Identifying Opacities in Lung X-Rays for Covid-19 Detection

A Multi-Classification and Bounding Box Regression Problem

Andrea Pertoldi

Julian Robinson-Tait

Applied Machine Learning Course, 2021
Niels Bohr Institute

Both team members contributed equally to the project

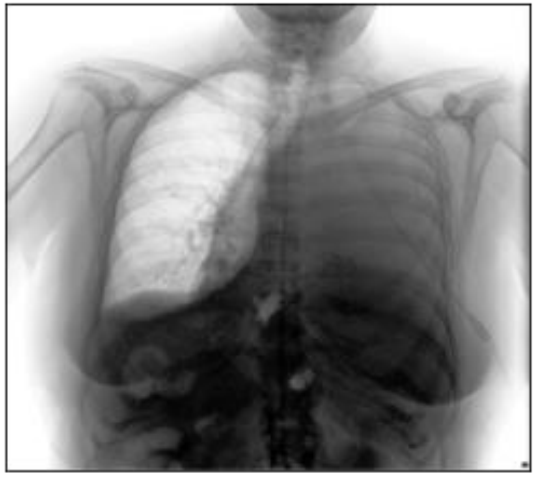
Kaggle Challenge: Covid-19 detection

a.

b.

c.

d.



Labels:

1. No Pneumonia
2. Typical Pneumonia
3. Atypical Pneumonia (Covid-19)
4. Indeterminate

Hint:

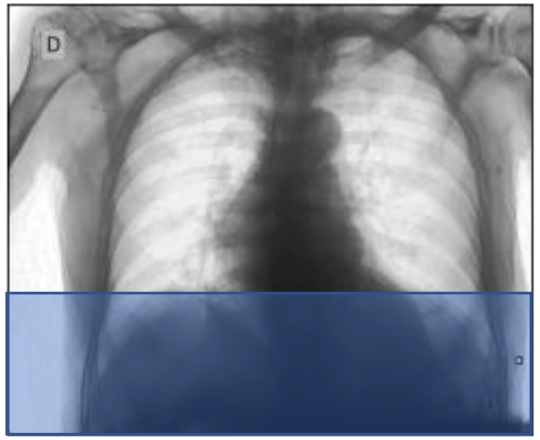
Pneumonia → Fluid build up in the alveoli

→ Scatter X-rays more

→ Darker regions

Problem Introduction: Classification AND Regression

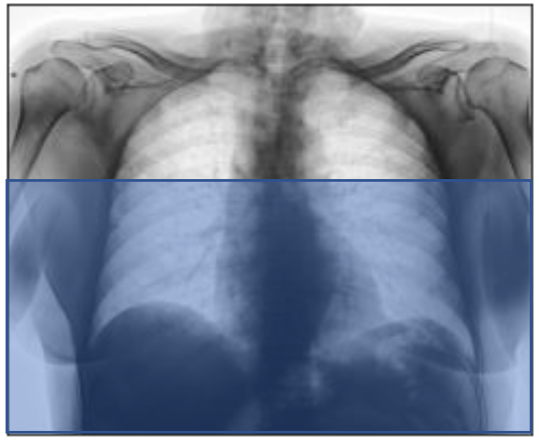
a.



No Pneumonia

1736 (27%)

b.



Typical

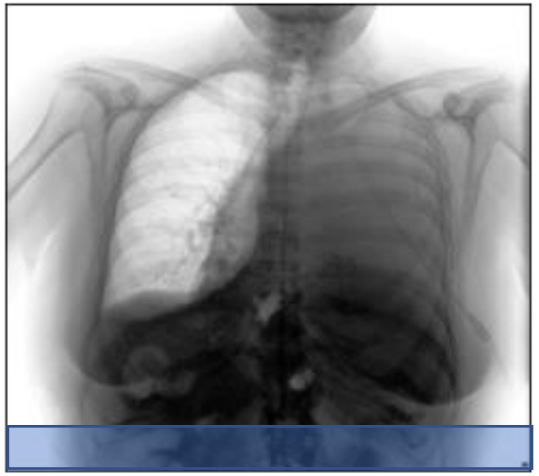
3007 (48%)

c.

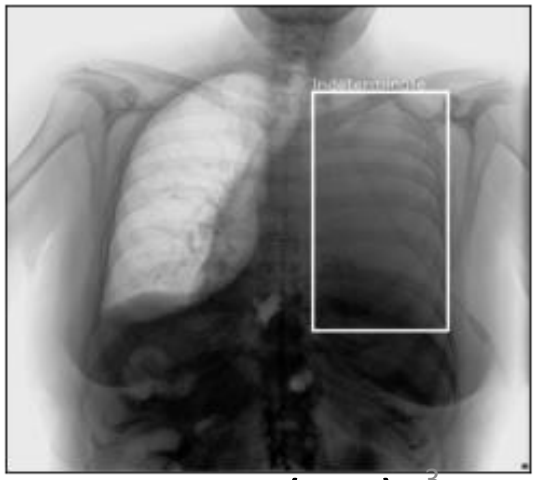
Atypical

1108 (17%)

d.



Indeterminate

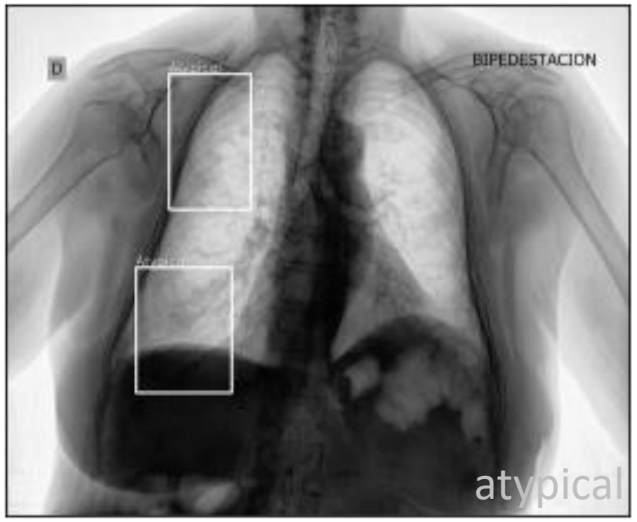


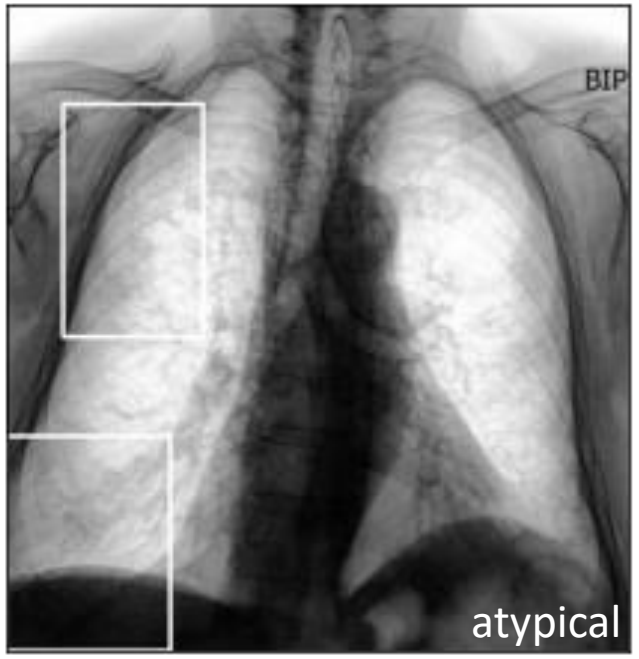
483 (8%)

Data pre-processing

Train dataset: 6334 medical DICOM images (120 GB):

Custom pixel range	→	Pixel range [0,255]
Different detectors	→	Invert images
Not always centred	→	Cut Images
Low contrast	→	Equalize histograms

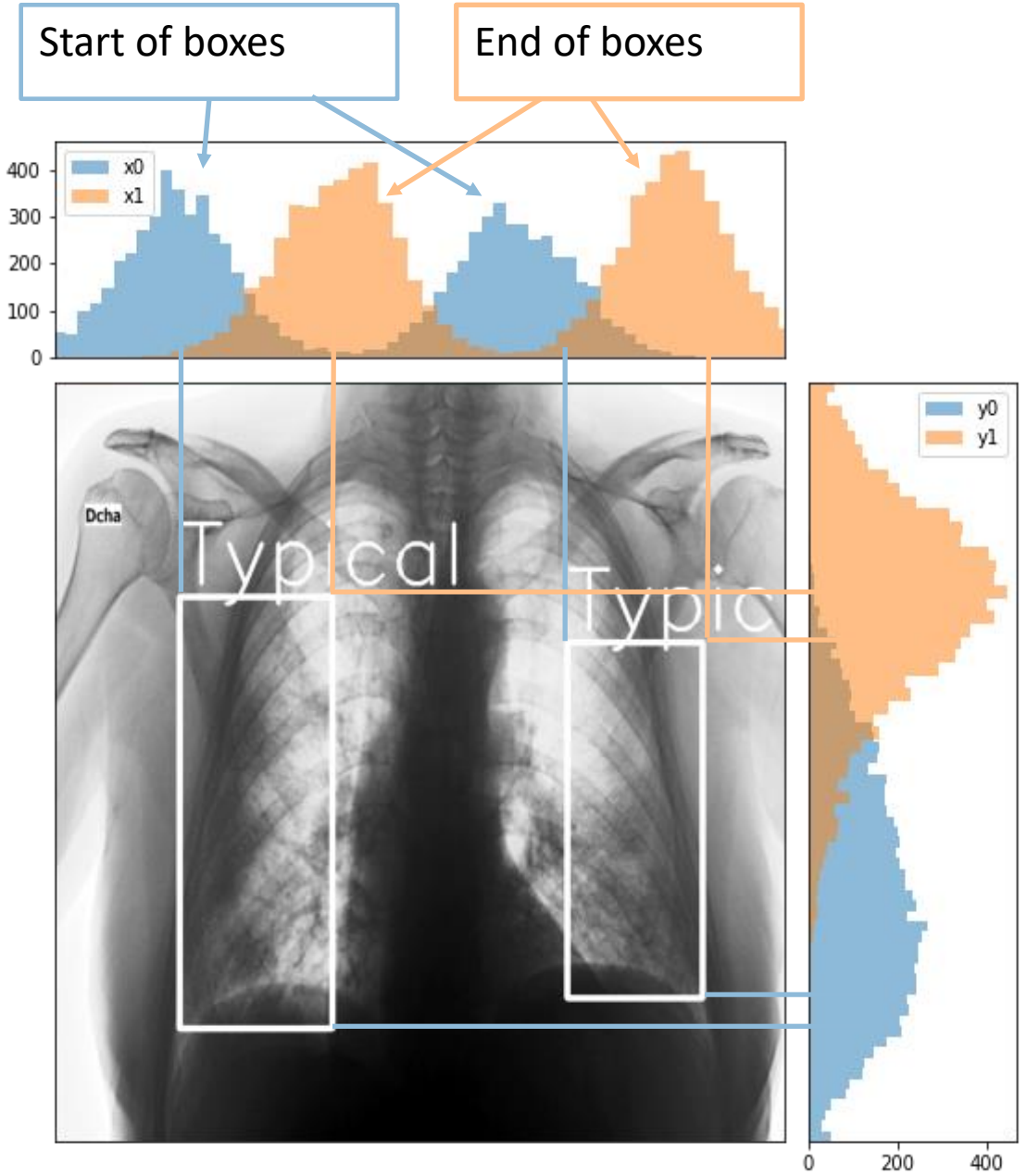




Data pre-processing

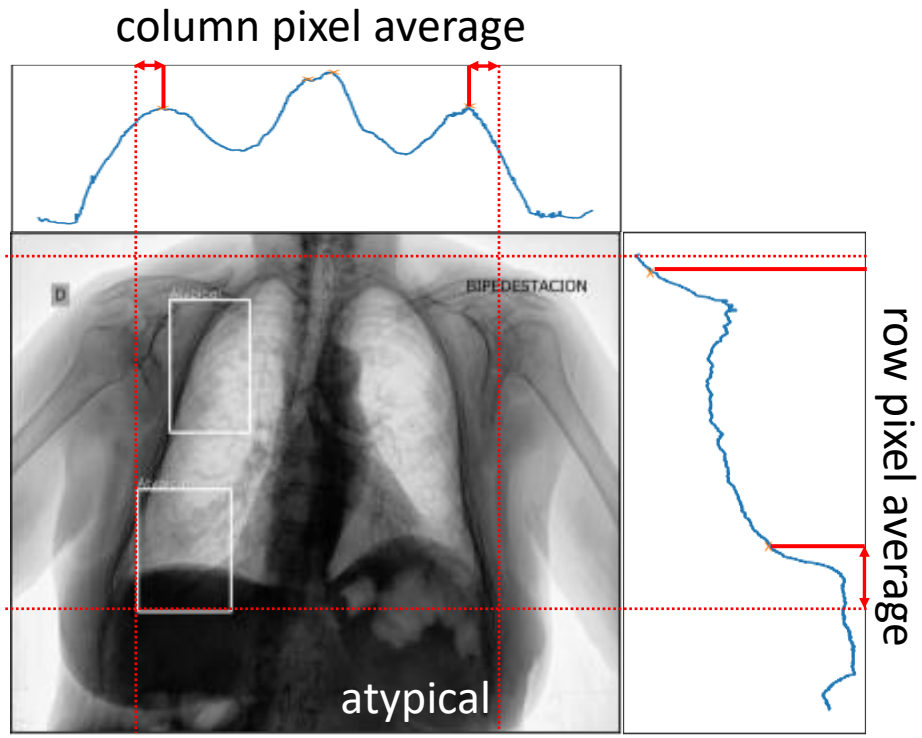
Distribution of bounding boxes

Motivate cutting images around the lungs



Data pre-processing

Identify lungs by the mean pixel value in the columns and rows



Data size reduced by almost 40 %

→

Model trains in 5h instead of 9!

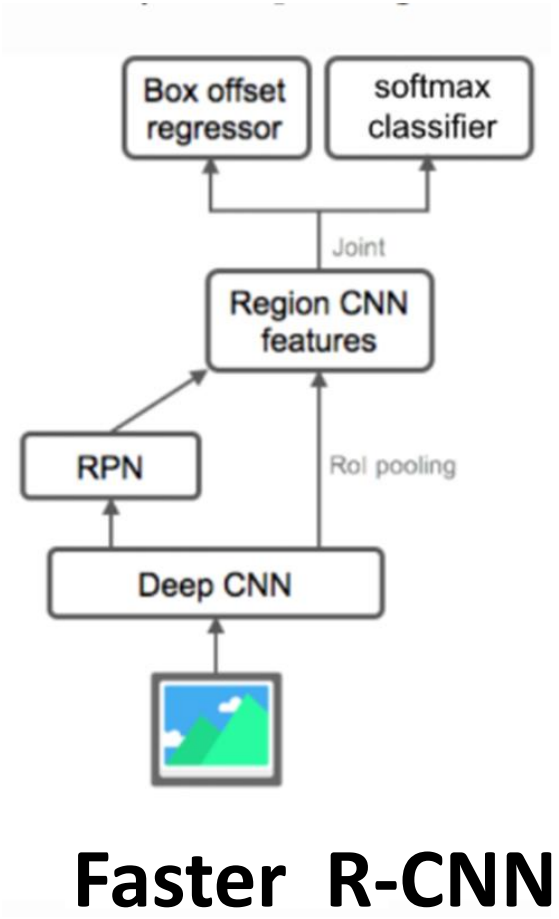
Model: `fasterrcnn_resnet50_fpn`

Parameters:

- # of Classes = 5 (4 + background)
- Optimizer: stochastic gradient descent
  - best hyperparameters: `lr=0.0001, momentum=0.9, weight_decay=0.0005`
- Ran on Kaggle GPUs

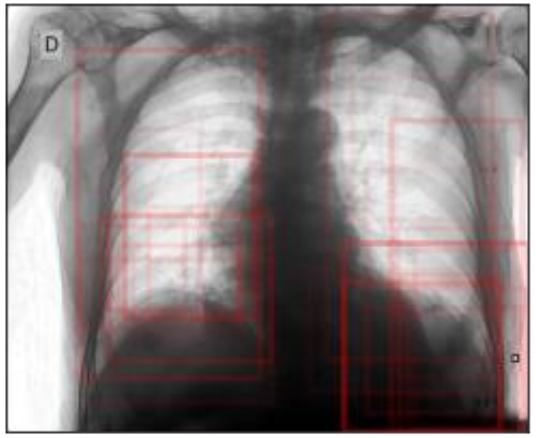
- Removed weight and biases
- Trained on specific problem!

- Pretrained (transfer learning)
- Keep general feature identification

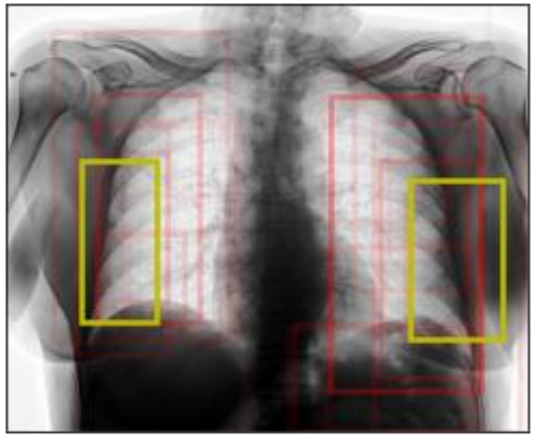


Model Evaluation

No Pneumonia



Typical



Atypical

Indeterminate

Regression:

Bounding boxes

1. `[x,y]` coordinates
2. Confidence
3. Yellow = ground truth

   Red = prediction

Classification:

Labels:

1. No Pneumonia
2. Typical Pneumonia
3. Atypical Pneumonia (Covid-19)
4. Indeterminate

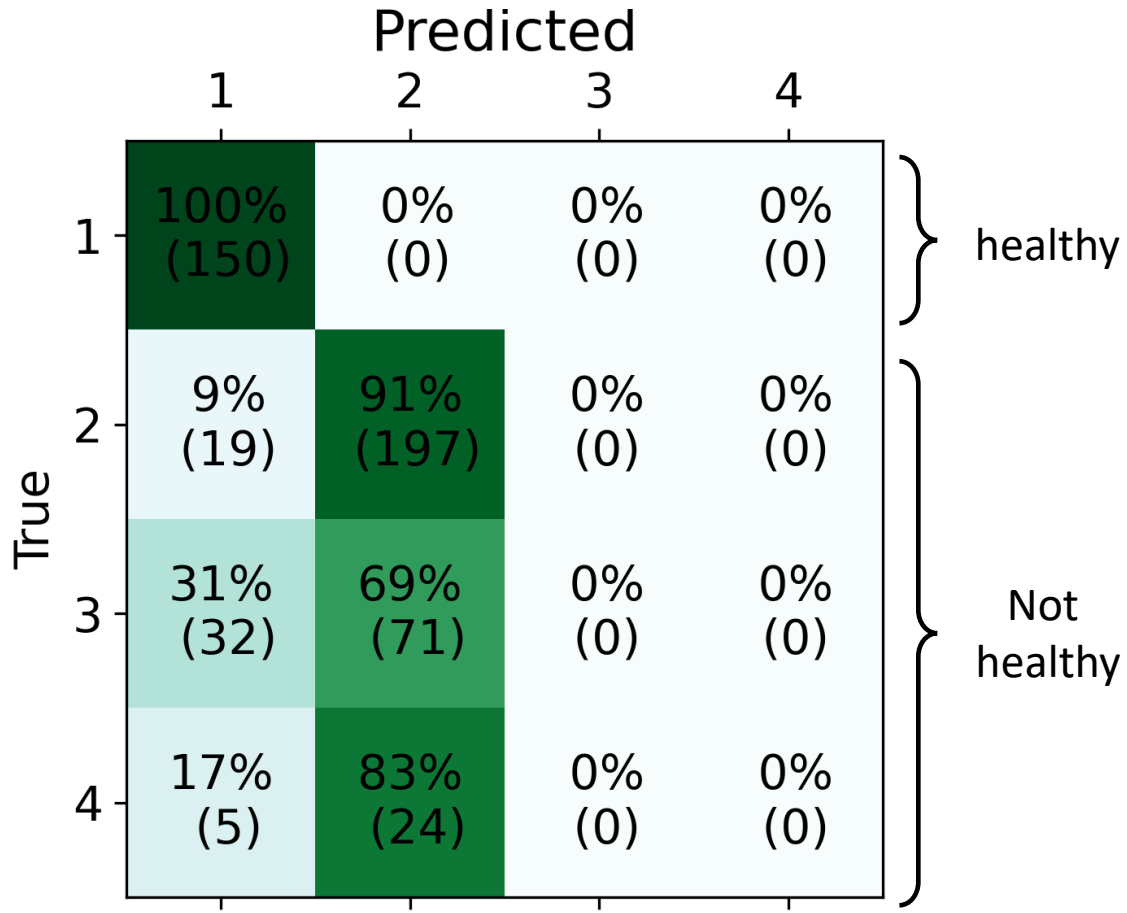
} Healthy

} Not healthy

Model evaluation – classification

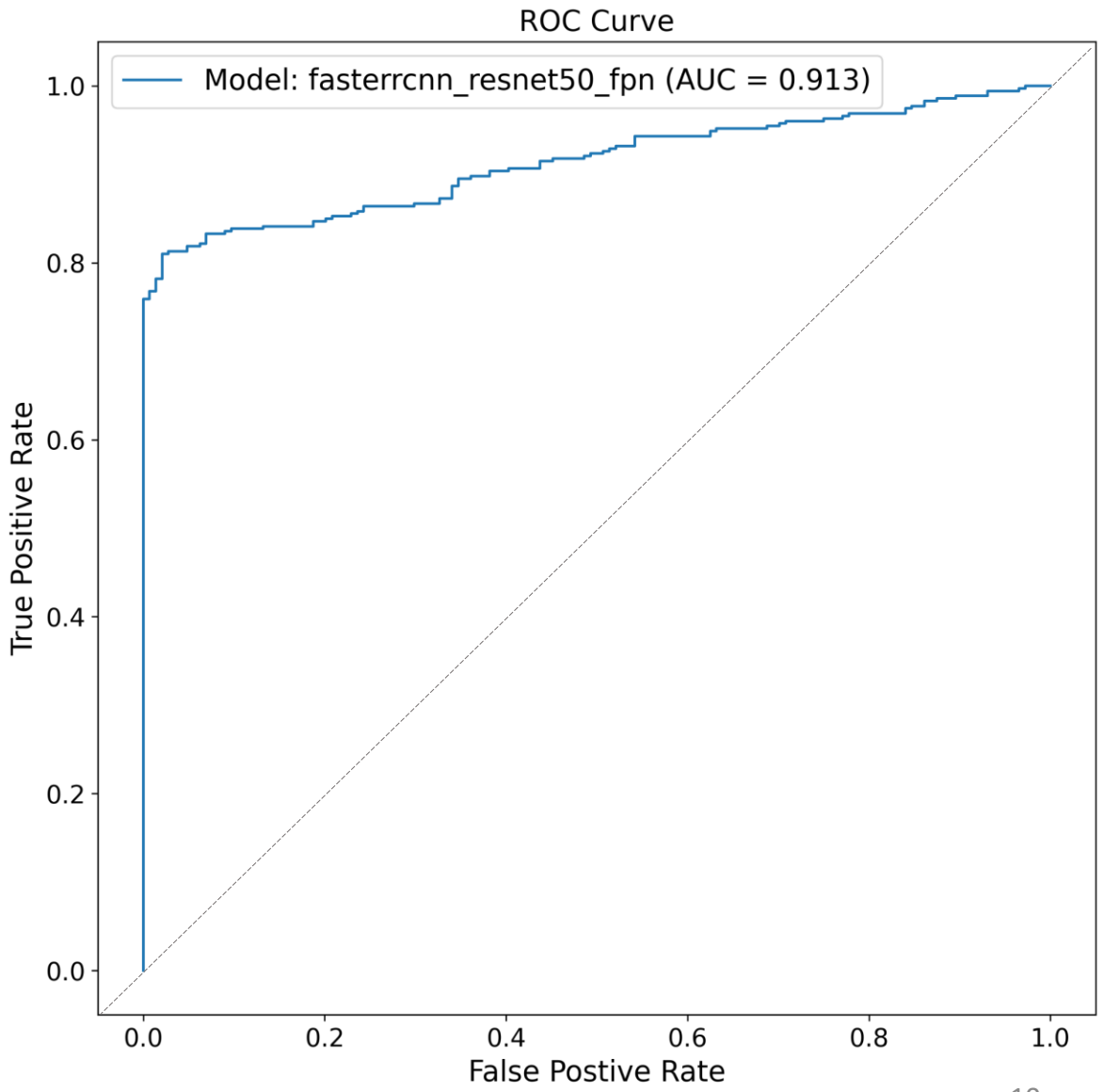
- 20% of the training data reserved for validation
- Label given by top confidence box per image
- Successfully differentiates between healthy and not healthy
- Model does not recognize labels 3 & 4
  - Low number of samples

Confusion matrix



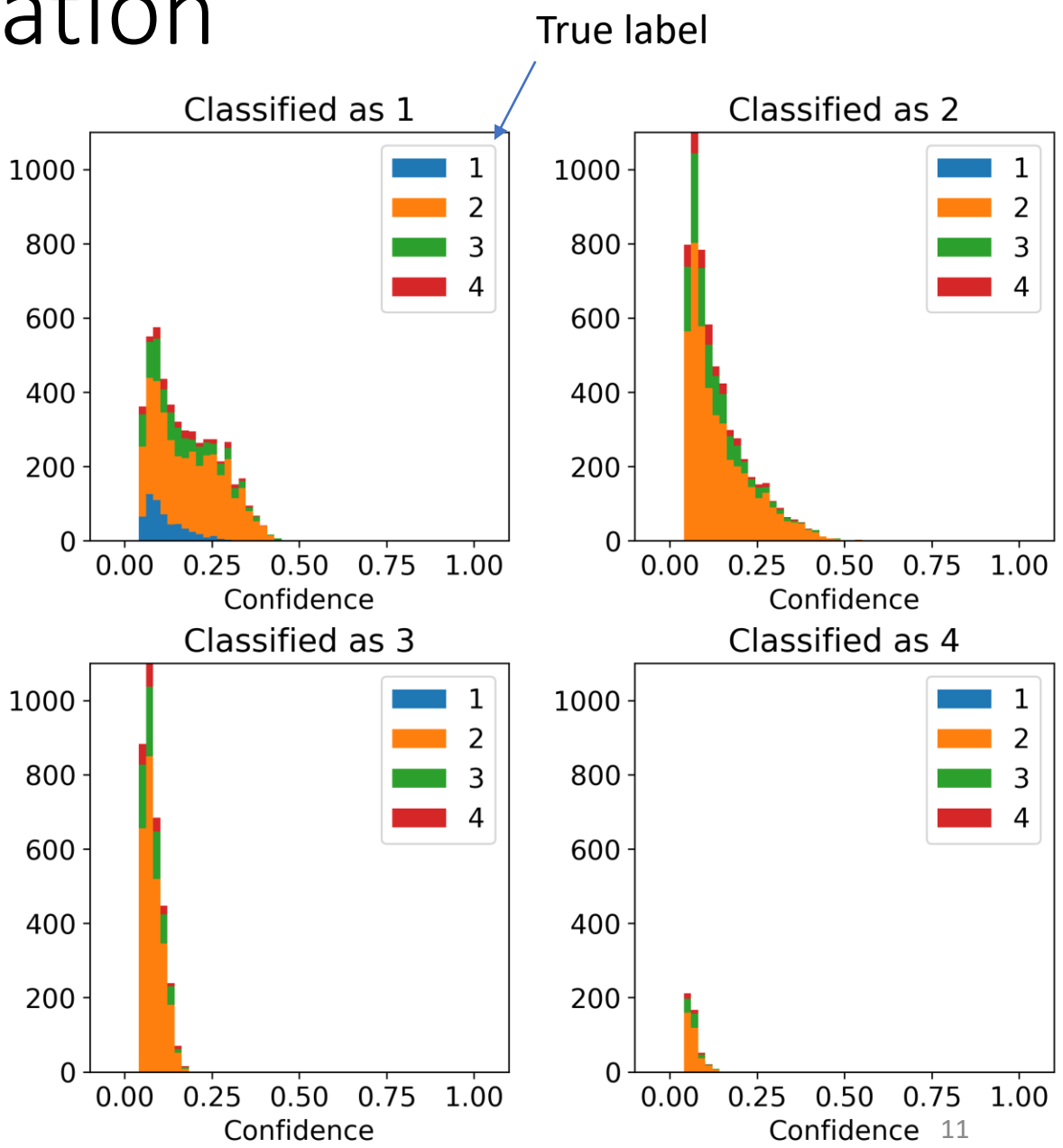
Model evaluation- classification

- Model is good at classifying healthy / not healthy images
- True positive: correctly identified healthy lungs
- Labels 2, 3, 4 grouped together

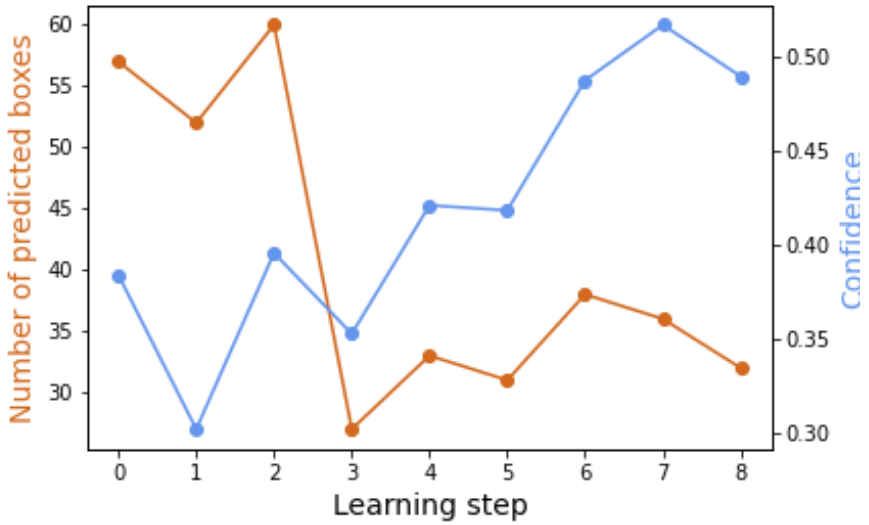


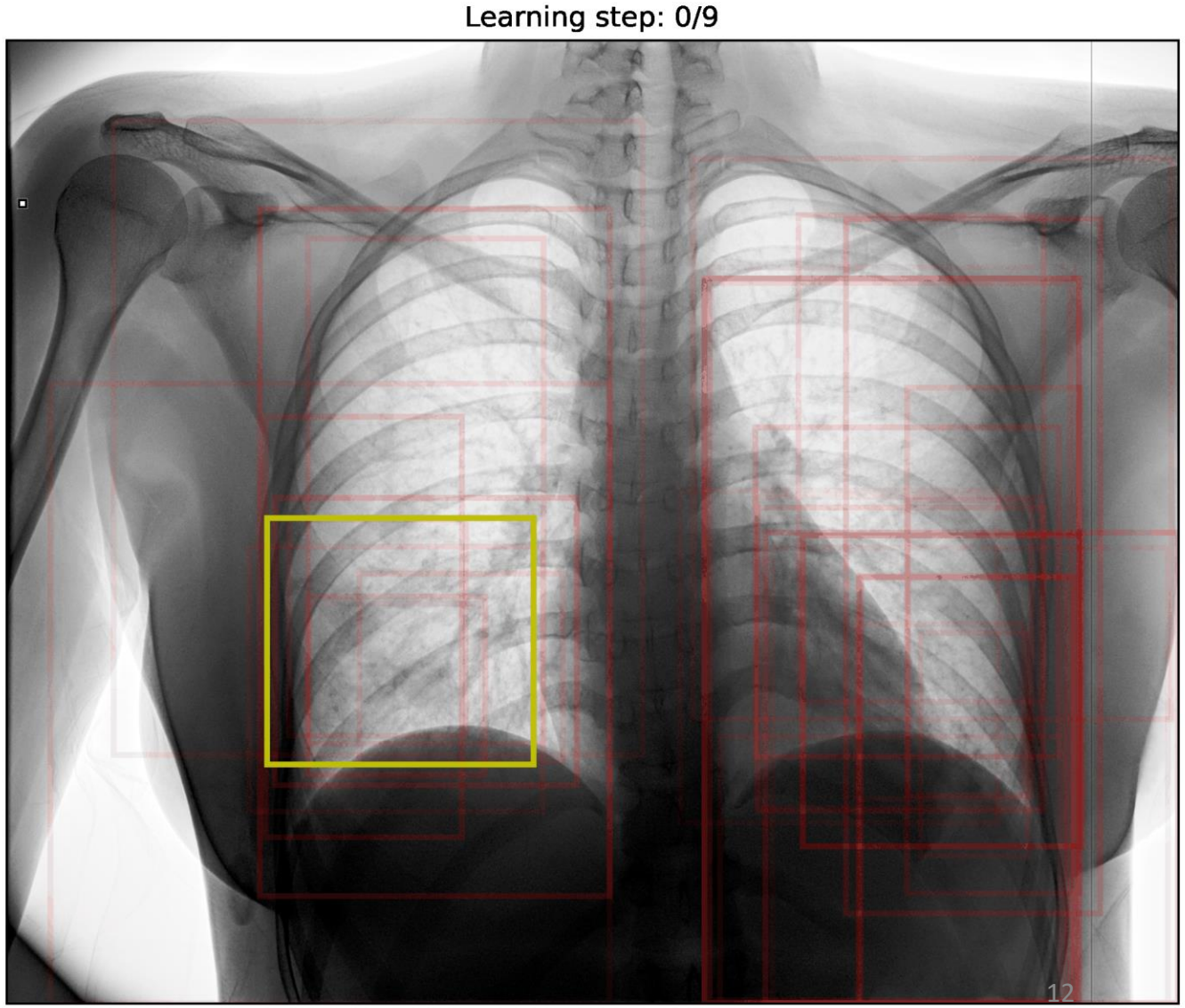
Model evaluation - classification

- Top 5 boxes of highest confidence per figure.
- Quantification of the confidence of the boxes
- Cannot use the confidence to classify images

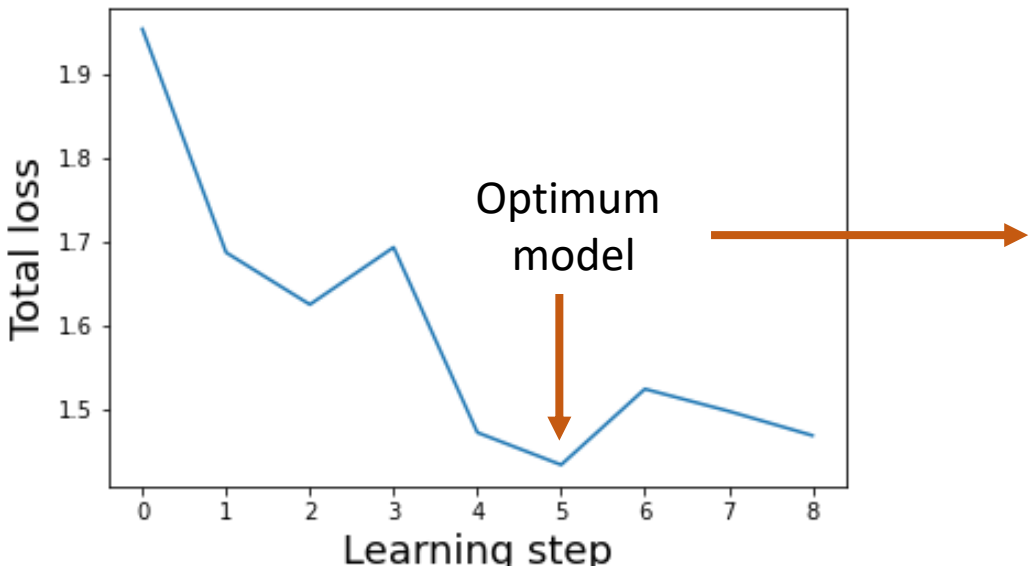


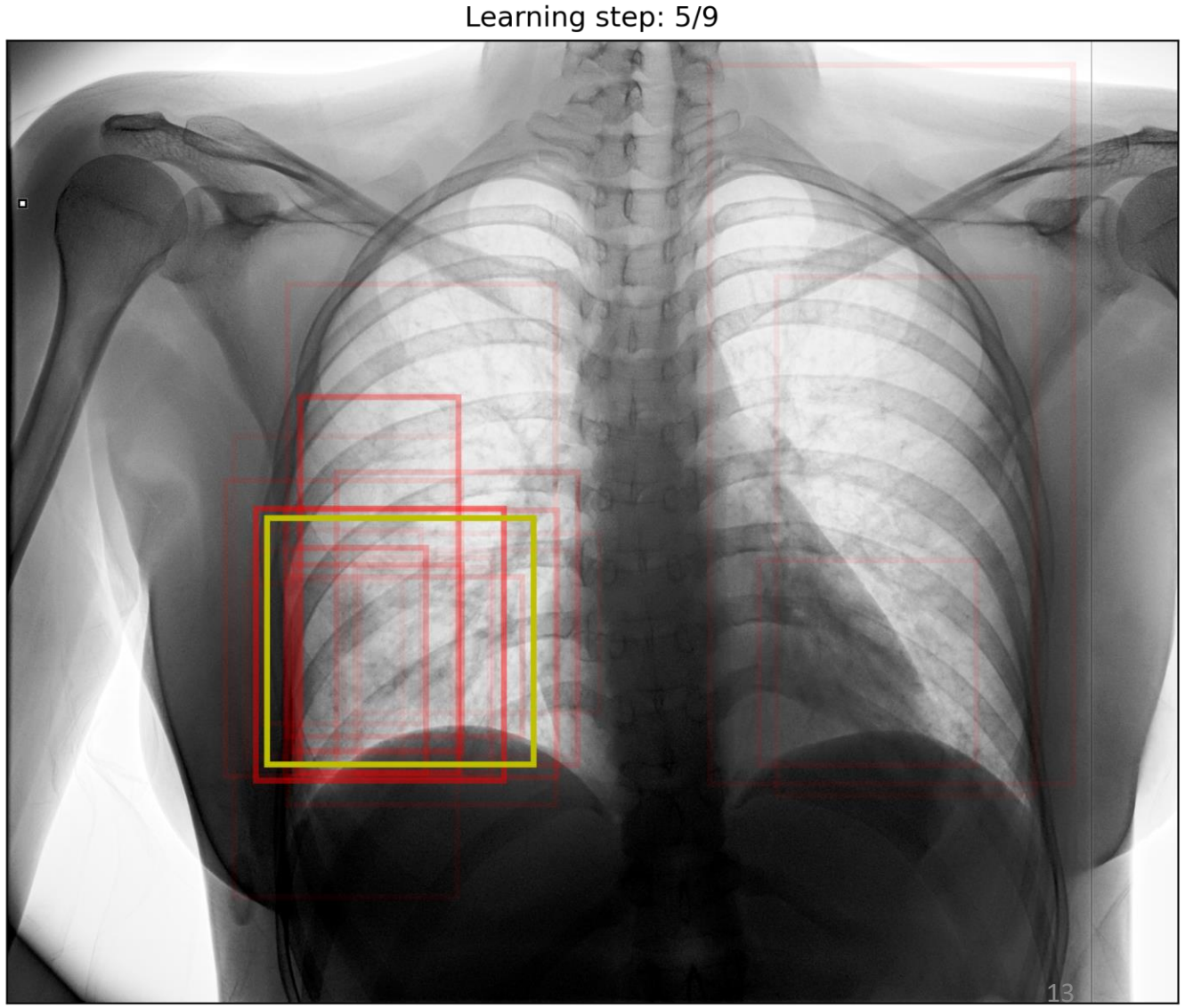
Model evaluation – regression





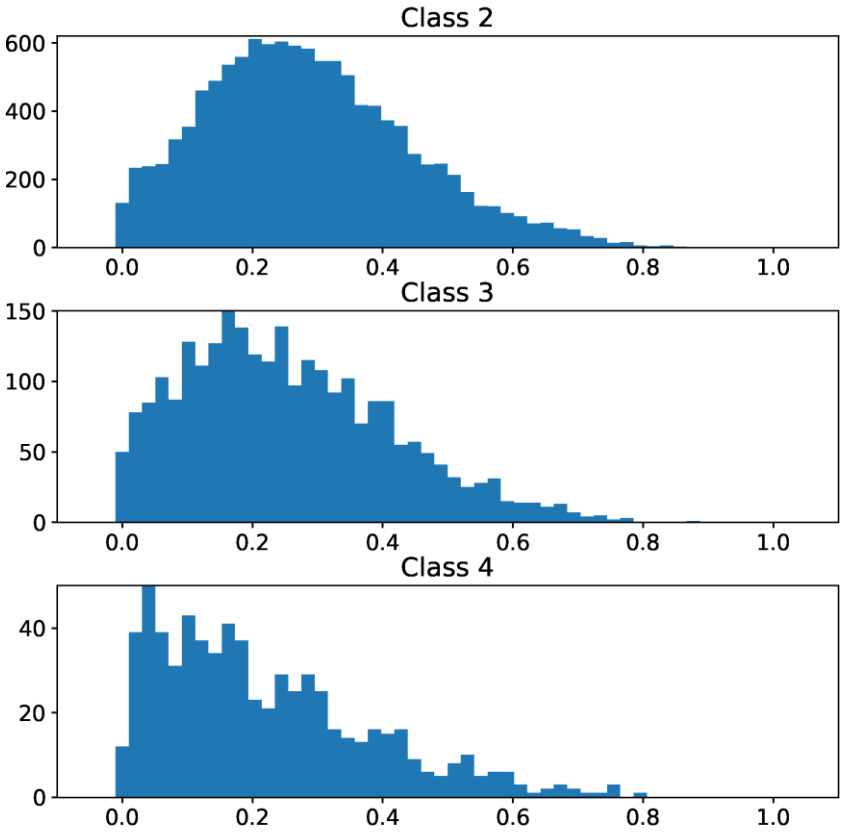
Model evaluation – regression



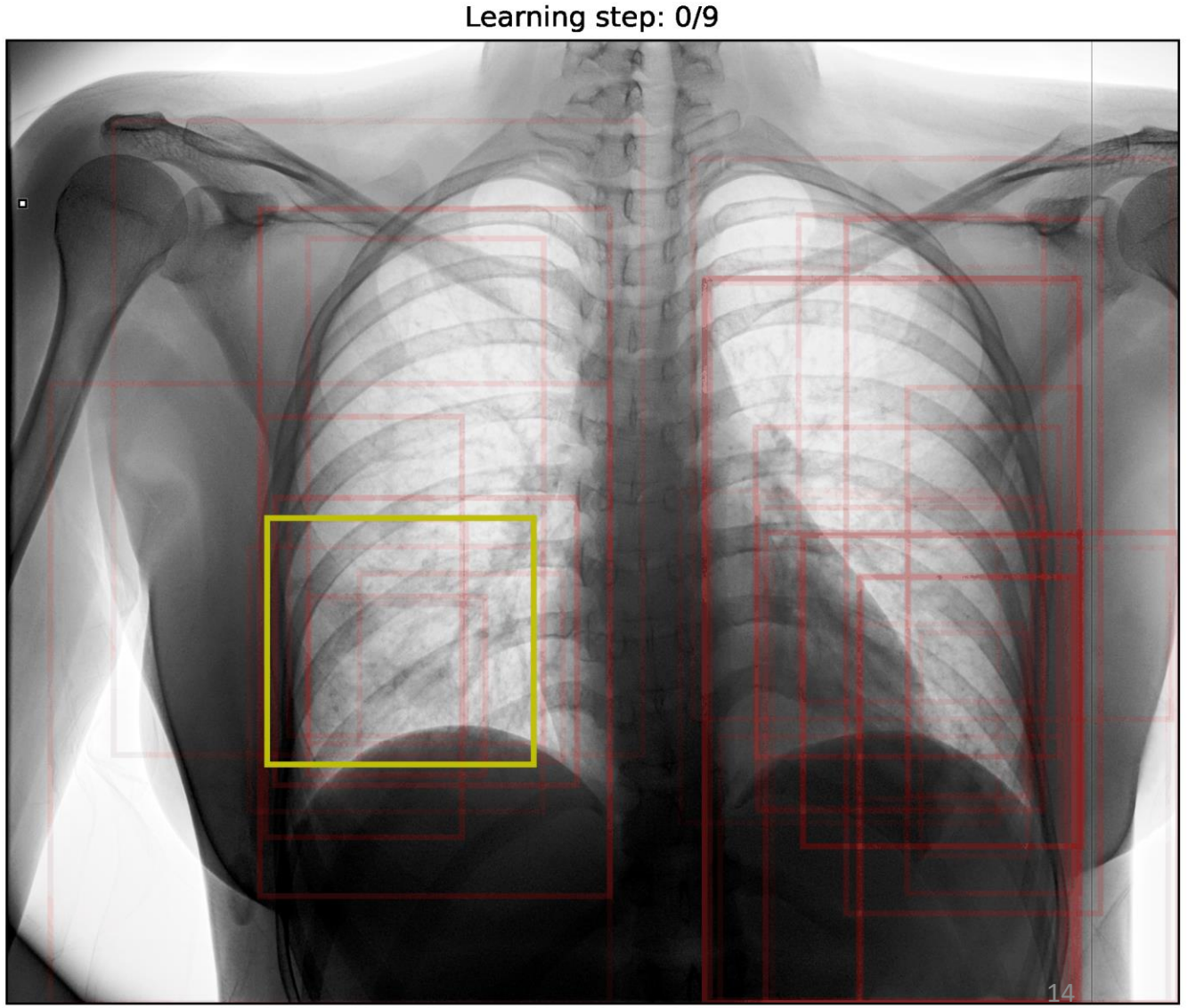


Model evaluation – regression

Evolution of IoUs across all validation dataset:



Intersection over Union (IoU)



Summary

- Implementation of R-CNN model `fasterrcnn_resnet50_fpn`
- Successful classification of healthy/not healthy lungs X-rays
- Somewhat meaningful boxing of sick areas
- Built a number of tools to characterize the model
- 150 h of Tesla P100 GPU time = 7.88 kg CO_2 eq. (= 32 km by car)

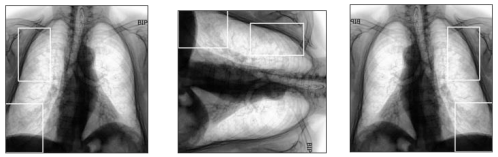
THANK YOU FOR YOUR ATTENTION

Appendix

Training the model

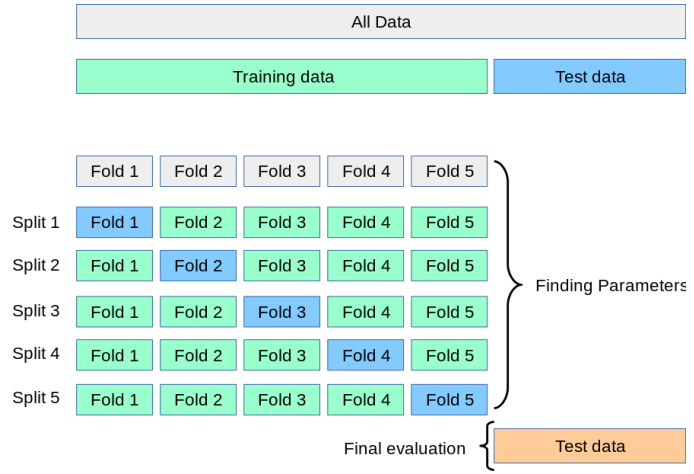
Imbalanced dataset:

→ Data augmentation (albumentation)



→ Use `groupKfold` splitting to feed the data

Do `n epochs` in each fold




```
losses =
sum([  1 * loss_objectness,
      10 * loss_classifier,
      10 * loss_rpn_box_reg,
     0.5 * loss_box_reg**2 ])
```

→