XAI system for Pneumonia detection using chest X-ray images

XAI Project

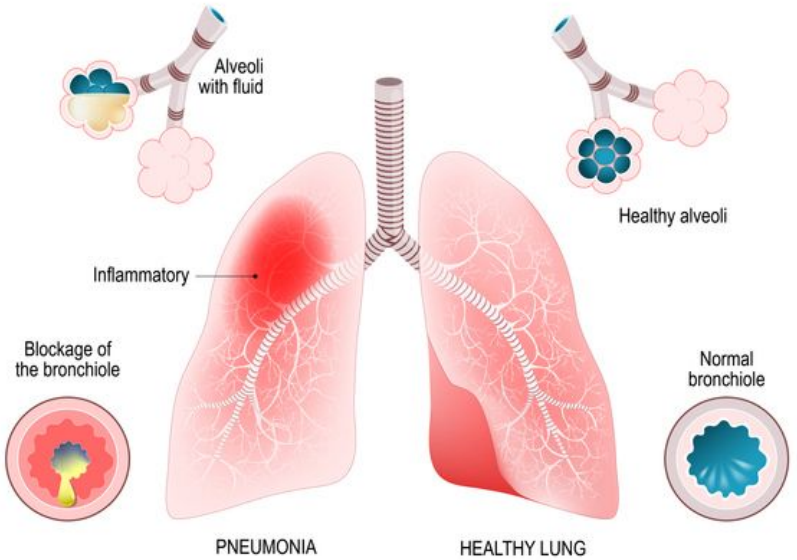
Alessia Fantini Lorenzo Marini Alessandro Quarta

Overview

- What is Pneumonia?
- The dataset
- Pipeline of ML process
  - Image preprocessing
  - Model construction
- XAI methodologies
- Results
- Conclusions and possible improvements
- References

What is Pneumonia?

- Pneumonia is an **inflammatory condition of the lung** affecting primarily the small air sacs known as alveoli.
- **Symptoms**
  typically include some combination of productive or dry cough, chest pain, fever and difficulty breathing.
- **Causes**
  Pneumonia is usually caused by infection with viruses or bacteria and less commonly by other microorganisms, certain medications or conditions such as autoimmune diseases.
- **Diagnosis**
  often based on symptoms and physical examination. **Chest X-ray**, blood tests, and culture of the sputum may help confirm the diagnosis

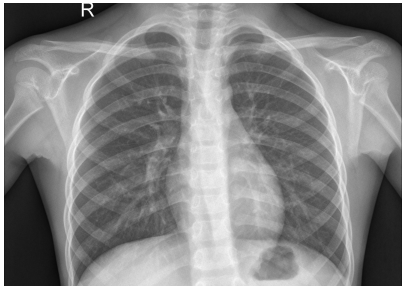


https://www.h-h-c.com/pneumonia-transmission-and-its-risk-factors/

Dataset

- The dataset is organized into 3 folders (train, test, val) and contains subfolders for each image category (Pneumonia/Normal). There are **5,863 X-Ray images** (JPEG) and **2 categories (Pneumonia/Normal)**.
- For the analysis of chest x-ray images, all chest radiographs were initially screened for quality control by removing all low-quality or unreadable scans.
- The diagnoses for the images were then graded by two expert physicians before being cleared for training in the AI system. In order to account for any grading errors, a third expert also checked the evaluation set.

Normal

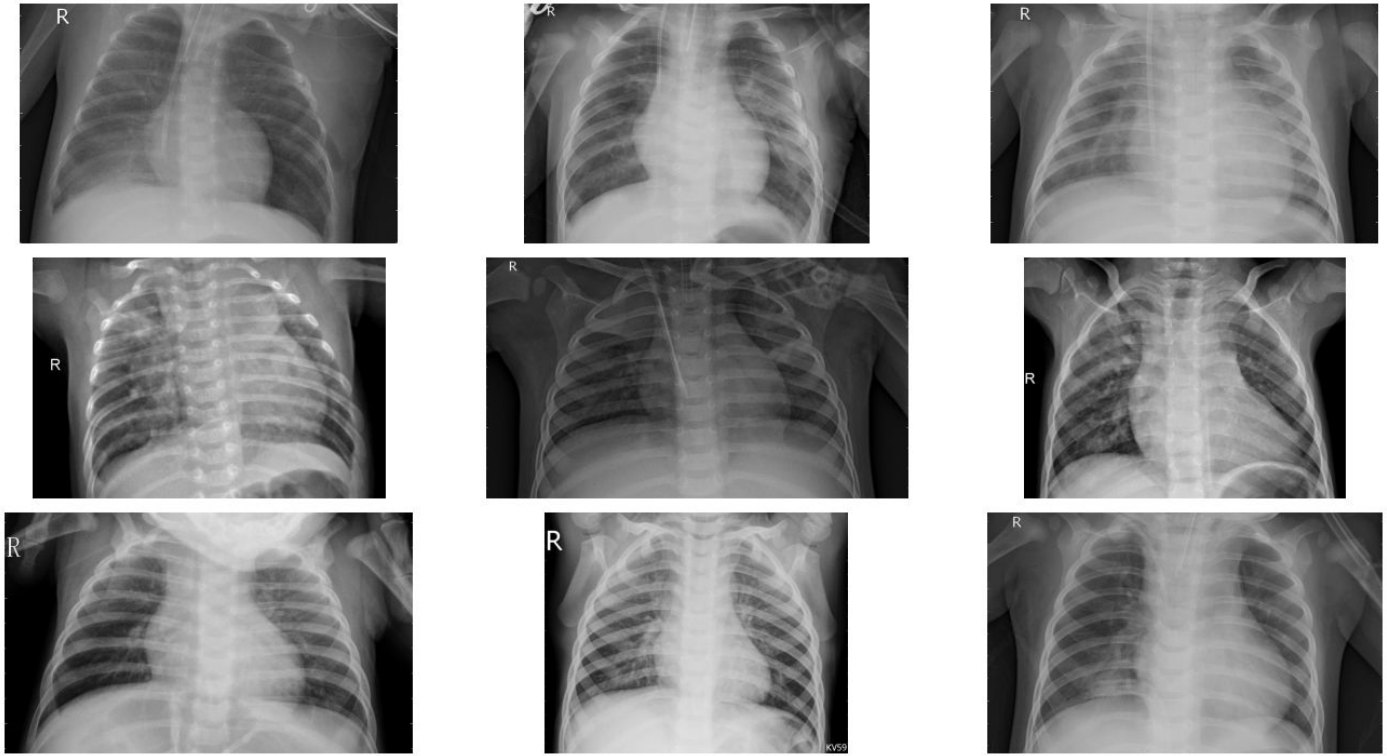


Pneumonia

```
dataset
├── test
│   ├── NORMAL
│   └── PNEUMONIA
├── train
│   ├── NORMAL
│   └── PNEUMONIA
└── val
    ├── NORMAL
    └── PNEUMONIA
```

https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia

Examples of *pneumonia* images



Dataset preprocessing

1) Validation dataset (only 8+8 images) << than the training and testing dataset (≈10^3) → we **merged training and validation** datasets
2) **Data augmentation** for mitigate the unbalanced class distribution

```
dataset
├── test
│   ├── NORMAL
│   └── PNEUMONIA
├── train
│   ├── NORMAL
│   └── PNEUMONIA
```

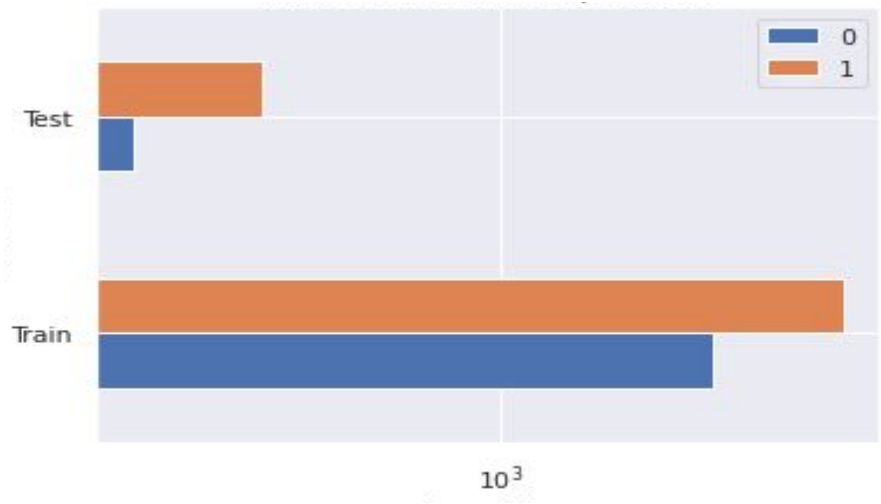
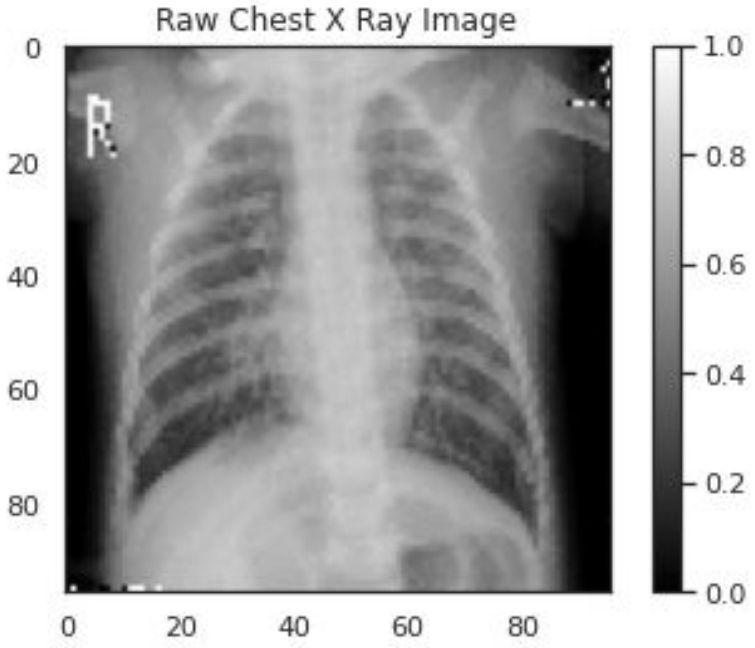


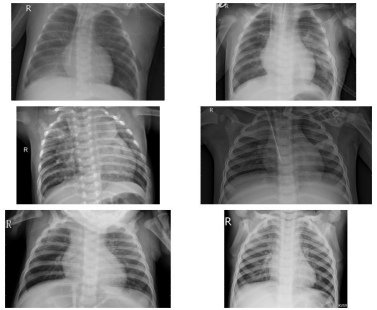
Image preprocessing

1) Resize → from **(2090×1858)** to **(96x96)**
2) Normalization (min/max) → from [0,255] to [0,1]

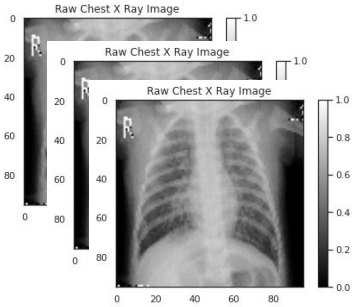


Pipeline of the analysis

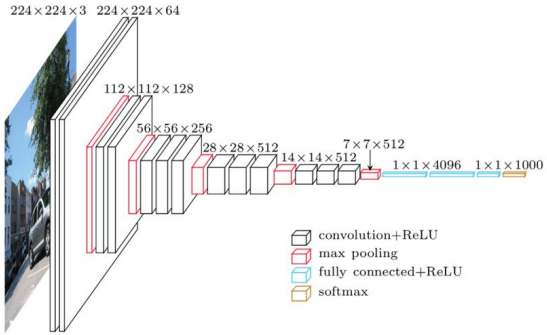
Initial dataset



Img prep + Data Aug

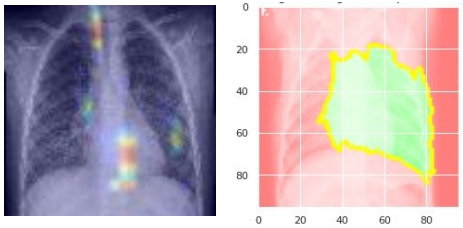


Transfer Learning (VGG16)



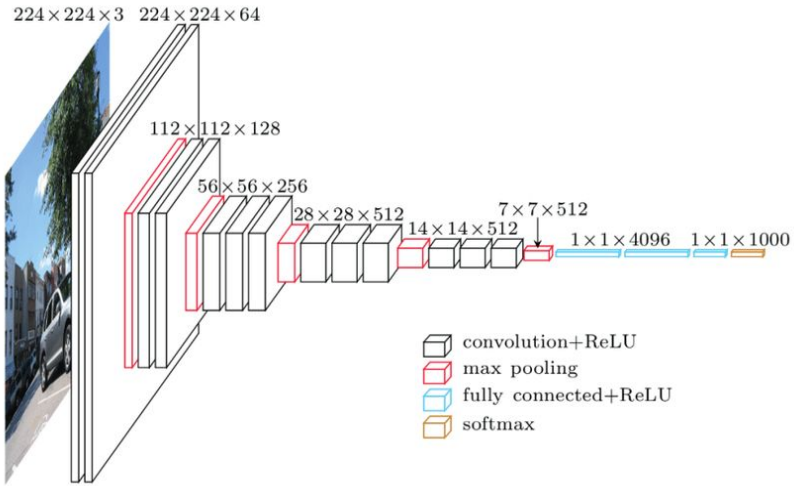
Training + Evaluation on test dataset

Explainability



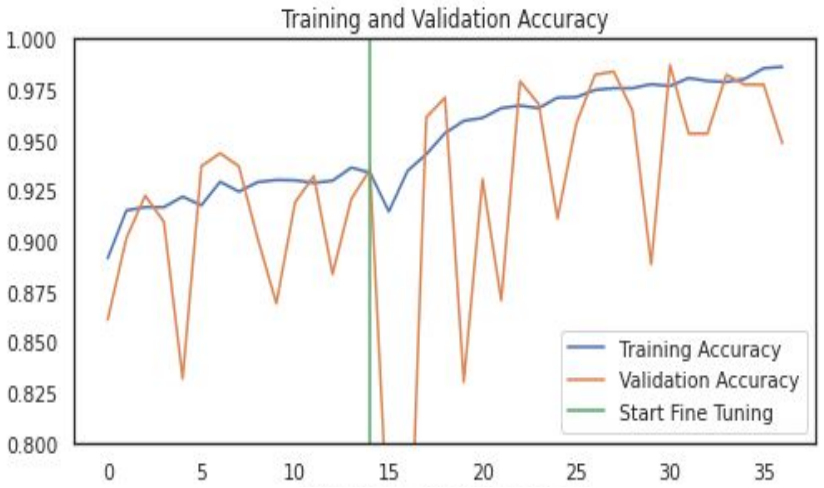
VGG16 model architecture

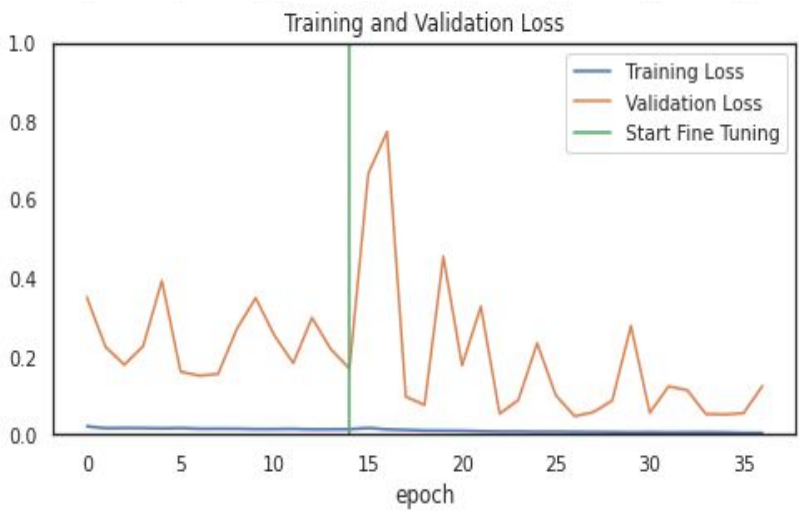
- VGG16, as its name suggests, is a **16-layer deep neural network**.
- VGG16 is thus a relatively extensive network with **138 million parameters**—it is huge even by today's standards. However, the simplicity of the VGGNet16 architecture is its main attraction.
- The VGGNet architecture incorporates the most important convolution neural network features.
- A VGG network consists of small convolution filters: **3 fully connected layers and 13 convolutional layers**.
- Further insights about a quick outline of the VGG architecture: datagen



Source: *ResearchGate*

History: loss and accuracy on training and validation





Evaluation of the model

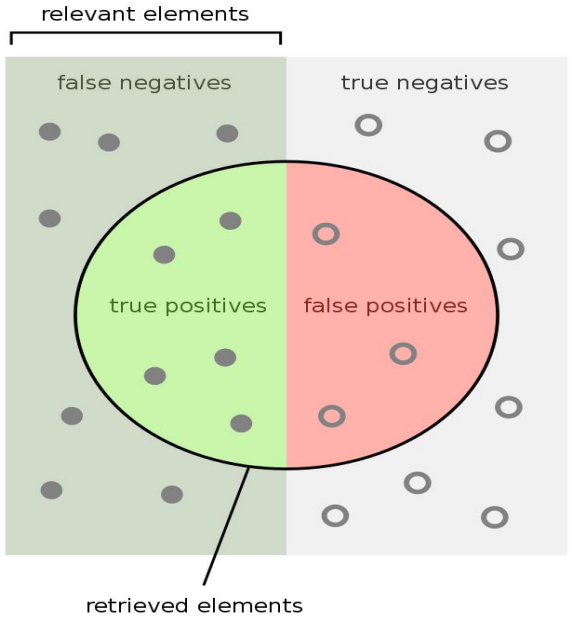
We evaluate the performance by means of:

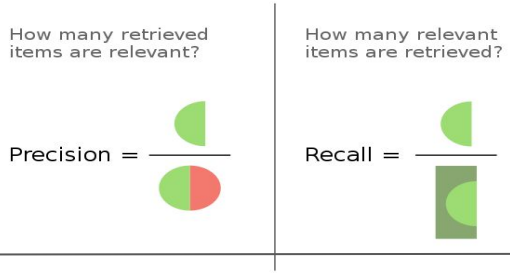
$$Acc = \frac{TP + FN}{TP + FP + TN + FN}$$

$$\mathrm{Pr} = \frac{TP}{TP + FP}$$

$$Rec = \frac{TP}{TP + FN}$$

$$F_1 = \frac{2}{Rec^{-1} + \mathrm{Pr}^{-1}}$$

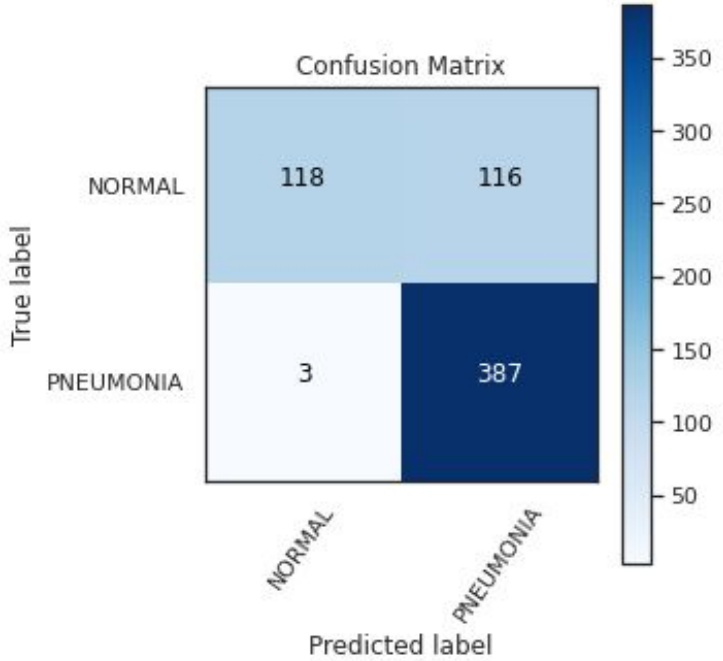


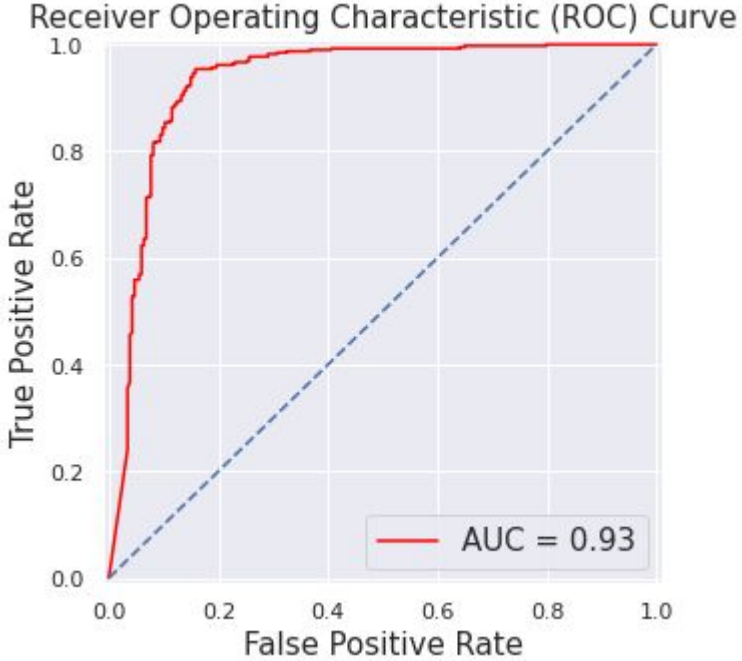


Performance of the model on the **test** dataset

	precision	recall	f1-score	support
NORMAL	0.98	0.50	0.66	234
PNEUMONIA	0.77	0.99	0.87	390
accuracy			0.81	624
macro avg	0.87	0.75	0.77	624
weighted avg	0.85	0.81	0.79	624

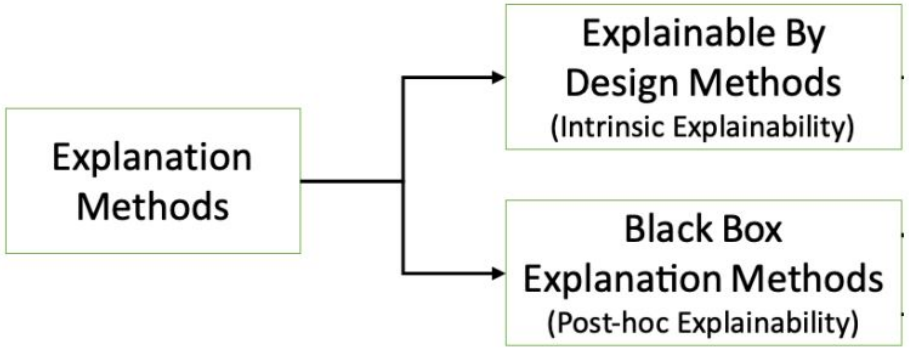
Confusion matrix and ROC curve

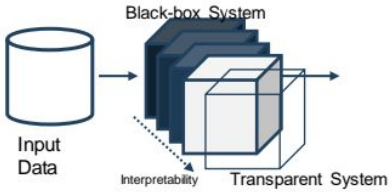


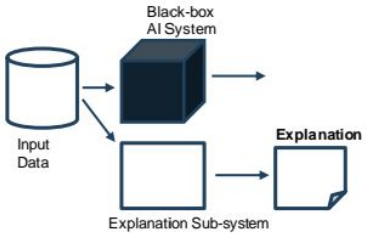


XAI methodologies

Explanation Methods

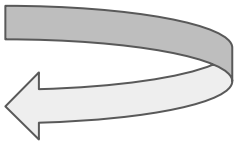






Local Explanation

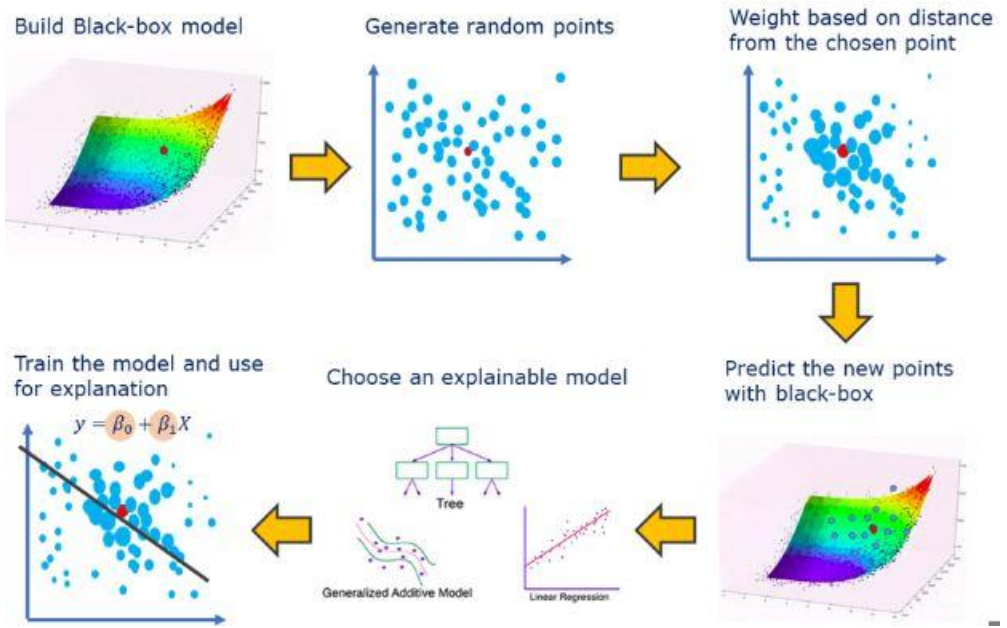
LIME
GRAD CAM



Explainer: Local Interpretable Model-Agnostic Explanation

LIME

Creates data with **perturbation** around the selected instance and defines a linear regression model to explain the classification criterion of that data locally.

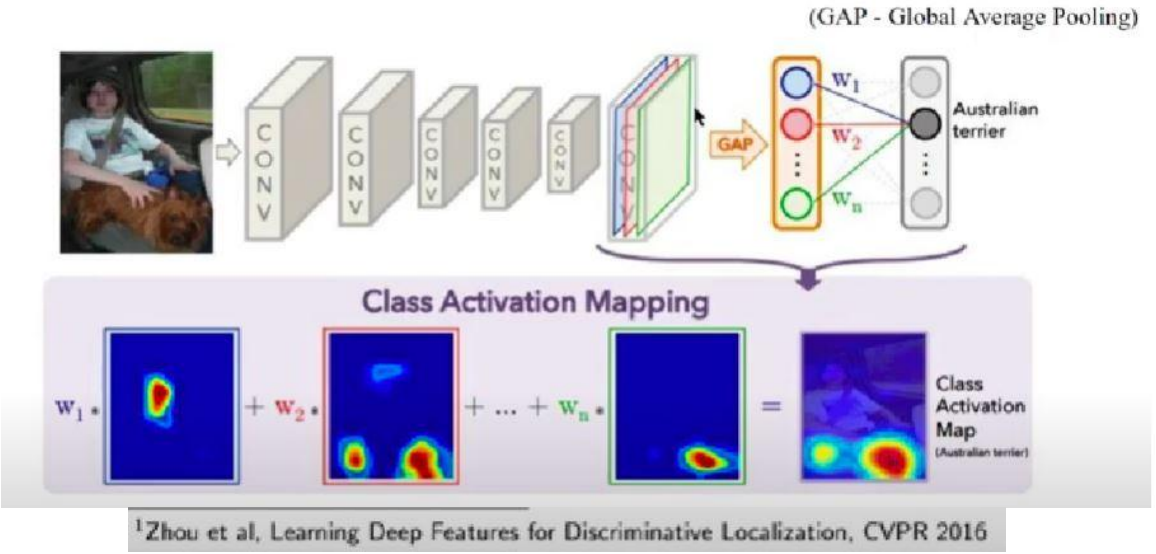


https://towardsdatascience.com/lime-explain-machine-learning-predictions-af8f18189bfe

Explainer: Gradient-weighted Class Activation Mapping

Grad-CAM

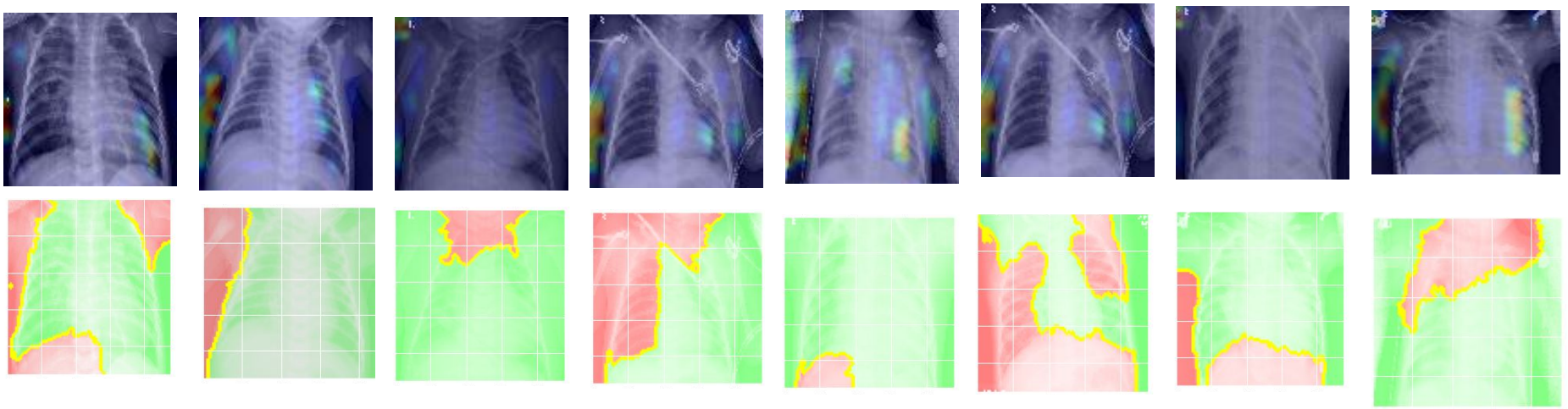
Utilizes the gradients of the classification score with respect to the ***final convolutional feature map***, to identify the parts of an input image that most impact the classification score.



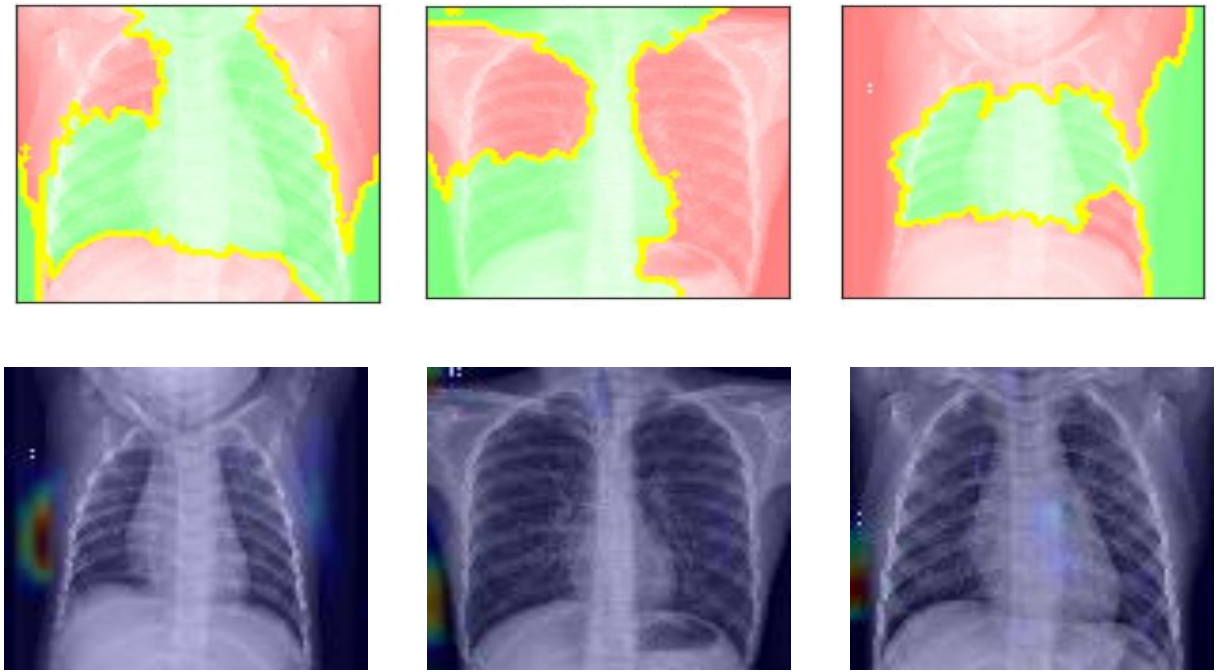
[1]Zhou et al, Learning Deep Features for Discriminative Localization, CVPR 2016

Results

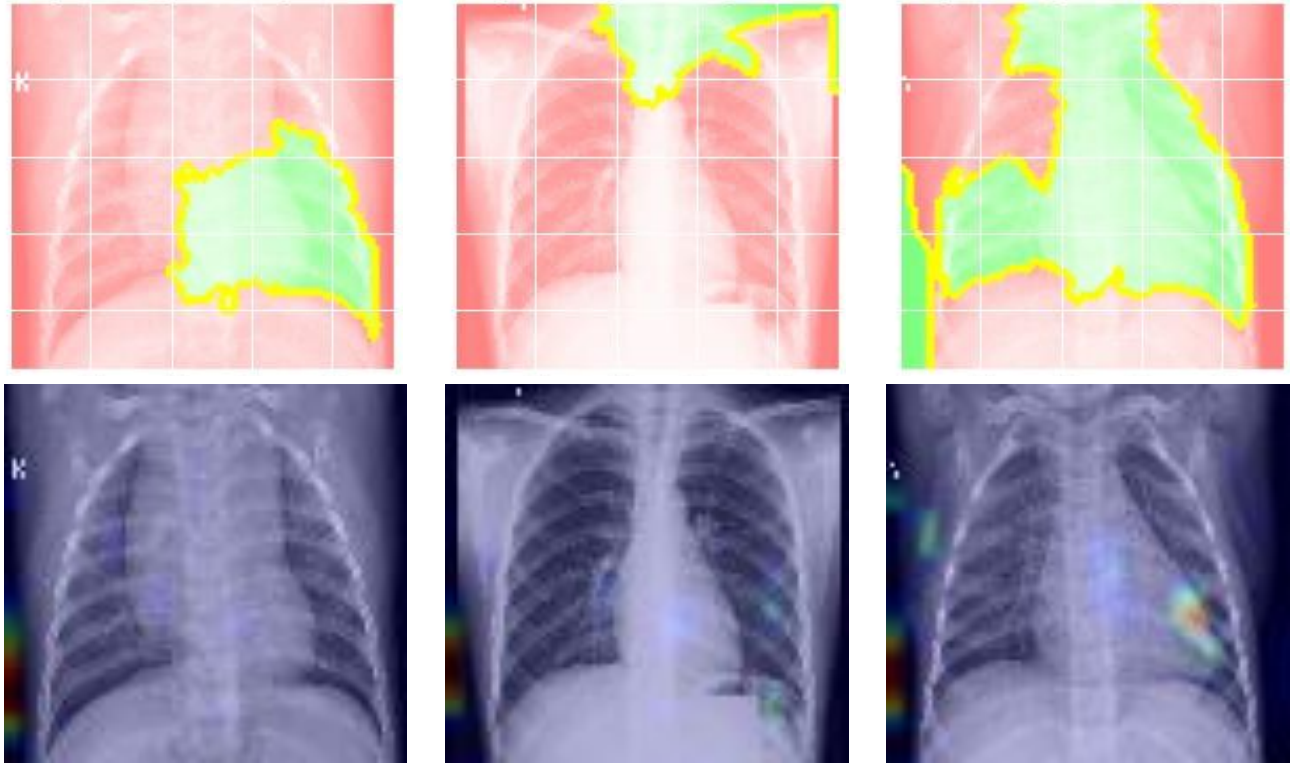
Results - LIME vs GradCAM (pred=1, ground truth=1)



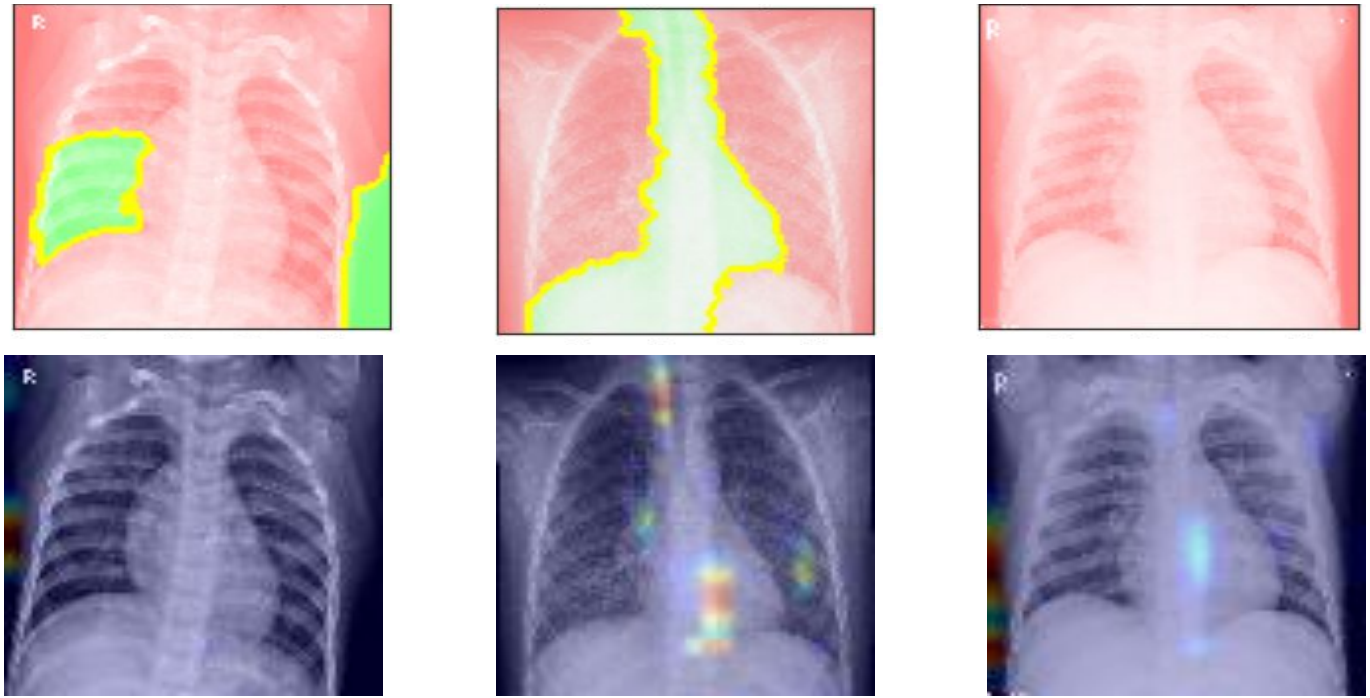
Results - LIME vs GradCAM (pred=0, ground truth=0)



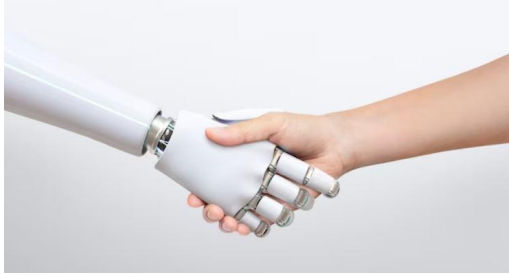
Results - LIME vs GradCAM (pred=1, ground truth=0)



Results - LIME vs GradCAM (pred=0, ground truth=1)



Designing an experiment with the final users



- Show the doctor the tool and **how it explains the results**
- Discuss the degree of agreement with the explanation provided by the algorithm, in order to **detect any doubts and critical issues**
- Imagine a test session in which the doctor and the algorithm label the images together and subsequently **assess the degree of perceived reliability** by administering an *ad hoc* questionnaire. Some questions to propose could be:

"Do you think the explanations provided are sufficient"?

"Would you use this tool in your clinical practice?"

"What limitations do you think the tool has?"

"What benefits do you think it could bring?"

Conclusions and possible improvements

In some cases, it seems that the model finds a certain pattern, but the heterogeneity is prevalent → *Maybe, are we biased in the pattern identification?*

Possible improvement:

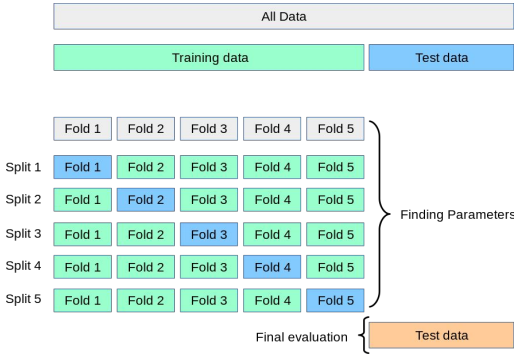
Automatic **segmentation** of lung to mitigate the effect of the image corners.

k-Cross validation techniques for unbiased evaluation of the performance.

Compare the results with different model (i.e., **ResNet50**, **GoogLeNet**)

- Implement other explainable methodologies (i.e., shap, counterfactual)
- ML-based approach:
  - image segmentation
  - **radiomic features** extraction
  - binary classification by machine learning model
  - XAI methods on **tabular data** (SHAP, LIME)

python + RADIOMICS

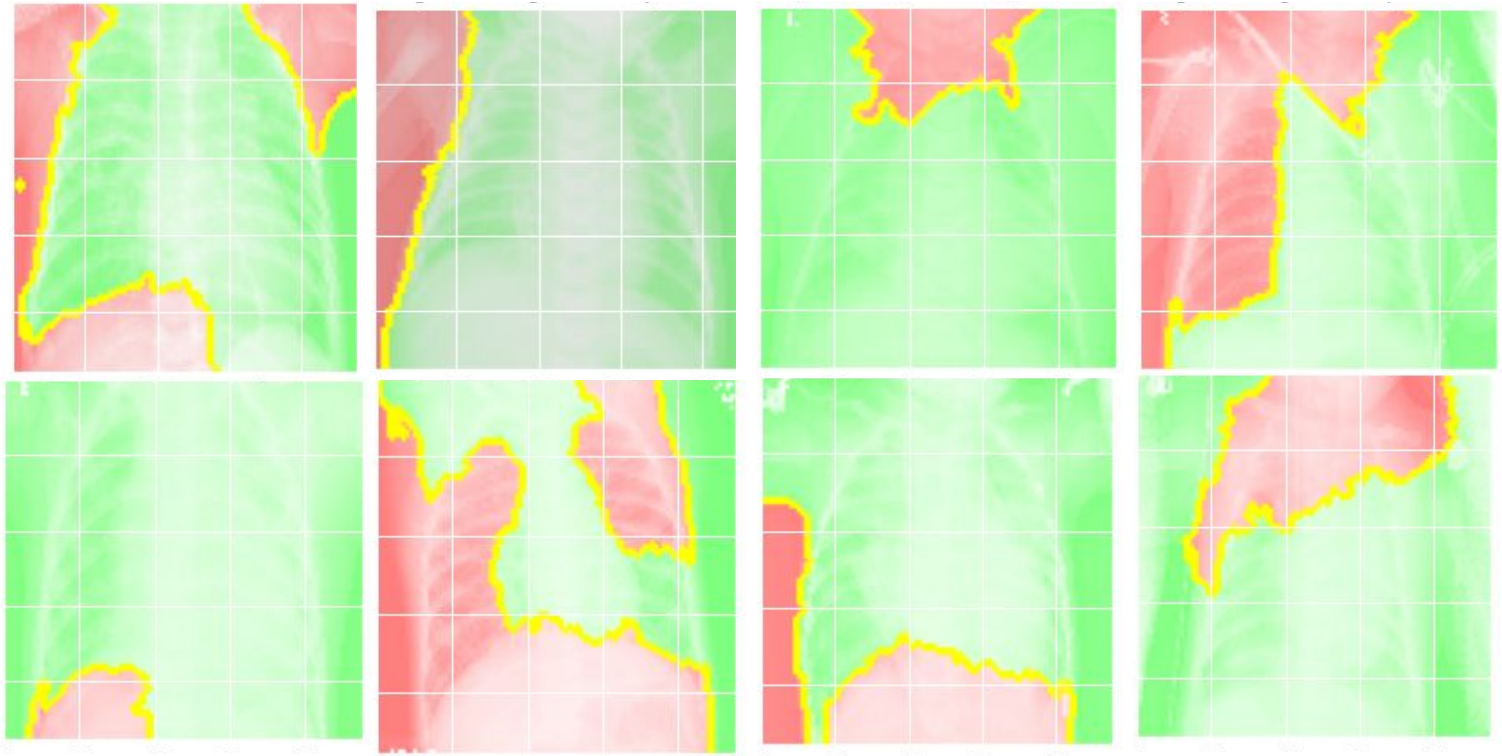


References

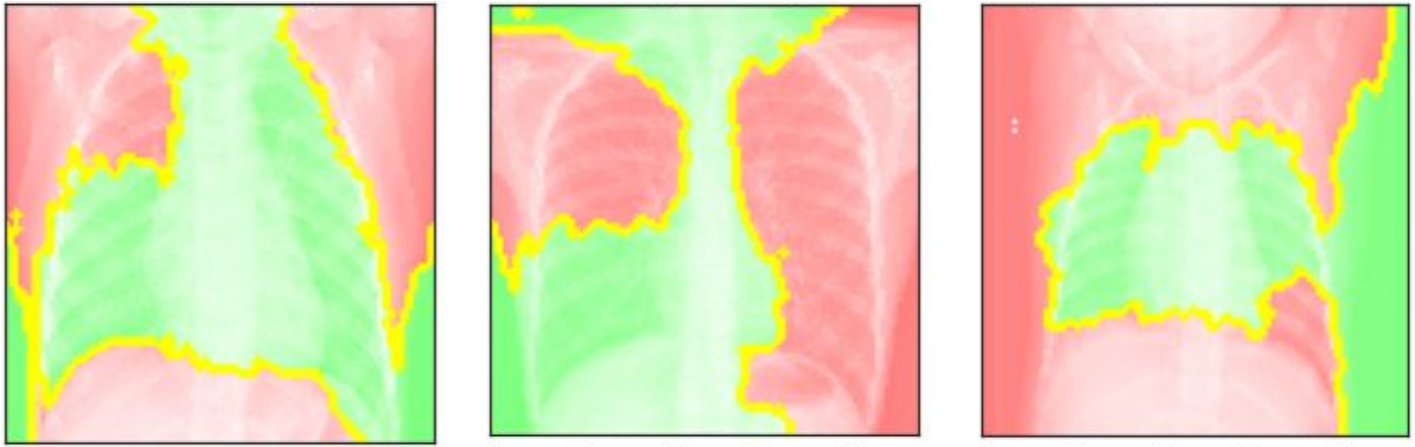
- **DATASET:** https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia
- **VGG16**: https://keras.io/api/applications/vgg/
- **LIME**: https://lime-ml.readthedocs.io/en/latest/
- **GRADCAM**: https://keras.io/examples/vision/grad_cam/

Backup slides

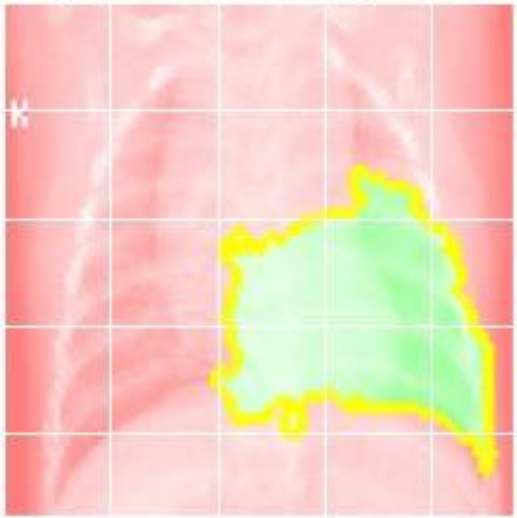
Results - LIME (pred=1, ground truth=1)

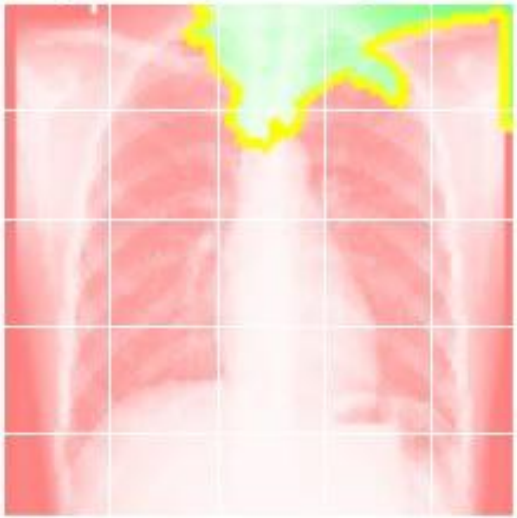


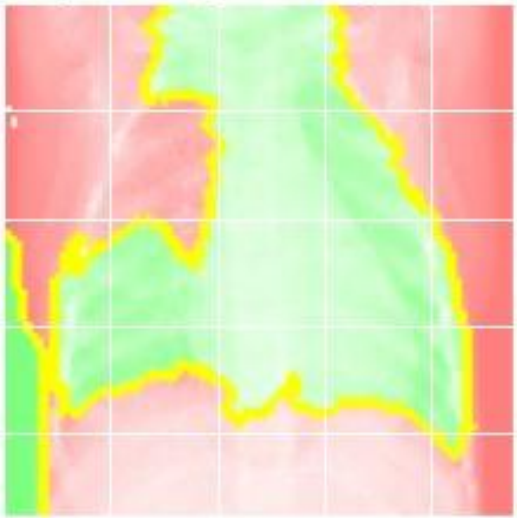
Results - LIME (pred=0, ground truth=0)



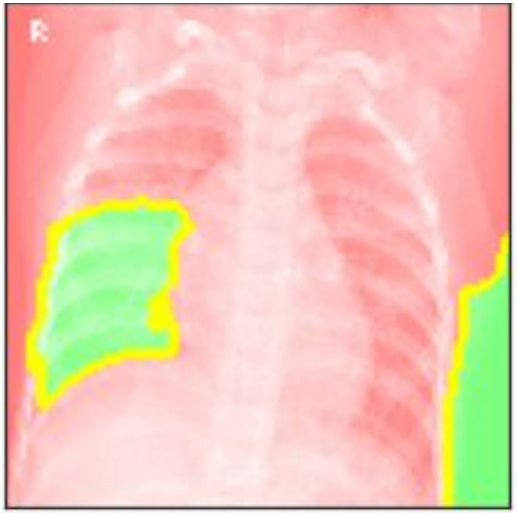
Results - LIME (pred=1, ground truth=0)



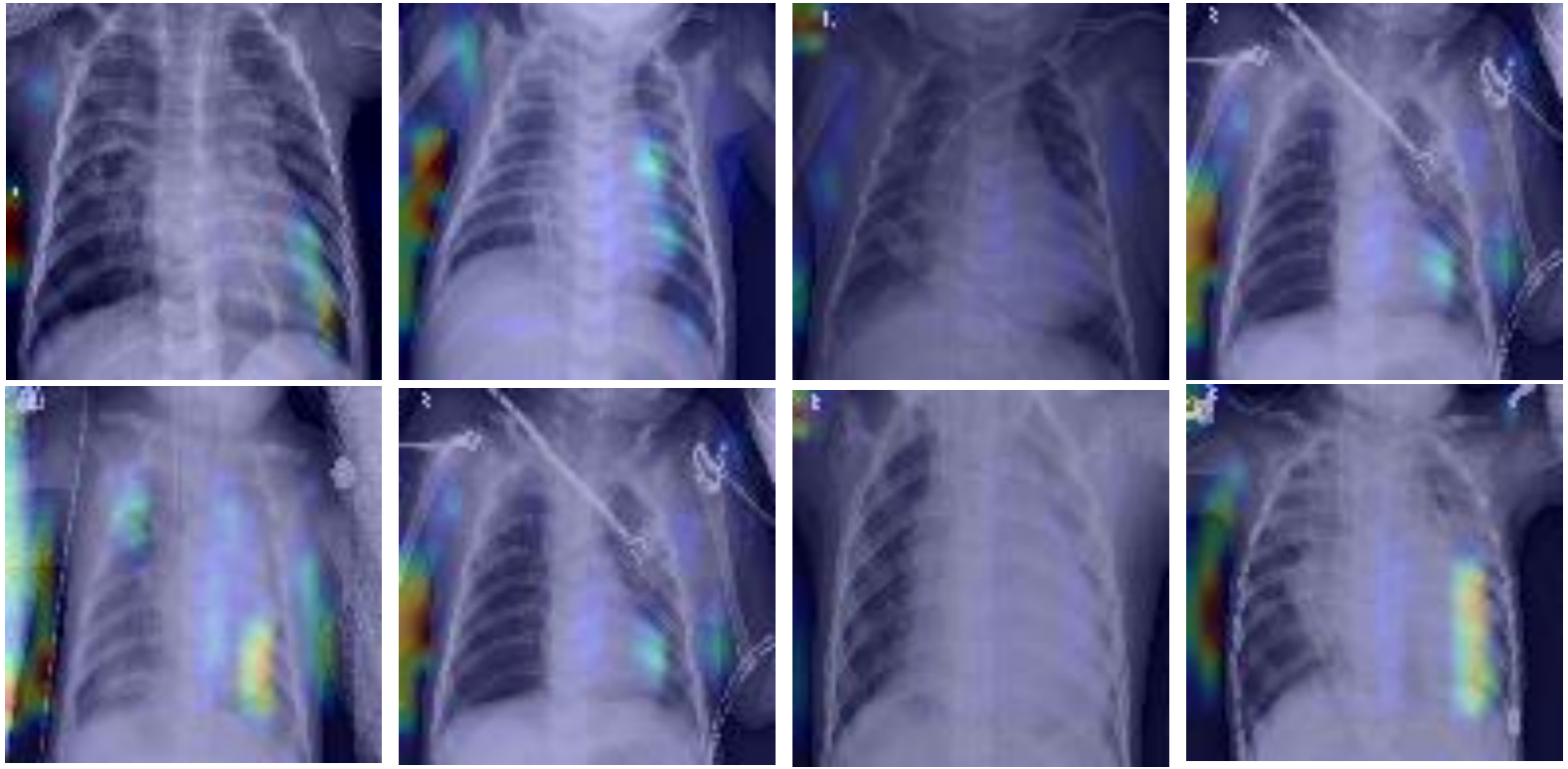




Results - LIME (pred=0, ground truth=1)



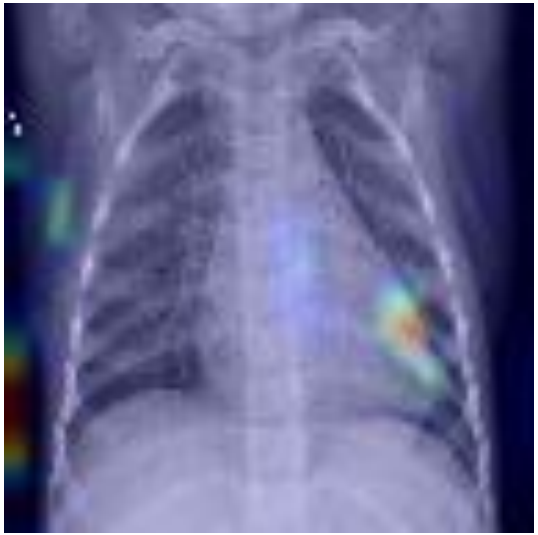
Results - GradCAM (pred=1, ground truth=1)



Results - GradCAM (pred=0, ground truth=0)

Results - GradCAM (pred=1, ground truth=0)





Results - GradCAM (pred=0, ground truth=1)

Explainable methods

Table 6: Explainers for black-boxes classifying image data sorted by explanation type: Saliency Maps (SM), Concept Attributions (CA), Counterfactuals (CF), and Prototypes (PR). For every method is indicated if is possible it for images (IMG) only, or for ANY type of data, if it is an Intrinsic (IN) or a Post-Hoc (PH) model, Local (L) or Global (G), and if it is model Agnostic (A) or model-Specific (S).

Type	Name	Ref.	Authors	Year	Data Type	IN/PH	G/L	A/S	Code
SM	SHAP	[84]	Lundberg et al.	2007	ANY	PH	L	A	link
	LIME	[102]	Ribeiro et al.	2016	ANY	PH	L	A	link
	ϵ-LRP	[17]	Bach et al.	2015	ANY	PH	L	S	link
	INTGRAD	[115]	Sundararajan et al.	2017	ANY	PH	L	S	link
	DEEPLIFT	[110]	Shrikumar et al.	2017	ANY	PH	L	S	link
	SMOOTHGRAD	[112]	Smilkov et al.	2017	IMG	PH	L	S	link
	XRAI	[70]	Kapishnikov et al.	2019	ANY	PH	L	S	link
	GRADCAM	[106]	Selvaraju et al.	2017	IMG	PH	L	S	link
	GRADCAM++	[27]	Chattopadhay et al.	2018	IMG	PH	L	S	link
	RISE	[97]	Petsiuk et al.	2018	IMG	PH	L	S	link
CA	TCAV	[75]	Kim et al.	2018	IMG	PH	L	A	link
	ACE	[49]	Ghorbani et al.	2019	IMG	PH	G	A	link
	CONCEPTSHAP	[129]	Yeh et al.	2020	IMG	PH	G	A	-
	CACE	[54]	Goyal et al.	2019	IMG	IN	G	A	-
CF	CEM	[40]	Dhurandhar, Amit, et al.	2018	IMG	PH	L	A	link
	ABELE	[57]	Guidotti et al.	2020	IMG	PH	L	A	link
	L2X	[29]	Chen et al.	2018	ANY	PH	L	A	link
	GUIDED PROTO	[118]	Van Looveren et al.	2019	IMG	PH	L	A	link
PR	MMD-CRITIC	[74]	Kim et al.	2016	ANY	IN	G	A	link
	-	[76]	Koh et al.	2017	ANY	PH	L	A	link
	PROTONET	[28]	Chen et al.	2019	IMG	IN	G	S	link