Exploring LRP and Grad-CAM visualization to interpret multi-label-multi-class pathology prediction using chest radiography

Mahbub Ul Alam[1], Jón Rúnar Baldvinsson[2], Yuxia Wang[3]

[1]Department of Computer and Systems Sciences, Stockholm University, Stockholm, Sweden

[2]Skatturinn (Iceland Revenue and Customs), Reykjavík, Iceland

[3]Qamcom Research and Technology, Stockholm, Sweden

mahbub@dsv.su.se

jon.r.baldvinsson@gmail.com

yuxia.wang@qamcom.se

Stockholm University

The importance of interpretability

- Interpretability as a **gateway** between **machine learning and society**
- Making complex models **acceptable for certain applications**
- **Retaining human decision** in order to **assign responsibility**
- Ensuring the **“right to explanation”**
- **Optimizing** models / architectures
- Detecting **flaws / biases** in the data
- Gaining **new insights** about the problem
- Making sure that machine learning models behave **“correctly”**

Interpretability: LRP Algorithm

Layer-wise Relevance Propagation (LRP)
(Bach et al. 2015)

input $\boldsymbol{x}$

Classifier

Rooster

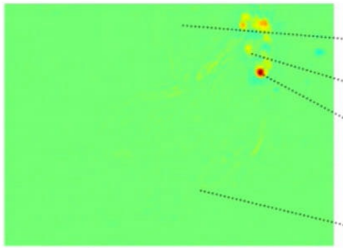
prediction $f(\boldsymbol{x})$

Explain prediction

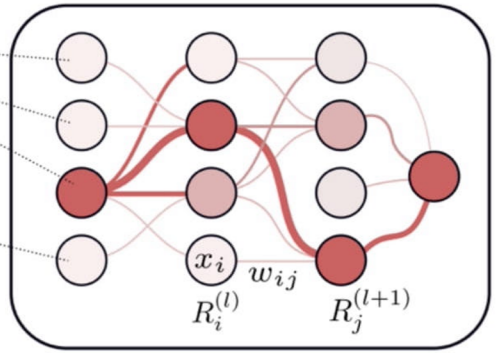
(how much each pixel contributes to prediction)

Idea: Decompose function

$$\sum_i R_i = f(x)$$



heatmap



"every neuron gets it's share of relevance depending on activation and strength of connection."

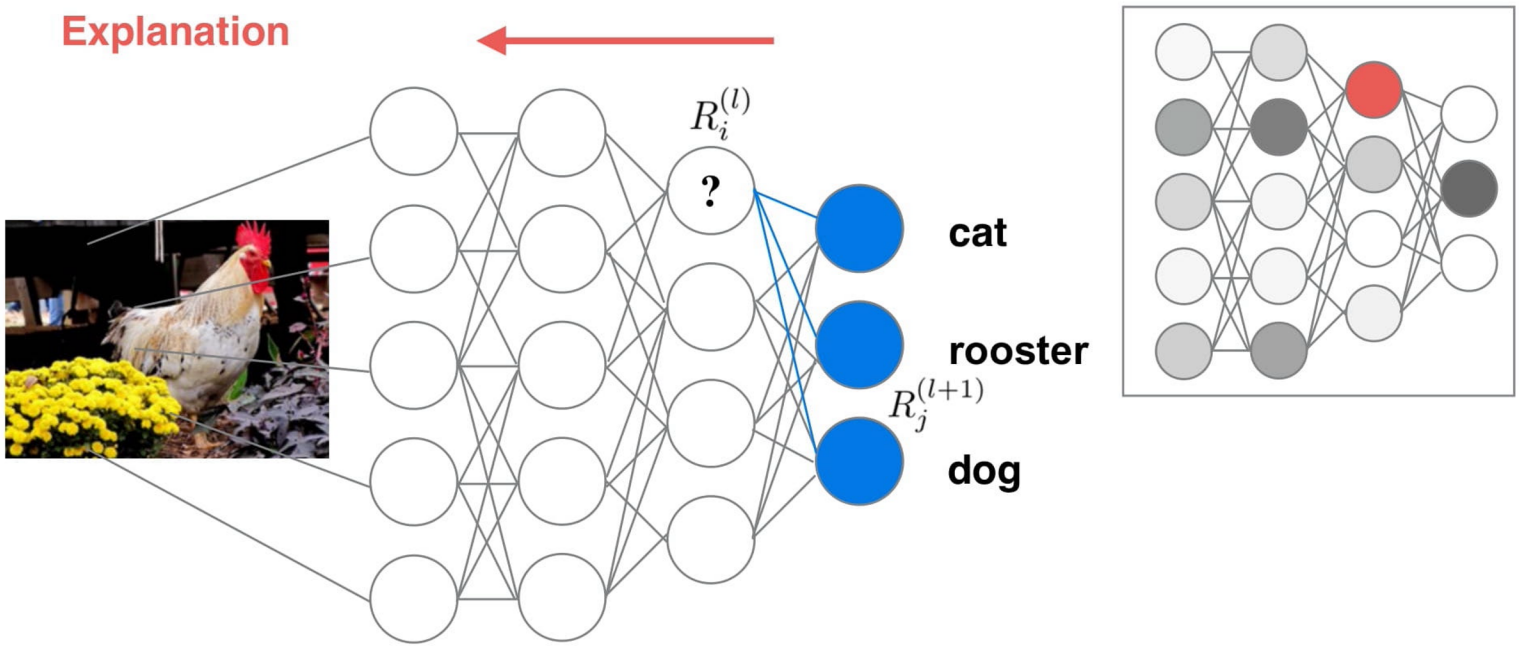
redistribute $f(\boldsymbol{x})$

Theoretical interpretation
Deep Taylor Decomposition
(Montavon et al., 2017)

[1]

Stockholm University

Interpretability: LRP Algorithm



Simple LRP rule (Bach et al. 2015)

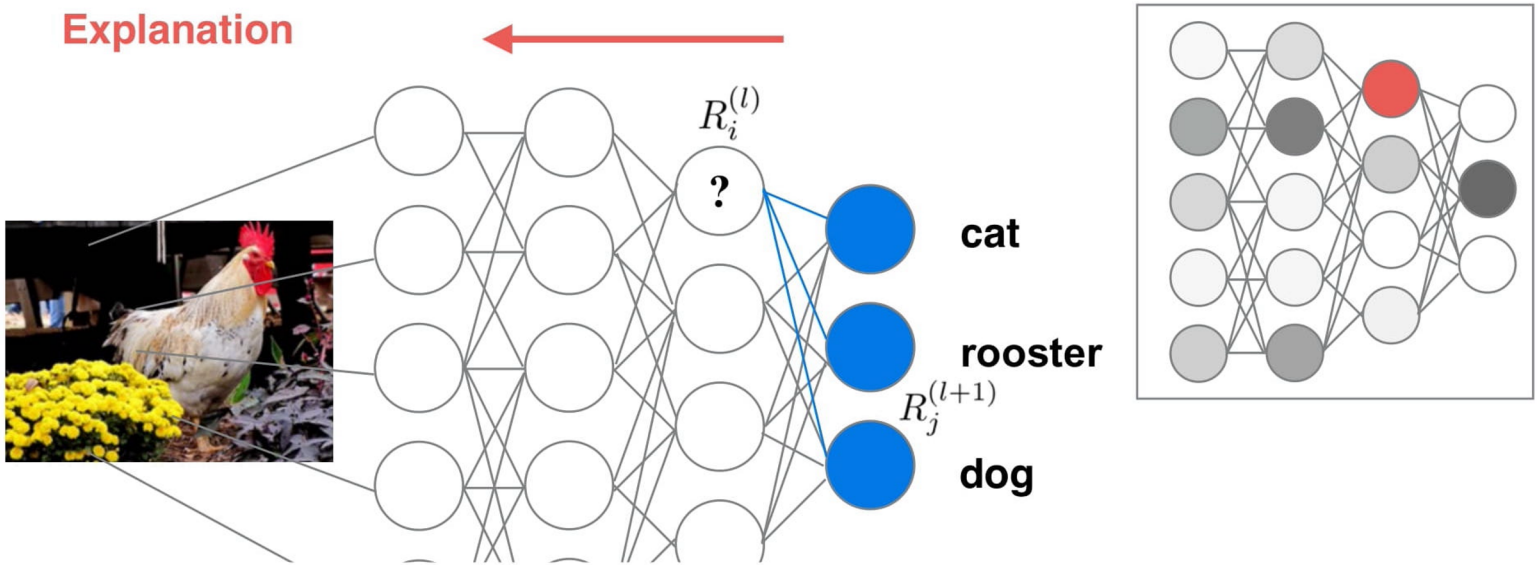
$$R_i^{(l)} = \sum_j \frac{x_i \cdot w_{ij}}{\sum_{i'} x_{i'} \cdot w_{i'j}} R_j^{(l+1)}$$

Every neuron gets its "share" of the redistributed relevance

[1]

Stockholm University

Interpretability: LRP Algorithm



special case
$\alpha = 1, \beta = 0$

Equivalent to redistribution rule proposed in Excitation Backprop (Zhang et al., 2016)

Theoretical interpretation
Deep Taylor Decomposition (Montavon et al., 2017)

[1]

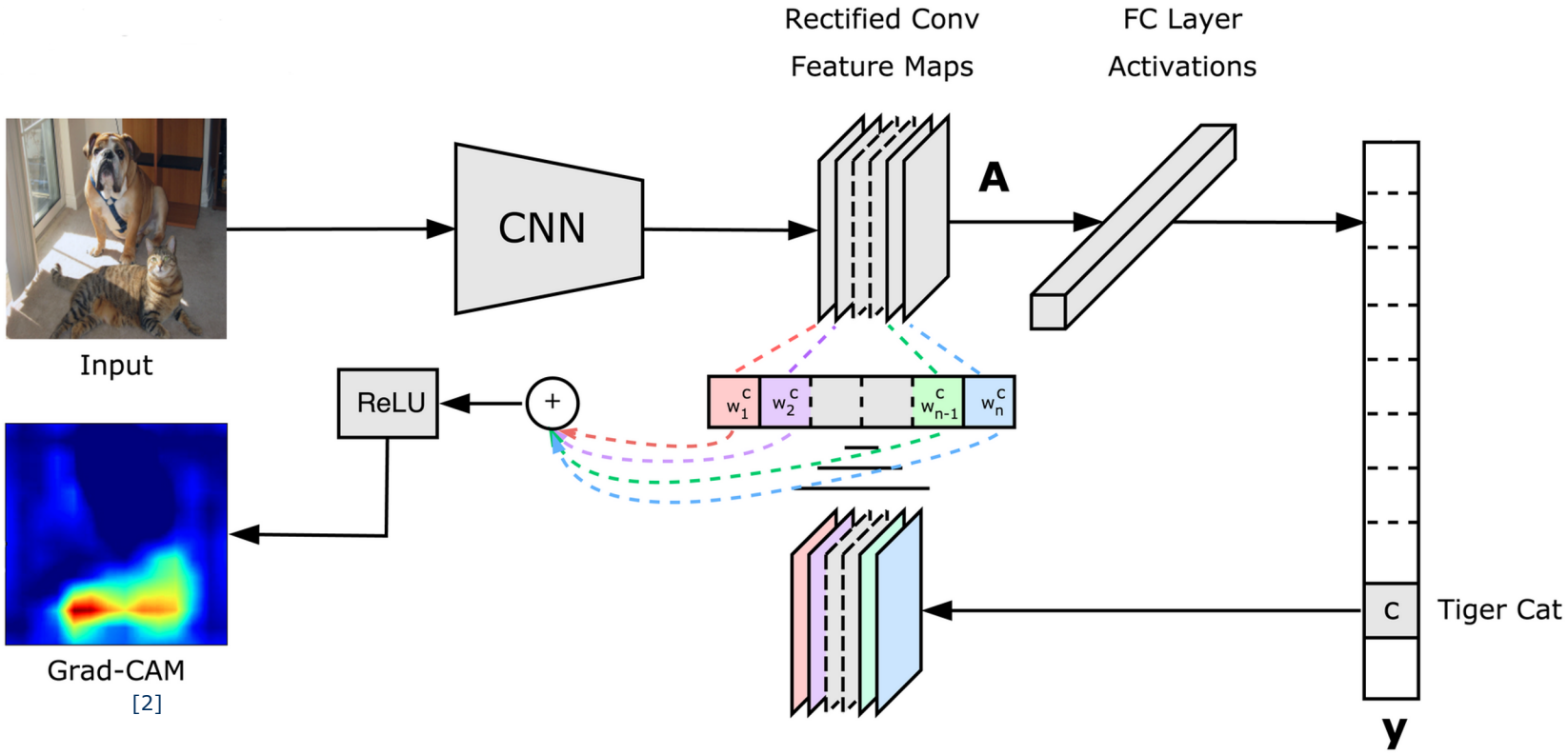
alpha-beta LRP rule (Bach et al. 2015)

$$R_i^{(l)} = \sum_j \left(\alpha \cdot \frac{(x_i \cdot w_{ij})^+}{\sum_{i'}(x_{i'} \cdot w_{i'j})^+} + \beta \cdot \frac{(x_i \cdot w_{ij})^-}{\sum_{i'}(x_{i'} \cdot w_{i'j})^-}\right) R_j^{(l+1)}$$

where $\alpha + \beta = 1$

Stockholm University

Interpretability: Grad-CAM (Gradient-weighted Class Activation Mapping) Algorithm



Stockholm University

Our Contribution

- Using both LRP and Grad-CAM for a pathology prediction task
- Using multi-label-multi-class chest radiography data
- Focusing on both the “correct” and “incorrect” predictions
- Code and Results are available: [Github.com/anondo1969/lrp-grad cam-chexpert]

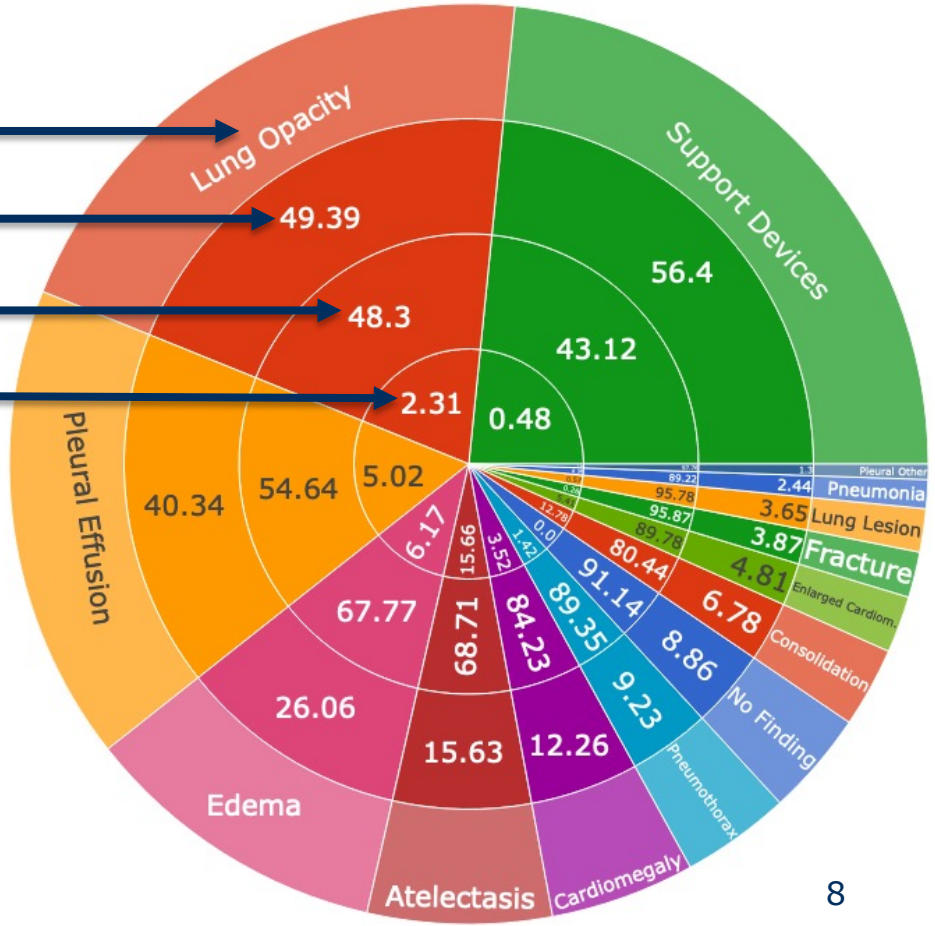
CheXpert Data Distribution [3]

Pathology/Class Name

% Positive Count

% Negative Count

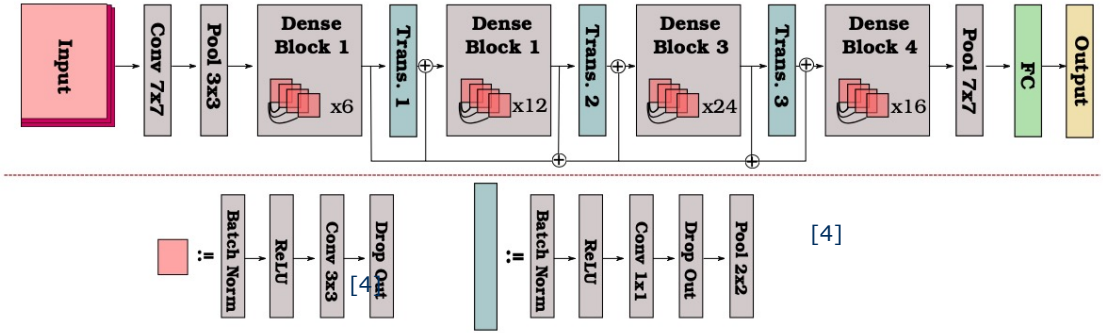
% Uncertain Count



Stockholm University

Experimental Setup

- Uncertain labels as negative
- Default train-test data split provided with the CheXpert dataset
- **DenseNet-121** architecture: transfer learning-based model training



[4]

- Pre-training weights: **ImageNet** weights
- Adam optimizer, mini-batch size of 16, total epochs = 100
- RGB heatmap representation 

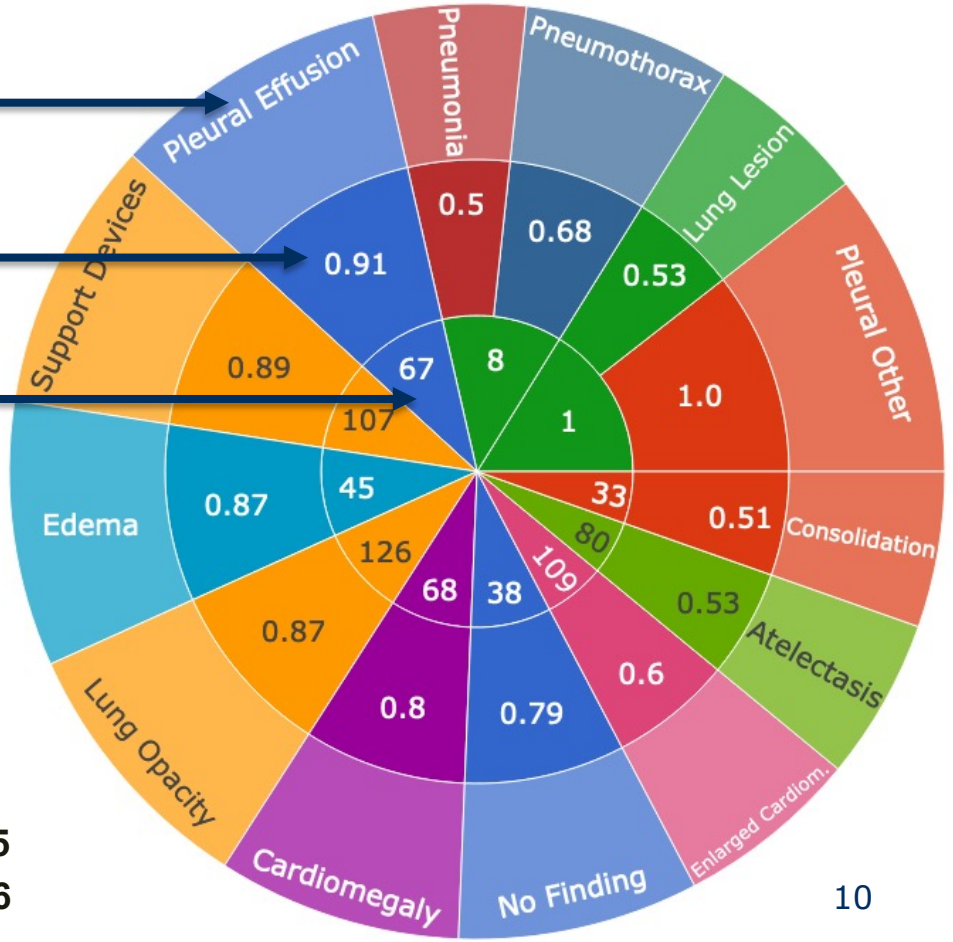
Stockholm University

Evaluation Results

Pathology/Class Name

AUROC Score

Total Positive Instances
(Fracture=0)

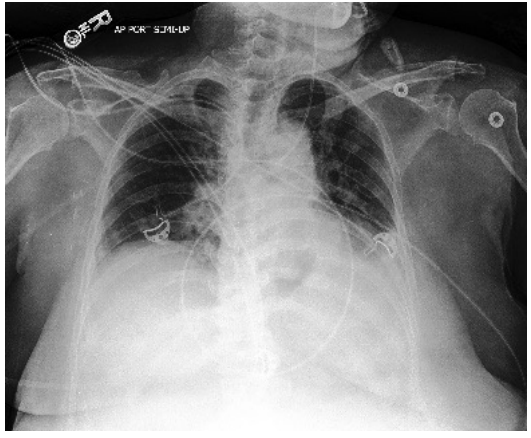


Average Weighted AUPRC Score: 0.65

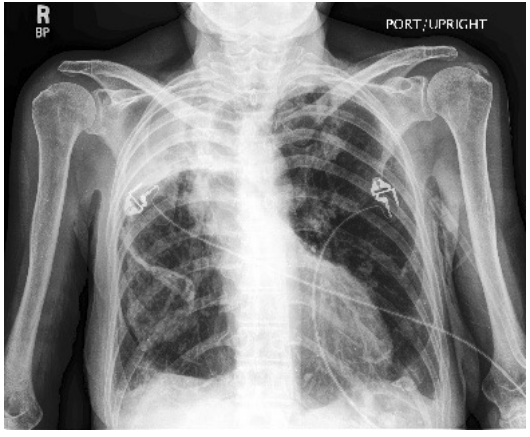
Average Weighted AUROC Score: 0.76

Stockholm University

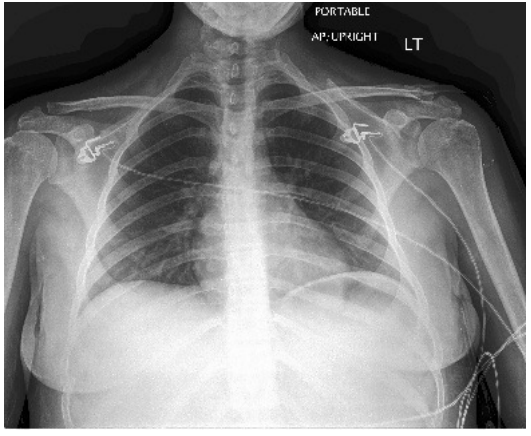
Selected Pathology/Classes for Visualization



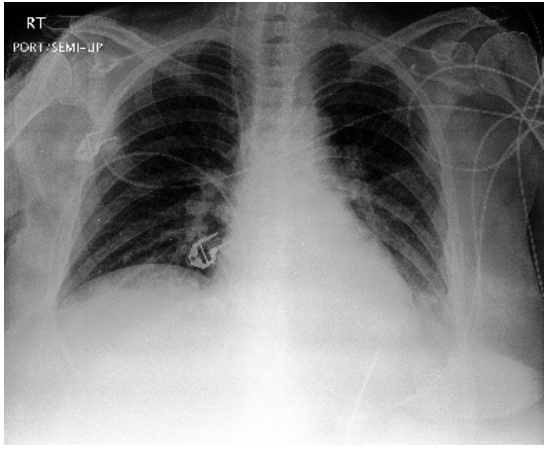
Pleural Effusion



Pneumothorax



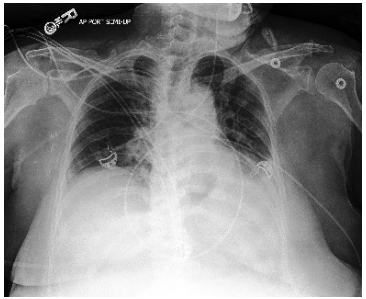
No Finding



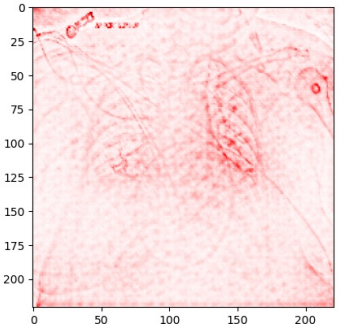
Support Devices
Lung Opacity
Pleural Effusion

Stockholm University

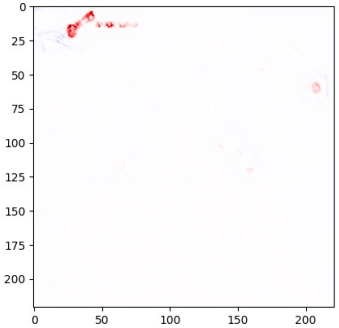
Heatmaps for Different LRP Algorithms



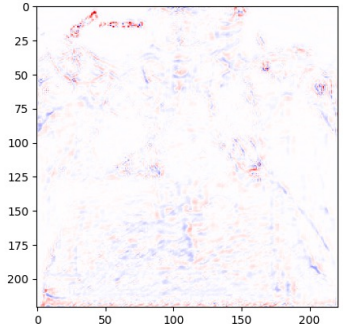
Pleural Effusion



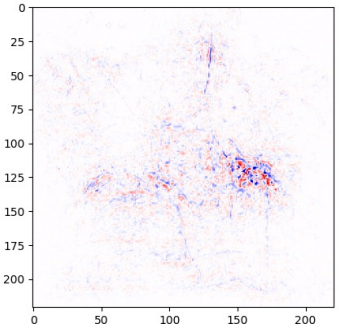
Deep Taylor decomposition



$LRP_{\alpha,\beta}$



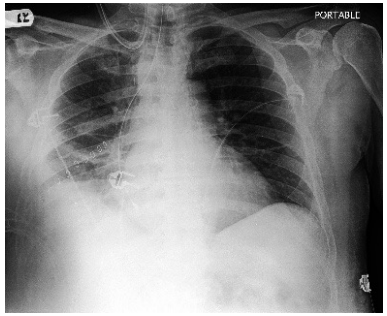
$LRP\epsilon$

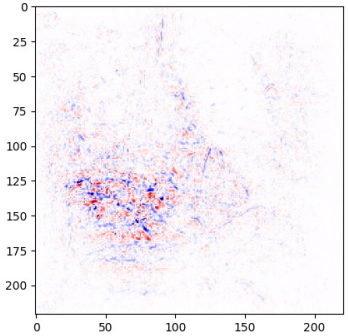


$LRPz$

Stockholm University

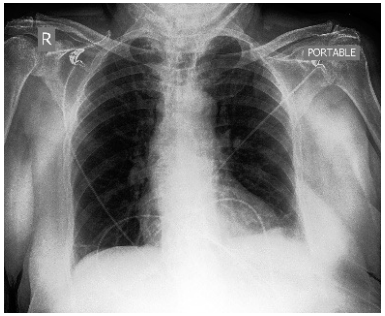
LRP Visualization

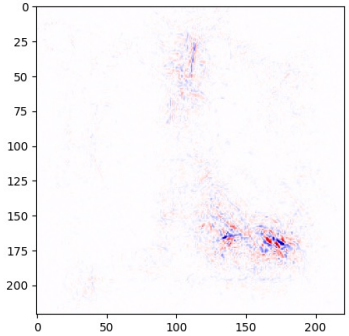




Pleural Effusion

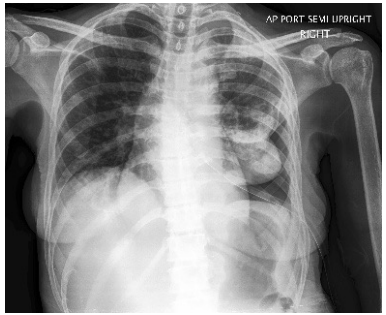
Correct

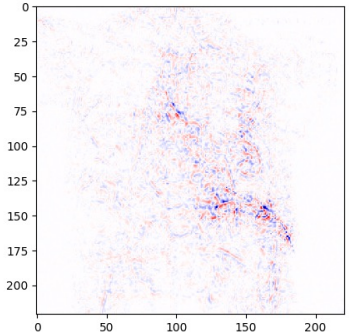




Pleural Effusion

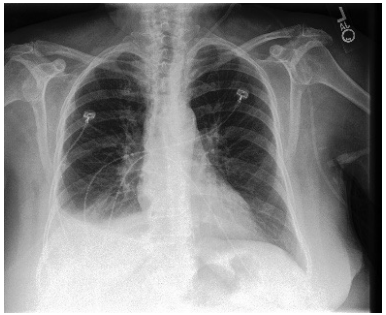
Incorrect

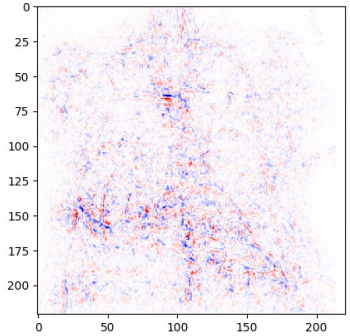




Pneumothorax

Correct



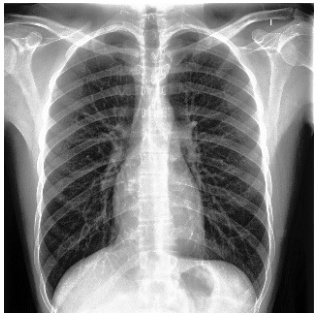


Pneumothorax

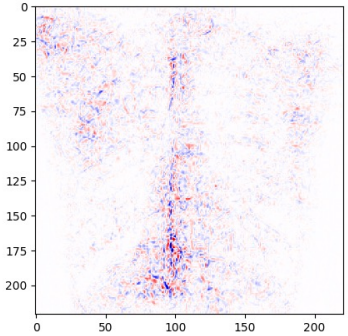
Incorrect

Stockholm University

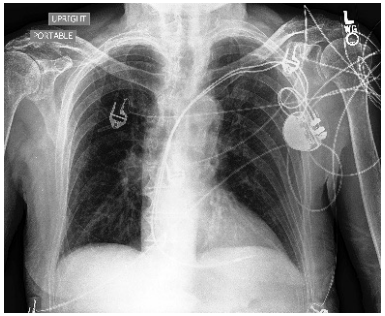
LRP Visualization



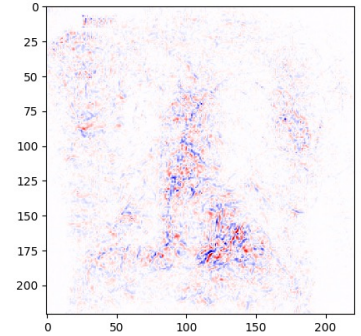
No Finding



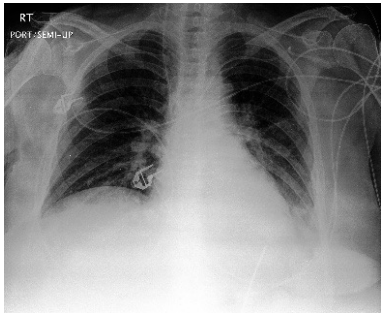
Correct



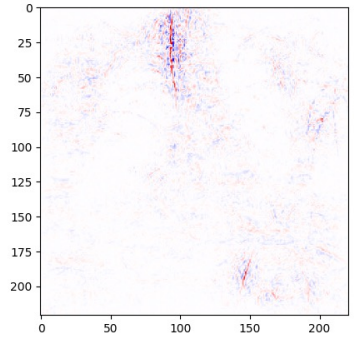
No Finding



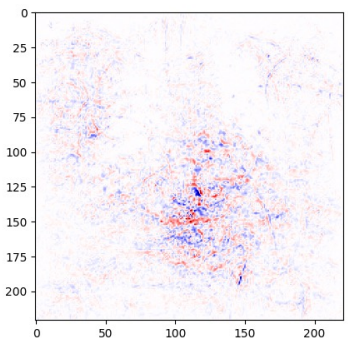
Incorrect



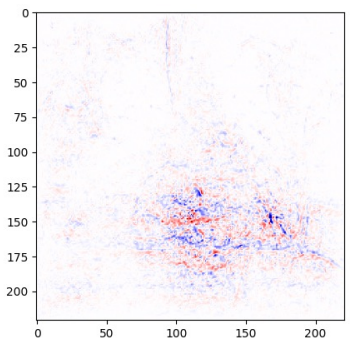
Multi-Label



Support Devices



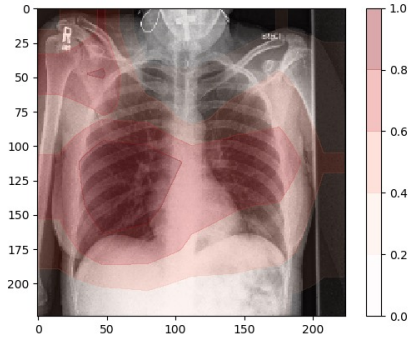
Lung Opacity



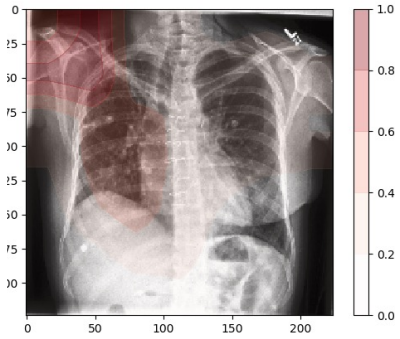
Pleural Effusion

Stockholm University

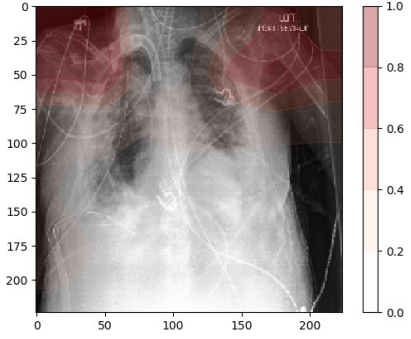
Grad-CAM Visualization



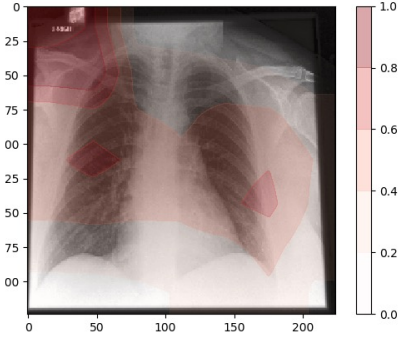
No Finding-Correct



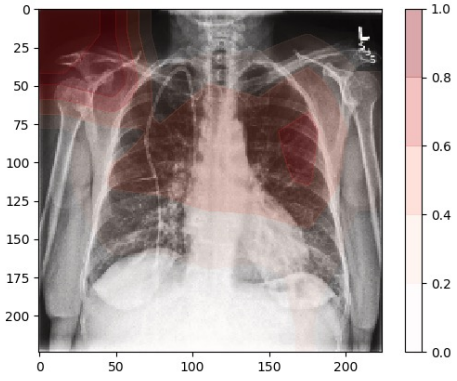
No Finding-Incorrect



Pneumothorax-Correct

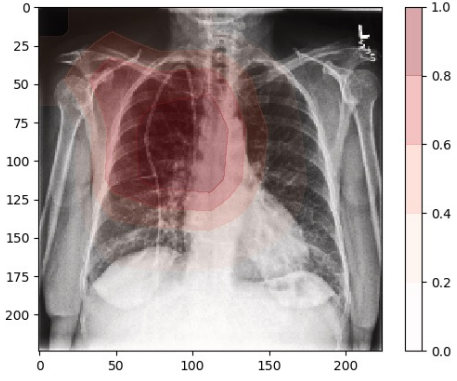


Pneumothorax-Incorrect



Lung Opacity

Multi-Label



Support Devices

Stockholm University

Key Insights

- Fine-tuning hyper-parameters
- Pathology detection versus segmentation in image
- Trade-off between the best classification and multiple classifications
- Negative contribution indication
- Combination of LRP and Grad-CAM

References

[1] http://www.heatmapping.org/slides/2017_ICASSP_3.pdf

[2] http://gradcam.cloudcv.org/

[3] https://stanfordmlgroup.github.io/competitions/chexpert/

[4] https://freidok.uni-freiburg.de/fedora/objects/freidok:149856/datastreams/FILE1/content

Thank You!

Stockholm University