0	(0.00%)
Infiltration:	26	(8.50%)
Mass:	22	(7.19%)
NoFinding:	0	(0.00%)
Nodule:	15	(4.90%)
Pleural_Thickening:	14	(4.58%)
Pneumonia:	0	(0.00%)
Pneumothorax:	147	(48.04%)



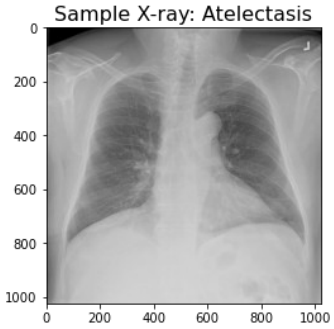
Heatmaps

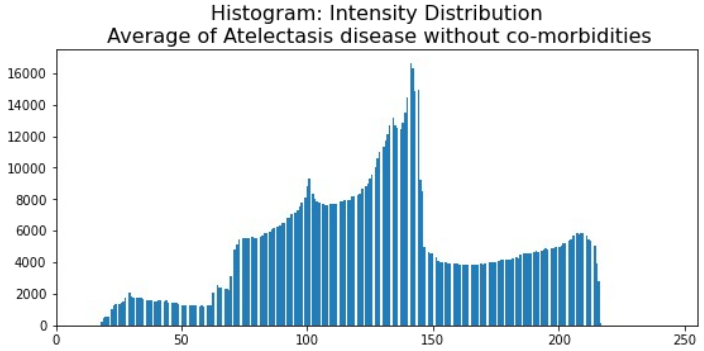

Heatmap for co-morbidities
Atelectasis 864 61 77 7 254 14 73 11 286 125 0 114 61 40 19
Cardiomegaly 61 429 11 10 172 13 11 3 64 22 0 33 21 6 1
Consolidation 77 11 119 7 68 1 20 1 70 69 0 69 45 8 5
Edema 7 10 7 20 19 1 1 0 30 5 0 10 2 8 3
Effusion 254 172 68 19 385 13 40 3 243 168 0 136 126 13 33
Emphysema 14 13 1 1 13 176 5 3 16 23 0 7 8 2 11
Fibrosis 73 11 20 1 40 5 300 4 77 32 0 47 47 1 10
Hernia 11 3 1 0 3 3 4 44 10 11 0 2 3 0 0
Infiltration 286 64 70 30 243 16 77 10 2372 165 0 283 93 31 26
Mass 125 22 69 5 168 23 32 11 165 589 0 208 100 12 22
No Finding 0 0 0 0 0 0 0 0 0 0 19631 0 0 0 0
Nodule 114 33 69 10 136 7 47 2 283 208 0 857 103 14 15
Pleural_Thickening 61 21 45 2 126 8 47 3 93 100 0 103 344 7 14
Pneumonia 40 6 8 8 13 2 1 0 31 12 0 14 7 48 0
Pneumothorax 19 1 5 3 33 11 10 0 26 22 0 15 14 0 147
Atelectasis
Cardiomegaly
Consolidation
Edema
Effusion
Emphysema
Fibrosis
Hernia
Infiltration
Mass
No Finding
Nodule
Pleural_Thickening
Pneumonia
Pneumothorax


Heatmap for correlation of co-morbidities
Atelectasis
Cardiomegaly
Consolidation
Edema
Effusion
Emphysema
Fibrosis
Hernia
Infiltration
Mass
No Finding
Nodule
Pleural_Thickening
Pneumonia
Pneumothorax

Unique diseases and no finding in numbers with histograms

Atelectasis

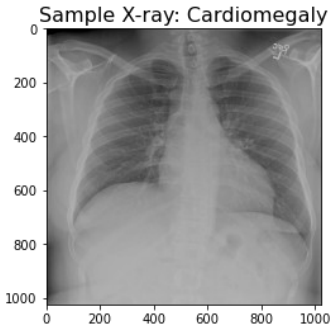


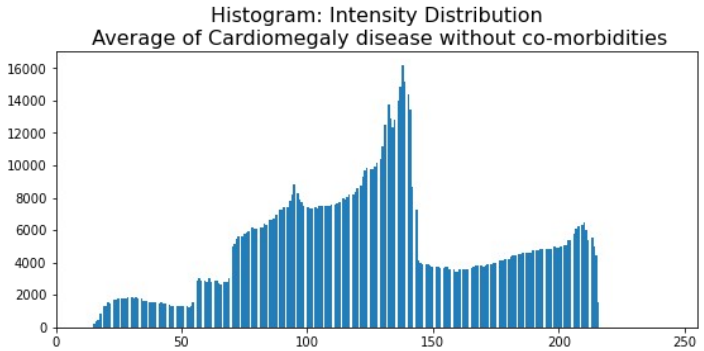


min: 18 | max: 217 | median: 130.00 | mean: 130.14 | avg: 130.14 | std: 44.33

percentiles: 18.0 - 75.0 - 93.0 - 106.0 - 119.0 - 130.0 - 139.0 - 146.0 - 172.0 - 196.0 - 217.0

Cardiomegaly

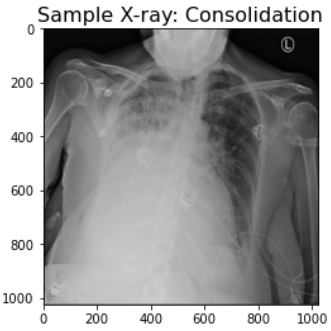


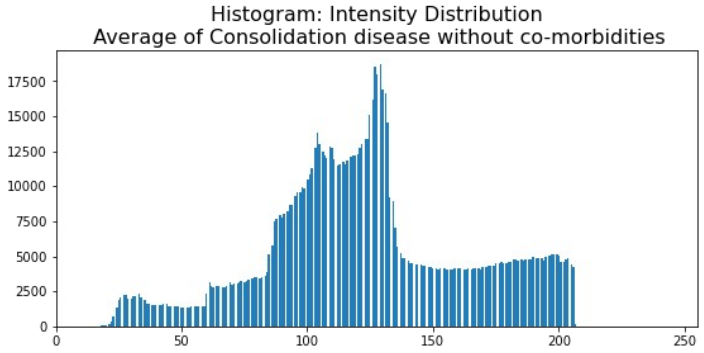


min: 15 | max: 216 | median: 127.00 | mean: 127.38 | avg: 127.38 | std: 46.11

percentiles: 15.0 - 71.0 - 88.0 - 102.0 - 115.0 - 127.0 - 136.0 - 145.0 - 173.0 - 196.0 - 216.0

Consolidation

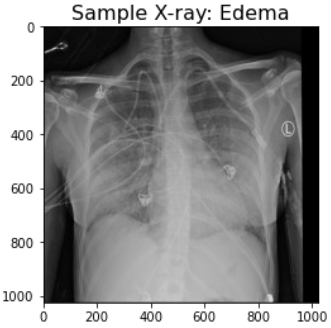


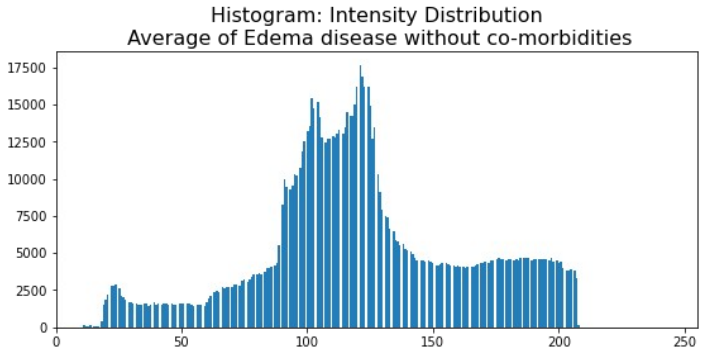


min: 18 | max: 207 | median: 121.00 | mean: 123.51 | avg: 123.51 | std: 40.30

percentiles: 18.0 - 75.0 - 94.0 - 104.0 - 112.0 - 121.0 - 128.0 - 137.0 - 161.0 - 185.0 - 207.0

Edema

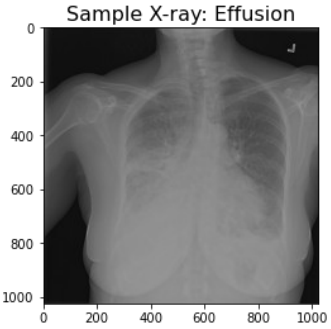


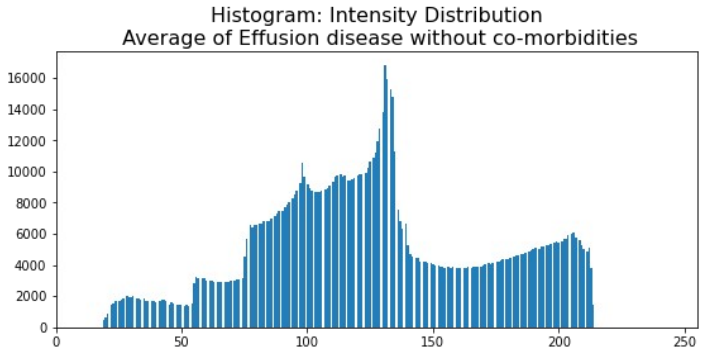


min: 11 | max: 209 | median: 118.00 | mean: 121.71 | avg: 121.71 | std: 41.12

percentiles: 11.0 - 73.0 - 94.0 - 103.0 - 111.0 - 118.0 - 125.0 - 137.0 - 159.0 - 184.0 - 209.0

Effusion

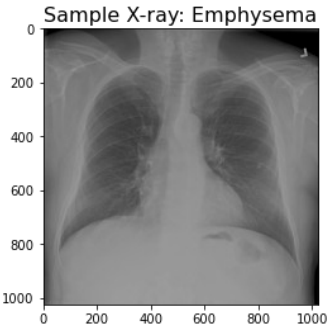


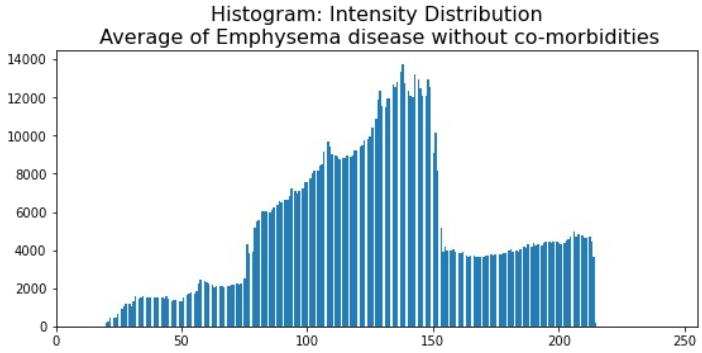


min: 18 | max: 214 | median: 123.00 | mean: 125.87 | avg: 125.87 | std: 45.16

percentiles: 18.0 - 70.0 - 89.0 - 101.0 - 112.0 - 123.0 - 132.0 - 145.0 - 172.0 - 194.0 - 214.0

Emphysema

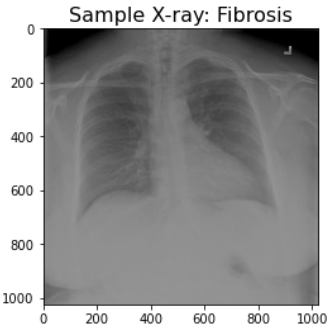


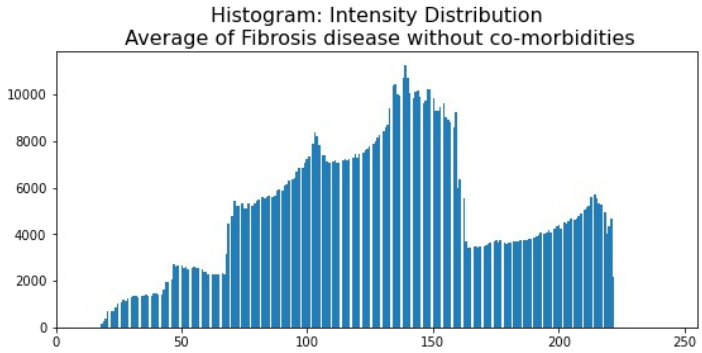


min: 19 | max: 215 | median: 130.00 | mean: 129.90 | avg: 129.90 | std: 41.29

percentiles: 19.0 - 79.0 - 96.0 - 109.0 - 120.0 - 130.0 - 139.0 - 147.0 - 164.0 - 191.0 - 215.0

Fibrosis

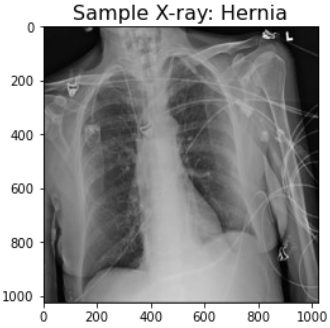


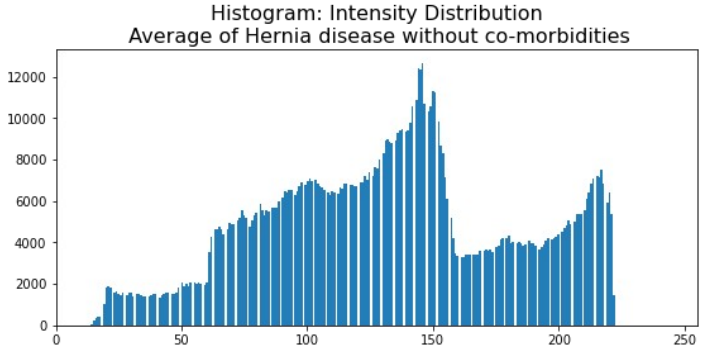


min: 17 | max: 222 | median: 133.00 | mean: 131.81 | avg: 131.81 | std: 46.35

percentiles: 17.0 - 72.0 - 91.0 - 106.0 - 120.0 - 133.0 - 143.0 - 154.0 - 173.0 - 200.0 – 222.0

Hernia

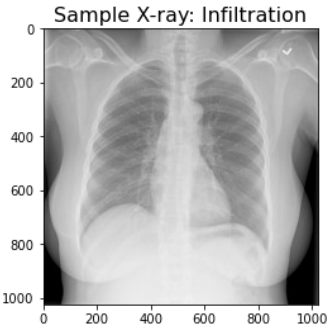


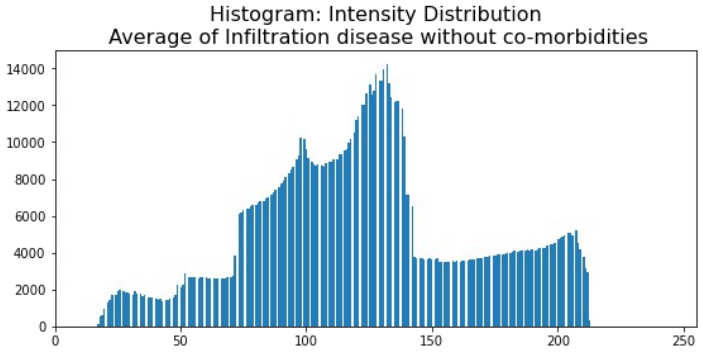


min: 14 | max: 223 | median: 133.00 | mean: 132.03 | avg: 132.03 | std: 48.92

percentiles: 14.0 - 68.0 - 88.0 - 104.0 - 120.0 - 133.0 - 144.0 - 154.0 - 179.0 - 205.0 - 223.0

Infiltration

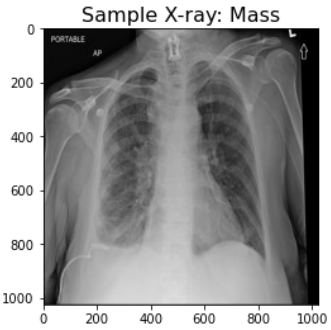


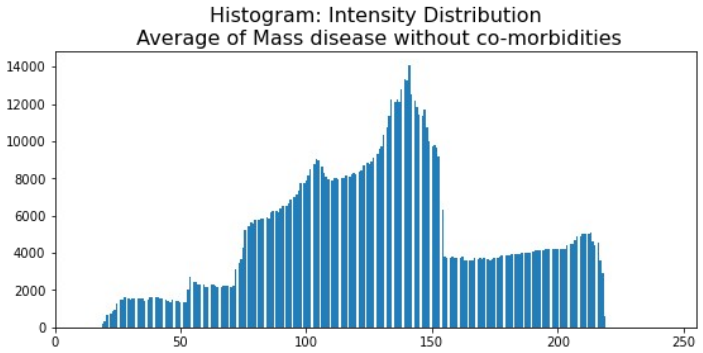


min: 17 | max: 213 | median: 122.00 | mean: 122.64 | avg: 122.64 | std: 43.29

percentiles: 17.0 - 70.0 - 87.0 - 100.0 - 111.0 - 122.0 - 130.0 - 138.0 - 161.0 - 189.0 - 213.0

Mass

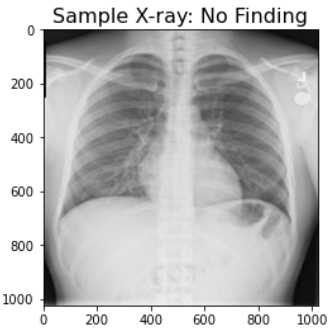


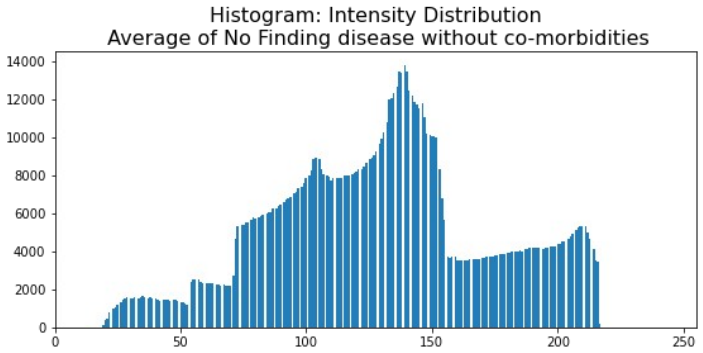


min: 18 | max: 219 | median: 131.00 | mean: 130.06 | avg: 130.06 | std: 43.72

percentiles: 18.0 - 76.0 - 93.0 - 106.0 - 119.0 - 131.0 - 140.0 - 148.0 - 168.0 - 195.0 - 219.0

No Finding

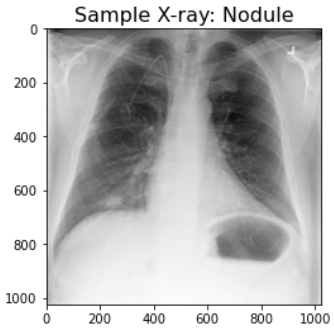


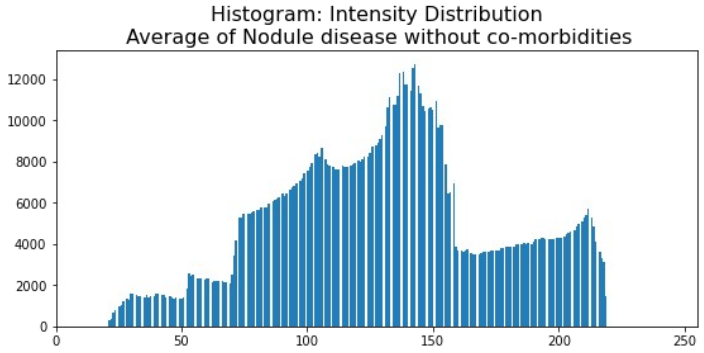


min: 19 | max: 217 | median: 131.00 | mean: 129.33 | avg: 129.33 | std: 43.34

percentiles: 19.0 - 75.0 - 92.0 - 106.0 - 119.0 - 131.0 - 139.0 - 148.0 - 167.0 - 193.0 - 217.0

Nodule

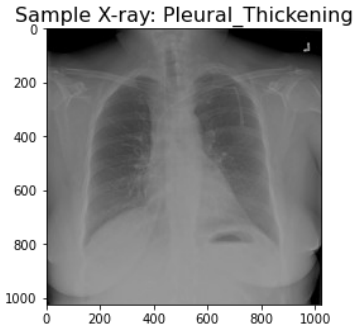


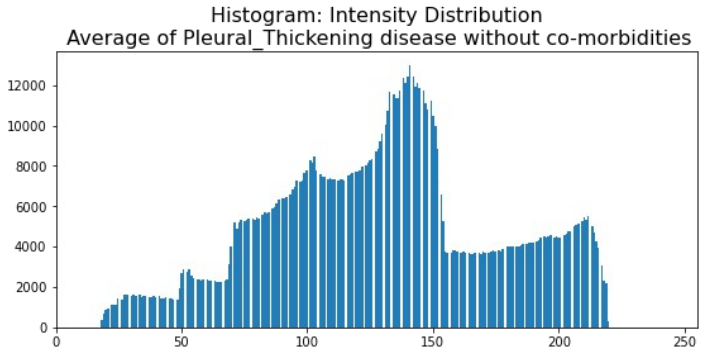


min: 21 | max: 219 | median: 132.00 | mean: 130.85 | avg: 130.85 | std: 43.82

percentiles: 21.0 - 75.0 - 93.0 - 107.0 - 120.0 - 132.0 - 141.0 - 151.0 - 169.0 - 196.0 - 219.0

Pleural Thickening

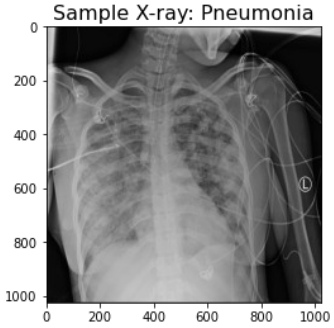


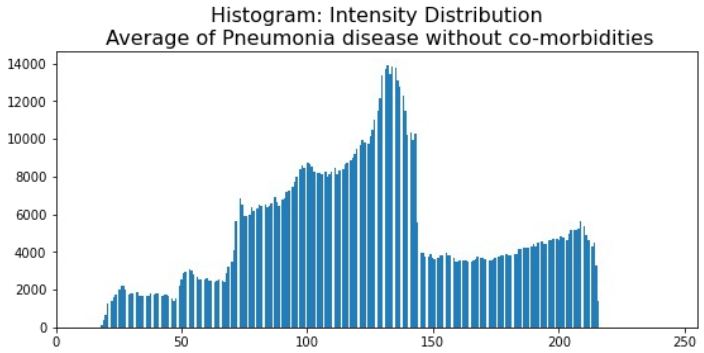


min: 18 | max: 220 | median: 131.00 | mean: 129.61 | avg: 129.61 | std: 45.21

percentiles: 18.0 - 72.0 - 91.0 - 105.0 - 119.0 - 131.0 - 140.0 - 149.0 - 170.0 - 196.0 - 220.0

Pneumonia

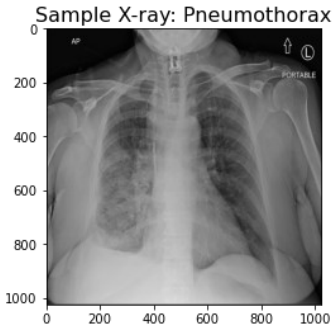


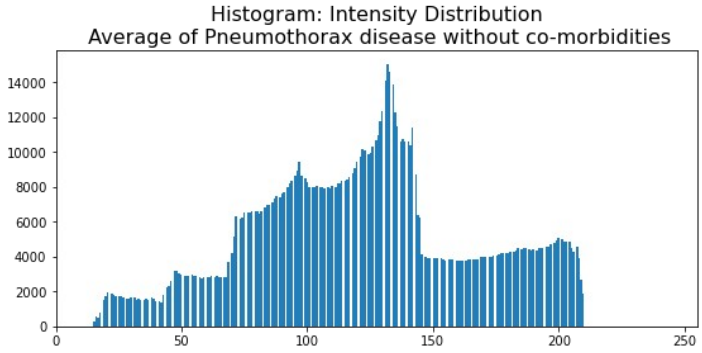


min: 18 | max: 216 | median: 125.00 | mean: 125.19 | avg: 125.19 | std: 45.20

percentiles: 18.0 - 69.0 - 87.0 - 101.0 - 113.0 - 125.0 - 133.0 - 142.0 - 167.0 - 194.0 - 216.0

Pneumothorax





min: 15 | max: 210 | median: 122.00 | mean: 121.60 | avg: 121.60 | std: 43.93

percentiles: 15.0 - 65.0 - 84.0 - 97.0 - 110.0 - 122.0 - 132.0 - 140.0 - 161.0 - 187.0 - 210.0

Description of Training Dataset:

The ratio of positive and negative cases is 50:50. Since the age of target population is between 21 and 90, data from every patient who does not fit in this range is removed. The datasets contain only chest X-ray images.

Disease	Total cases
Atelectasis	2356
Cardiomegaly	1050
Consolidation	548
Edema	104
Effusion	1764
Emphysema	372
Fibrosis	810
Hernia	118
Infiltration	4814
Mass	1746
Nodule	2318
Pleural Thickening	1048
Pneumonia	208
Pneumothorax	354

Description of Train Dataset:

The ratio of positive and negative cases is 20:80, that refers to the ratio of real life cases.

Disease	Total cases
Atelectasis	1635
Cardiomegaly	730
Consolidation	380
Edema	70
Effusion	1225
Emphysema	260
Fibrosis	565
Hernia	85
Infiltration	3345
Mass	1210
Nodule	1610
Pleural Thickening	730
Pneumonia	145
Pneumothorax	245

Description of Validation Dataset:

The ratio of positive and negative cases is 20:80, that refers to the ratio of real life cases.

Disease	Total cases
Atelectasis	655
Cardiomegaly	290
Consolidation	155
Edema	30
Effusion	490
Emphysema	105
Fibrosis	225
Hernia	35
Infiltration	1340
Mass	485
Nodule	645
Pleural Thickening	290
Pneumonia	60
Pneumothorax	100

5. Ground Truth

The training and validation dataset comes from a publicly available dataset from Kaggle known as NIH Chest X-ray Dataset made by National Institutes of Health or from AcademicTorrents. Dataset contains 14 different disease labels and 1 label for negative cases. Images can be labeled with multiple labels if the patient suffers from different diseases.

The creators use Natural Language Processing to text-mine disease classifications from the associated radiological reports. The labels are expected to be >90% accurate. This is a quite accurate dataset, however the biggest limitation is, that the original associated radiological reports are not publicly available. That's why we cannot make a cross-validation process or get a second opinion about data.

Paper are available here:
http://openaccess.thecvf.com/content_cvpr_2017/papers/Wang_ChestX-ray8_Hospital-Scale_Chest_CVPR_2017_paper.pdf

There is a big disadvantage of the dataset, that can be eliminated with pre-processing. It contains some bad records about patients' age, such as: '148, 149, 150, 151, 152, 153, 154, 155, 411, 412, 413 and 414'.

The greatest advantage of dataset is the original associated radiological reports were made by radiologists, according to the paper.

6. FDA Validation Plan

Patient Population Description for FDA Validation Dataset:

For the FDA Validation Dataset, the data shall be comes with these population parameters:

- sex: male or female
- age: between 21 and 90 (inclusive)

Ground Truth Acquisition Methodology:

Multiple approaches exist to make a golden standard study for obtain ground truth.

- CT scan
- Sputum test
- Pleural fluid culture test

CT scan is not the best since the patient had already taken radiation from an X-ray machine. The next two tests are expensive and their reliability is quite low without any prior knowledge about state of lung or the patient's general health condition. In this scenario there is enough a silver standard for obtaining ground truth. This can be as the average of three practicing radiologists. The original paper mentioned this method as well. Since the purpose of the algorithm is assisting the radiologists and not replacing them, this approach is necessary and enough to get ground truth.

Algorithm Performance Standard:

The base value of the performance standard is the F1 score. There are a lot of models for pneumonia detection around the world. The goal of development is to reach a higher F1 score, than human radiologists exceed. According to the CheXNet paper, the average of human radiologists is 0.387.

Paper available here: https://arxiv.org/abs/1711.05225

There is adviced to set confidence interval as 95%. It can be measure the performance better, than using a single value without any confidence score.