Advanced Deep Learning - Introduction to Explainable Artificial Intelligence (XAI)

XAI methods & metrics

STASSIN Sédrick

Professor: MAHMOUDI Sidi Ahmed

https://www.perficient.com/what-we-do/data-and-intelligence/artificial-intelligence

UMONS
Université de Mons

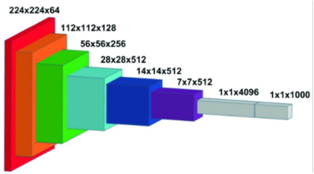
INTRODUCTION

Training

Dataset

Train

DL Model



Train

Train metric result

95%

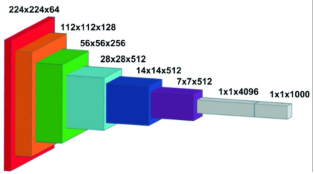
Testing

Image

Green Lizard

Test

Trained DL Model



Test

Test *metric result*

90% - Green Lizard

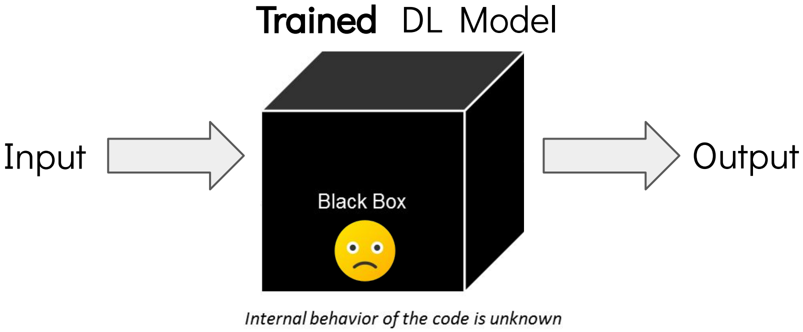
High result

Congratulations!

Is that enough ?

POLYTECH MONS

INTRODUCTION



Black Box problem of deep neural networks

European Commission

2024

The **AI Act**

Trustworthy AI systems must be considered:

- **Lawful –** Operating within the limits of law
- **Ethical –** Fair models that do not discriminate
- **Robust –** Delivering <u>reliable</u> results in all considered situations

http://artificialintelligencemania.com/2019/01/10/the-black-box-problem/

INTRODUCTION

XAI

Explainable Artificial Intelligence

How does this system work ?

Can I trust this AI model ?

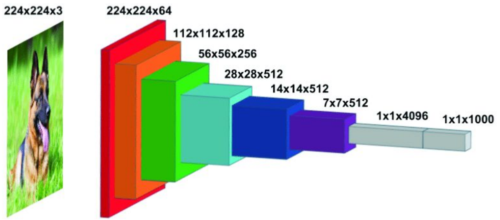
Under what circumstances ?

How to evaluate this explainability ?

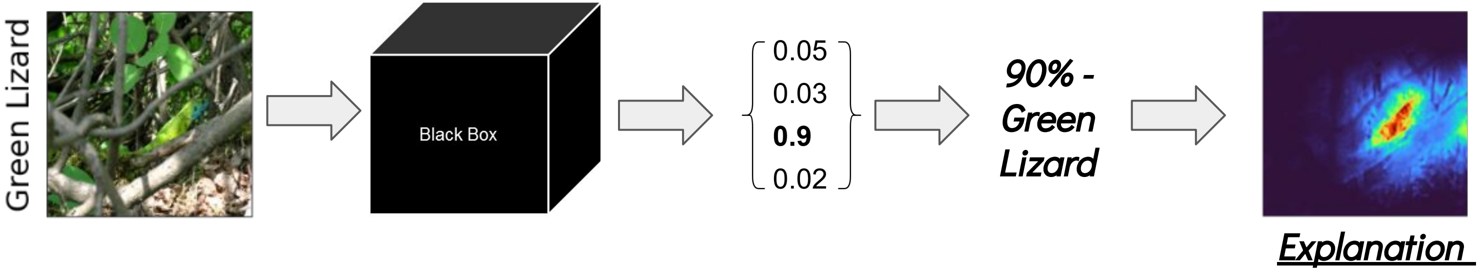
EXplainable Artificial Intelligence (XAI)

XAI in computer vision – Images

applied to CNN



- Most used - attribution-based methods = ***saliency methods***



CONTENTS

02

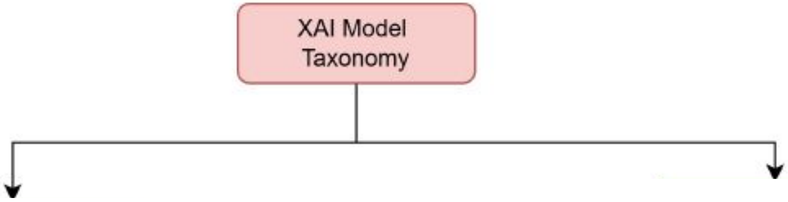
BACKGROUND

XAI Taxonomy

02

XAI Taxonomy

Model Taxonomy



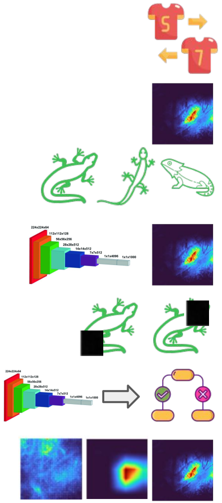
02

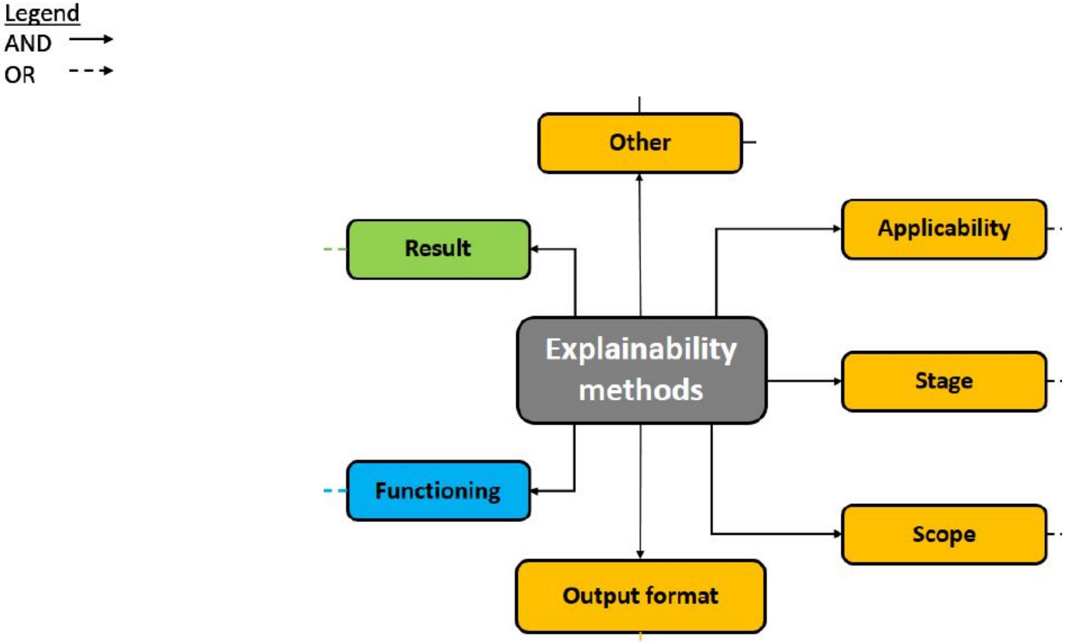
XAI Taxonomy

Method Taxonomy

Taxonomy:

- Conceptual
- Result-based
- Function-based

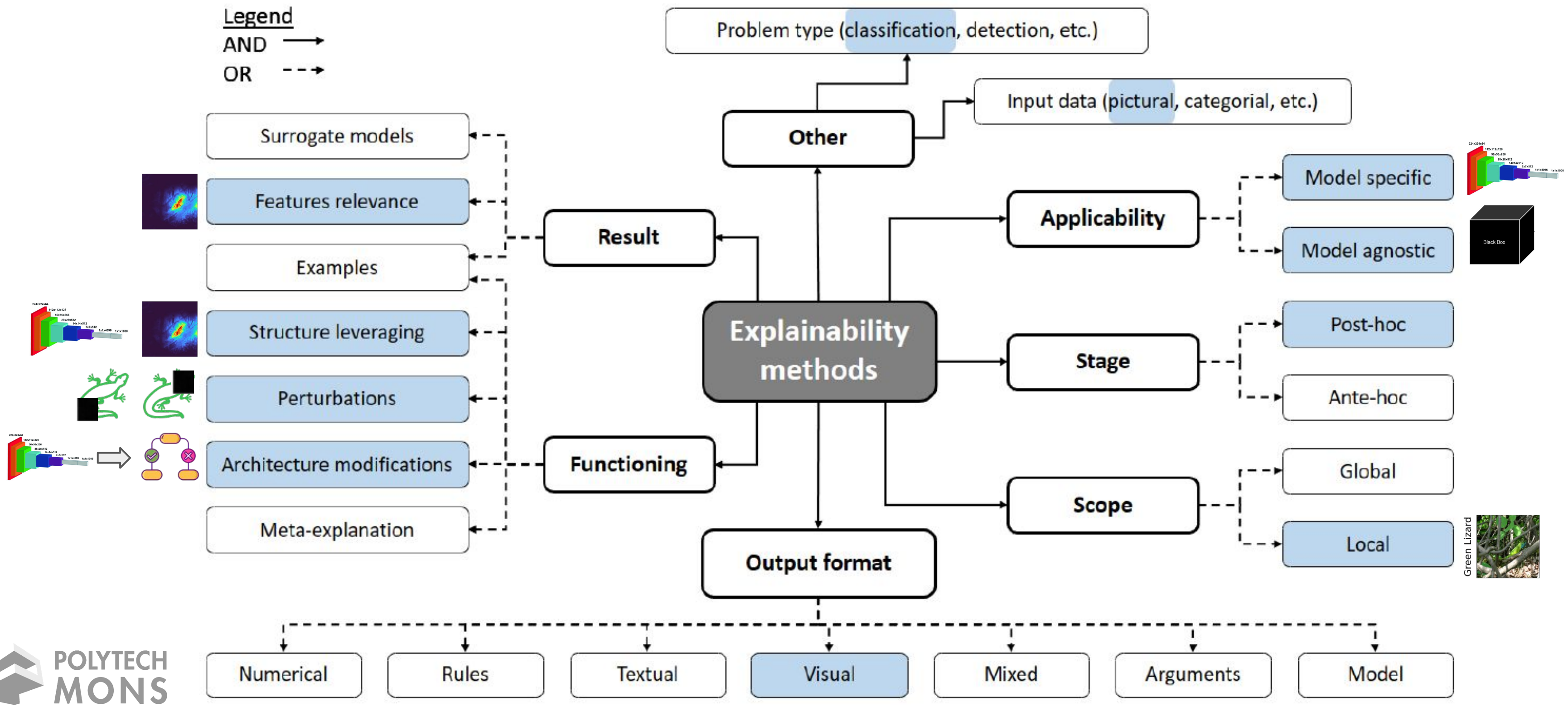




02

XAI Taxonomy

Selected Method Taxonomy



UMONS
Université de Mons

03

EXPLAINABILITY AND BIAS DETECTION

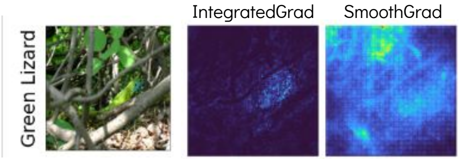
POLYTECH MONS

03a Saliency methods

Gradient-based methods :

[Smilkov et al., 2017] [Sundarajan et al., 2017] ...

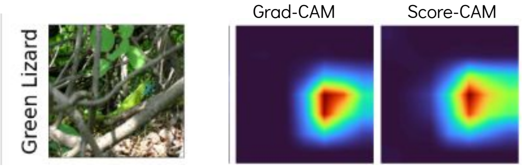
- Use the gradient (e.g. back-propagated) to picture the derivative of the model output w.r.t the input image.



CAM-based methods :

[Selvaraju et al., 2017] [Wang et al., 2020] ...

- Produce a weighted sum of the activations from a convolutional layer



Perturbation-based methods :

[Petsiuk et al., 2018] ...

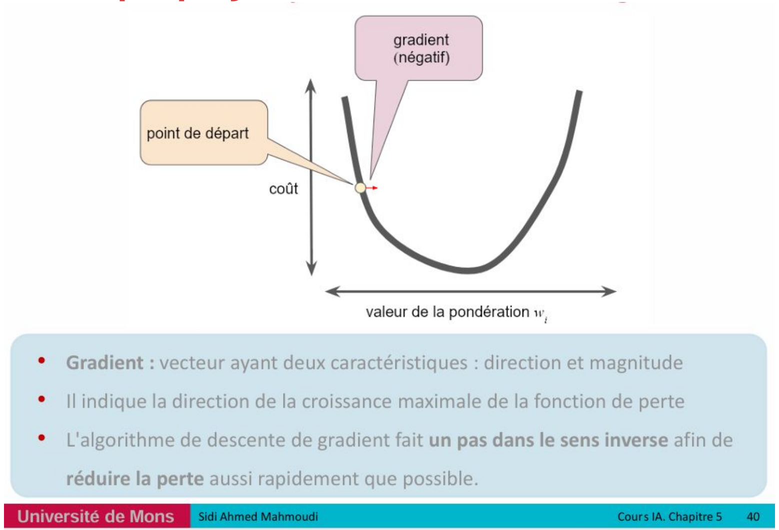
- Study the output model response to small changes in the input.

03a Gradient-based Methods

Gradients

[Simonyan et al., 2013]

Neural Network Backpropagation



Loss function

$$w_{t+1} = w_t - \alpha \frac{\partial \mathcal{L}}{\partial w_t}$$

learning rate

weight

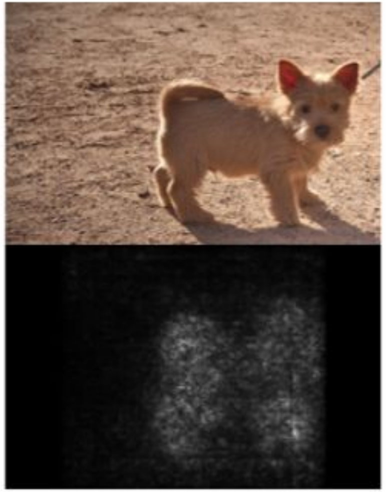
Formula

Loss function

$$R^c = \frac{\partial \mathcal{L}_c(x)}{\partial x}$$

Relevance of class c

Input



SmoothGrad: Add ***Gaussian noise*** to ***n*** samples of the image - Compute gradients - Average

Integrated Gradients: *Create* ***n*** images - ranging linearly from a ***baseline image*** *to the* ***input image*** - Average the gradients

UMONS
Université de Mons

03a Relevance-based Methods

Layer-wise Relevance Propagation (LRP)

[Bach et al., 2015]

Conservation property

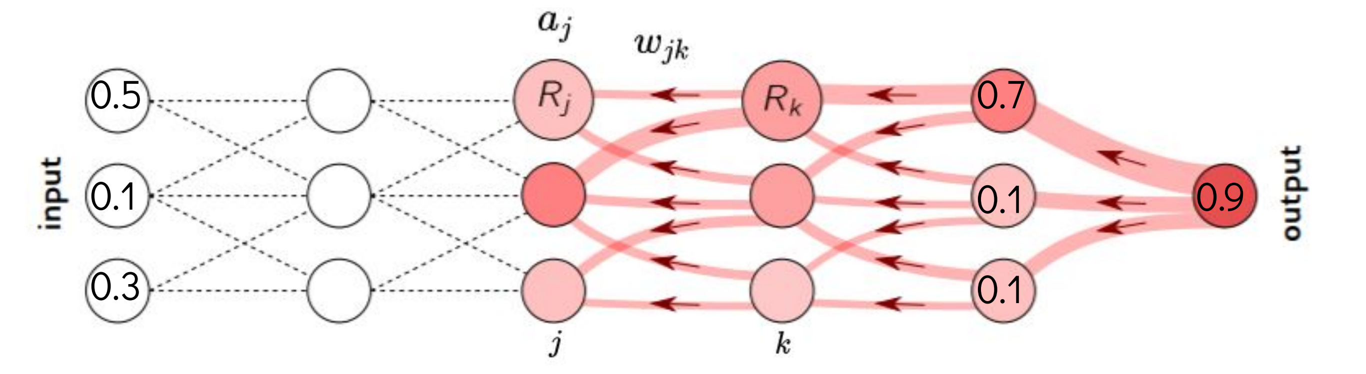
What has been received by a neuron must be redistributed to the lower layer in equal amount

j k Consecutive layers

R_j R_k Neuron relevances

a_j Neuron ***j*** activation

w_{jk} Weight between neuron ***j*** and ***k***



$$R_j = \sum_k \frac{a_j w_{jk}}{\sum_{0,j} a_j w_{jk}} R_k$$

POLYTECH MONS

03a Perturbation-based Methods

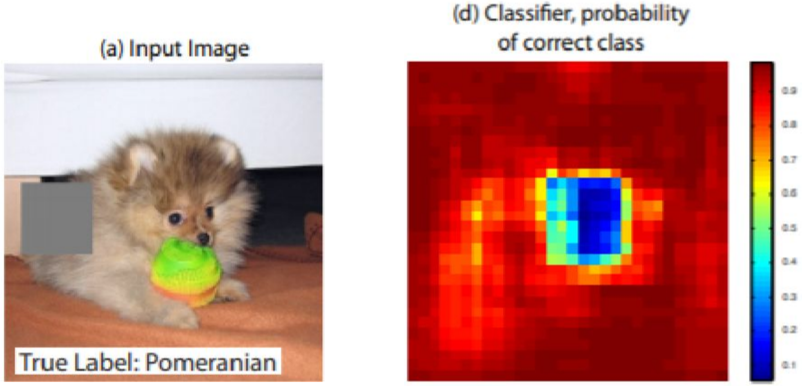
Occlusion

[Zeiler and Fergus, 2014]

Main steps

- Choose an image, a size of square
- For all possible positions of the square in the image
  - Occlude
  - Compute score

Pixels are important if the class score drops significantly



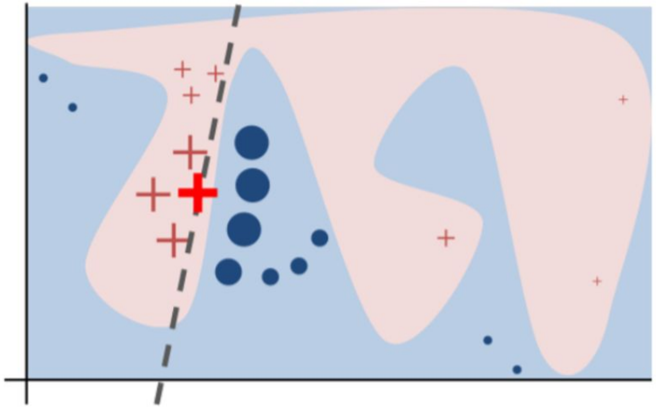
03a Perturbation-based Methods

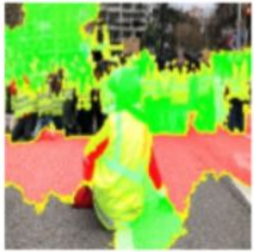
Local Interpretable Model-agnostic Explanation (LIME)

[Ribero et al., 2016]

Principles

- Fit a linear model to ***n*** perturbed samples of an image
    - Predicted by the complex model
    - With distances as weights
- The linear model using the ***m*** best features provides a locally faithful explanation for the image





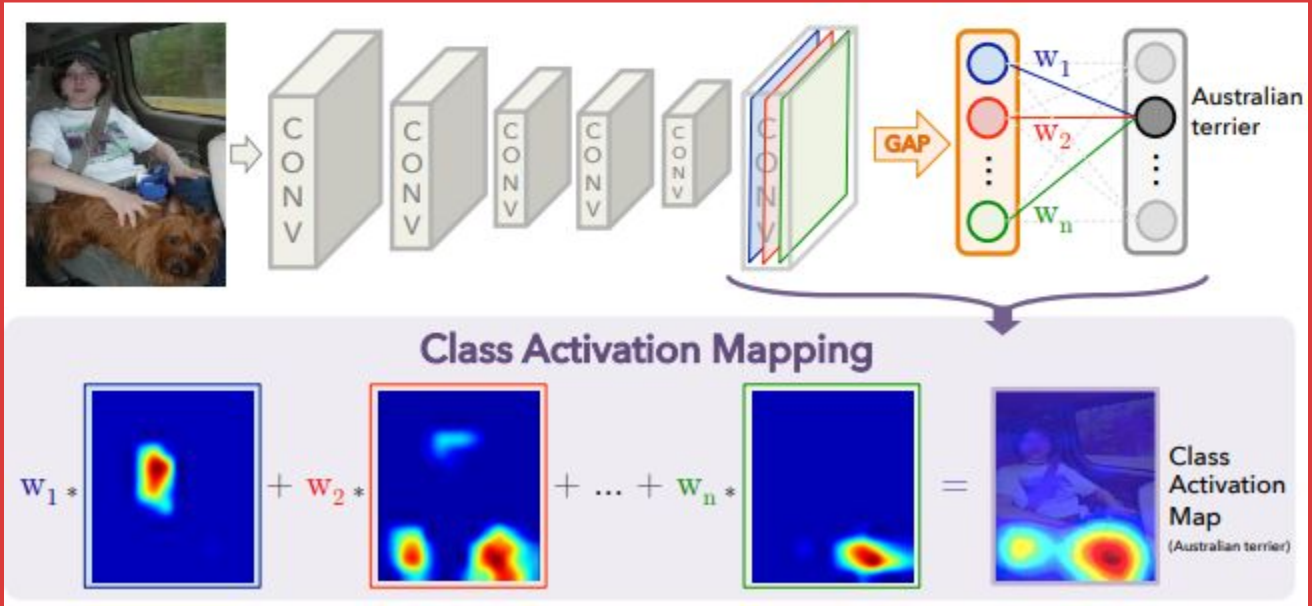
CAM-based Methods

Class Activation Mapping (CAM)

[Zhou et al. 2016]

Class Activation map for class **c**

$$R^c = \sum_{n=1}^{N} w_{n,c} A_n$$



03a CAM-based Methods

Grad-Class Activation Mapping (Grad-CAM)

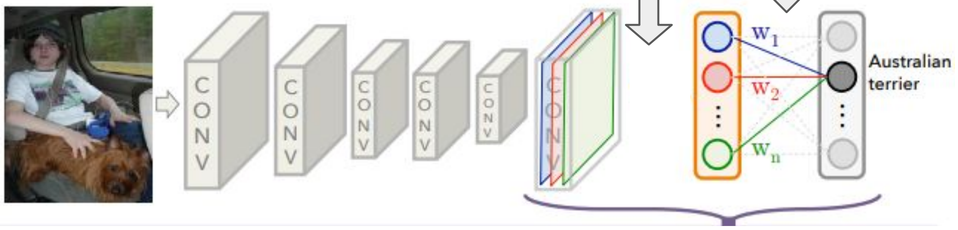
[Selvaraju et al., 2017]

Main differences

1. Generalization to multiple dense layers
   a. **Gradient** w.r.t feature maps of the last convolutional layer
2. Not constrained to models with GAP
   a. **GAP** is used to obtain neuron importance weights
3. **ReLU** activation to only keep positive contributions

Not constrained to use GAP

Multiple dense layers accepted



$$w_{n,c} = \overbrace{\underbrace{\frac{\partial y_c}{\partial A_{n,i,j}}}_{\text{gradients via backprop}}}^{\text{global average pooling}}$$

$$R^c = \underbrace{\left(\sum_{n=1}^{N} w_{n,c} A_n\right)}_{\text{linear combination}}$$

UMONS
Université de Mons

03b COVID-19 CLASSIFICATION

Problem

Covid-19 Image Classication

Dataset

2621 training (90%) – 284 testing (10%)
3 classes

Models

VGG16 – VGG19 – ResNet50 – DenseNet121 – DenseNet201

Analysis

1- XAI on predicted class
2- Model comparison

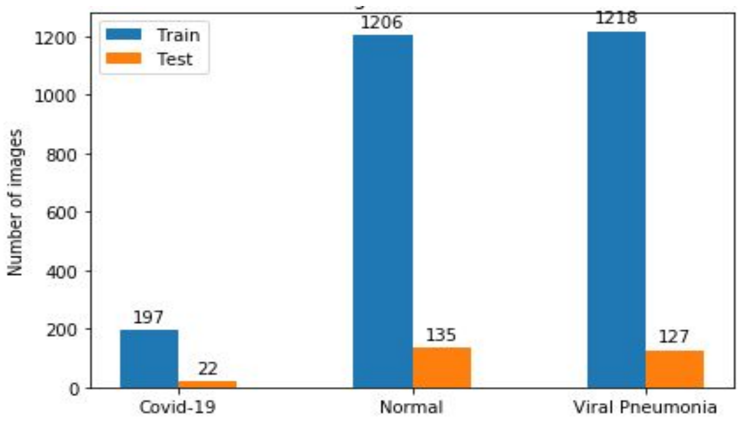
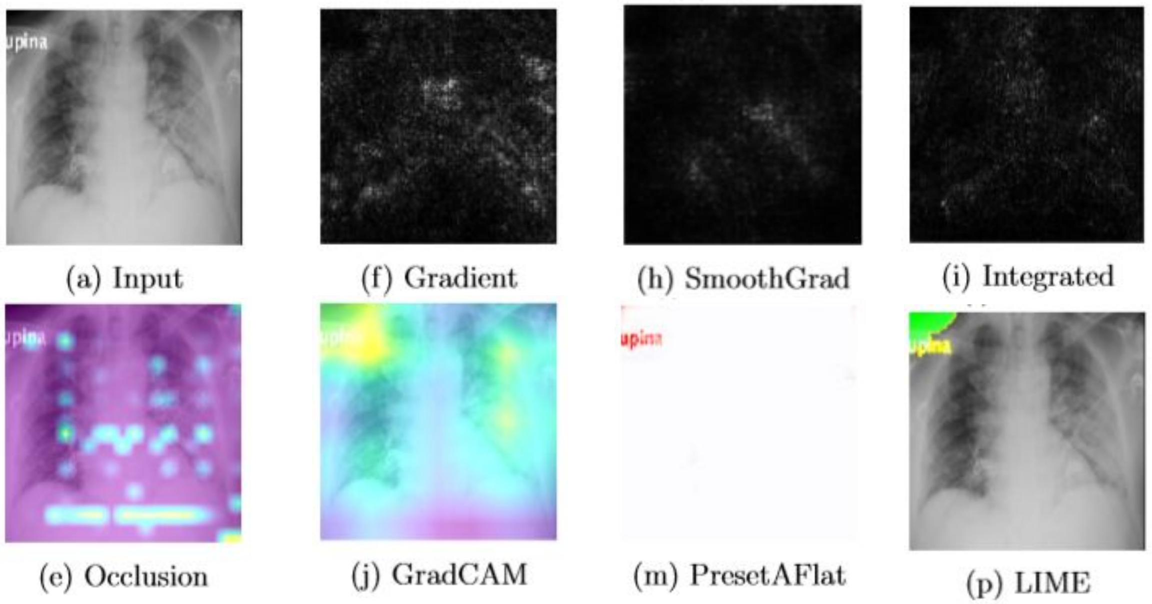


Image distribution among existing classes

03b COVID-19 CLASSIFICATION

XAI : Method comparison - Predicted class



DenseNet121 explanation for a Covid-19 x-ray image

- Gradients + Occlusion: noisy results
- Best visual results: LRP (Preset) + LIME + GradCAM

→ Letter Detection ?

XAI helps to detect biases in models

Model bias ?

03b COVID-19 CLASSIFICATION

XAI : Model Comparison

Models

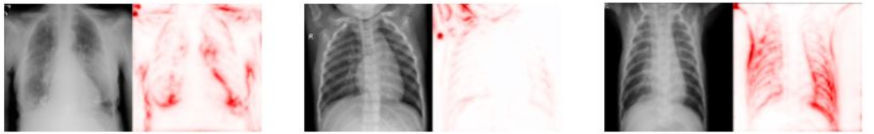
- DenseNet201

(a) Covid-19 Image (b) Normal Image (c) Viral Pneumonia Image

- ResNet50

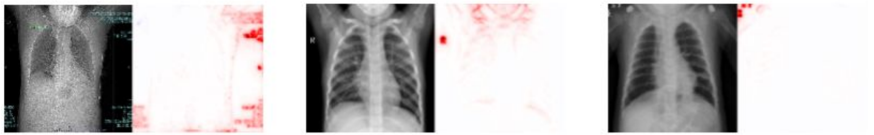
(a) Covid-19 Image (b) Normal Image (c) Viral Pneumonia Image

- VGG16



(a) Covid-19 Image (b) Normal Image (c) Viral Pneumonia Image

- VGG19



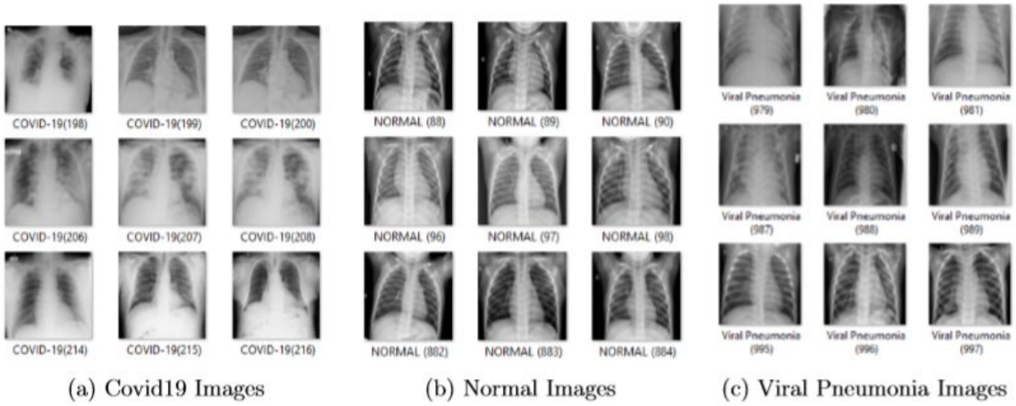
(a) Covid-19 Image (b) Normal Image (c) Viral Pneumonia Image

- Every model uses zones out of interest
- VGG-16 seems less reliant

→ **XAI can guide model selection**

- <u>Normal Images:</u> Detection towards the head ?

03b COVID-19 CLASSIFICATION



Visual comparison between 9 images by class

Biases due to:

- Camera positioning
- Arm positioning and patient characteristic

XAI helps to understand the sources of bias

UMONS
Université de Mons

03b COVID-19 CLASSIFICATION

CT-Scan Images classification

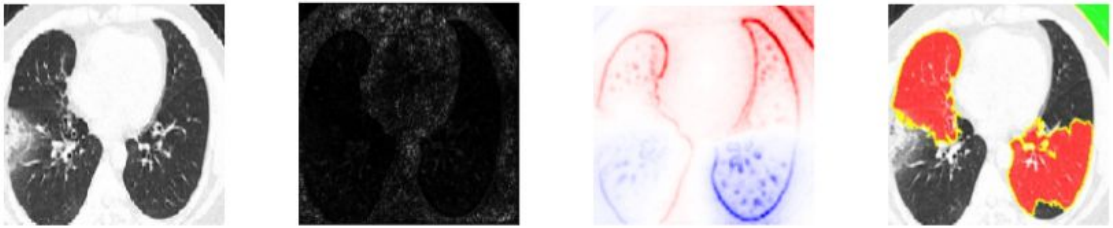
XAI : Method comparison - Predicted class

Dataset - Model

- 349 Covid-19 CT
- 297 Normal CT

VGG-16
Classification

89% test accuracy



(a) Unsegmented Covid-19 image (b) Integrated Gradients (c) LRP PresetAFlat (d) LIME Proxy Model

Biases

Focus outside lungs (integrated)

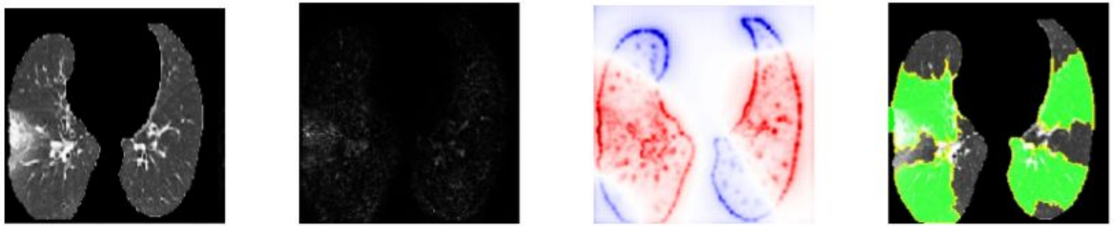
Top right Corner Bias

Lung Segmentation preprocessing

- 233 Covid-19 CT
- 293 Normal CT

VGG-16
Classification

70% Test Accuracy



(a) Segmented Covid image (b) Integrated Gradients (c) LRP PresetAFlat (d) LIME Proxy Model

Unsegmented and Segmented Covid-19 CT Scan explained with two VGG-16 model

Focus inside lungs (integrated)

No apparent bias

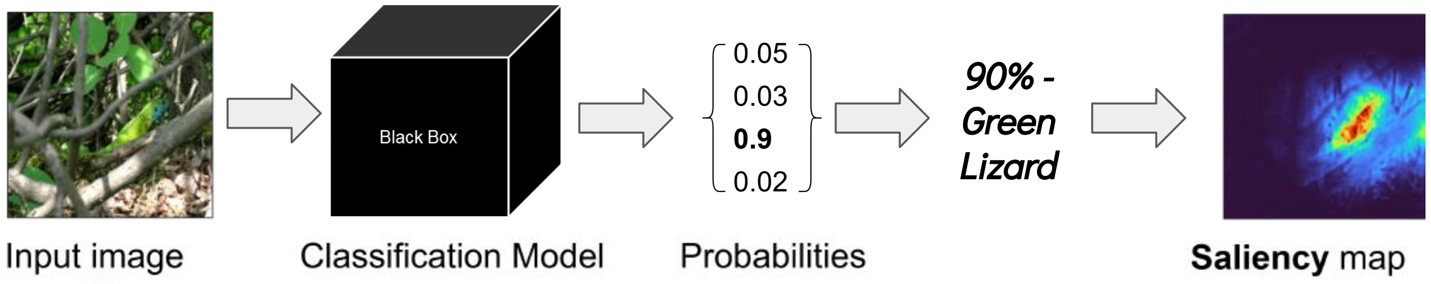
XAI helps to verify the model learning

POLYTECH MONS

04

INTRODUCTION – EVALUATION OF EXPLAINABILITY METHODS

04 Evaluation of Explainability Methods



How to evaluate XAI methods ?

04 Evaluation **metrics**

Deletion → does the accuracy drop?

Faithfulness metrics :

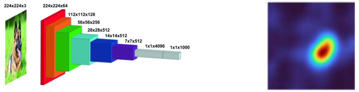
[Bhatt et al., 2020] [Petsiuk et al., 2018] ...

- Measure how explanations follow the **predictive** behavior of the model

Randomization → Is the saliency map modified ?

Randomization metrics :

[Sixt et al., 2020] ...



- Measure change in saliency map as a function of **parameter randomization**

Small perturbation → Is the saliency map stable ?

Robustness metrics :

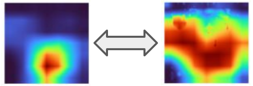
[Yeh et al., 2019] ...

- Measure the **stability** of explanations with respect to small input **perturbations**

Is the saliency map concise ?

Complexity metrics :

[Chalasani et al., 2020] ...

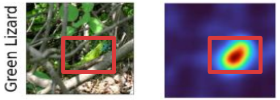


- Measure explanation **conciseness**

Does it fit into the bounding box ?

Localization metrics :

[Zhang et al., 2018] ...



- Measure whether explanations fit into the delimitation of a region of interest (e.g. bounding-boxes)

ADL - Intro to XAI

XAI is a **valuable tool** to:

- Detect biases
- Guide model selection
- Understand the bias source
- Verify the model learning

XAI methods can be evaluated:

- Visually (subjective !)
- Through metrics (properties!)

Gradient-based **Methods**

Gradients

[Simonyan et al., 2013]

Interpretation

“Which pixels need to be changed the least to affect the class score the most”

Gradient-based Methods

SmoothGrad - Integrated Gradients

[Smilkov et al., 2017] [Sundarajan et al., 2017]

SmoothGrad

" Removing noise by adding noise "

- Image ***x*** - class ***c*** - ***n*** samples - noise parameters
- Add ***Gaussian noise*** to ***n*** samples of the image
- Calculate the gradient of the class for each noisy image
- Average the gradients

$$\hat{M}_c(x) = \frac{1}{n}\sum_{1}^{n} M_c(x + \mathcal{N}(0, \sigma^2))$$

$M_c(x) = \partial S_c(x)/\partial x$

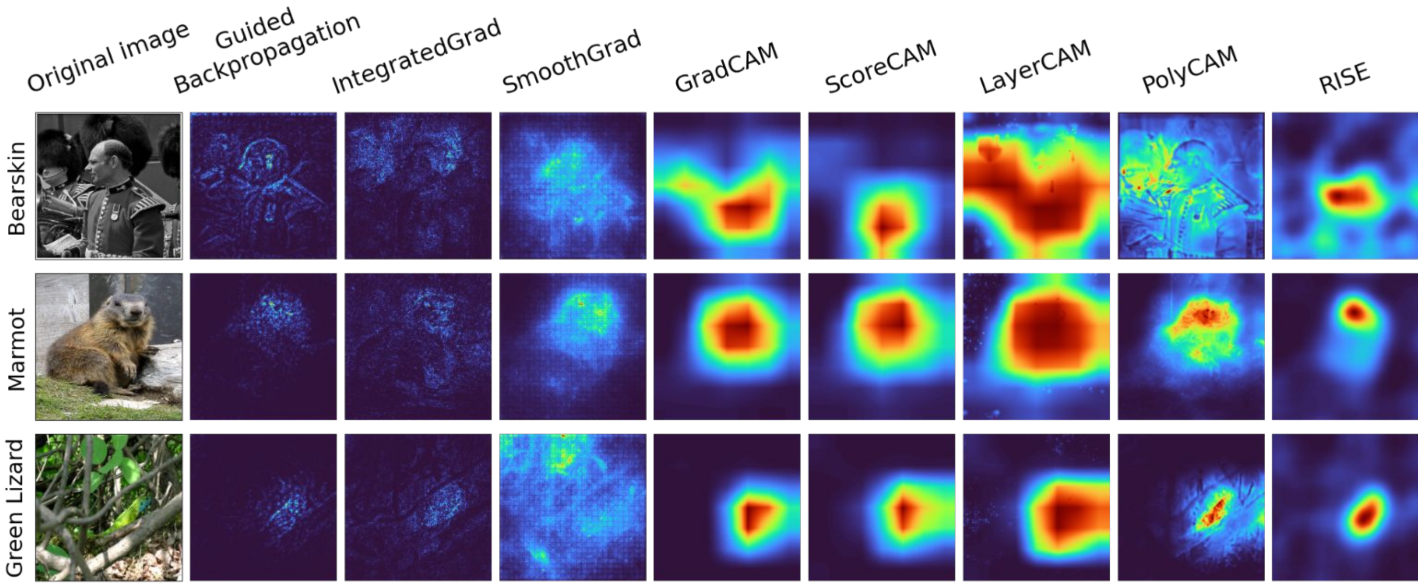
Integrated Gradients

"Path from baseline to input"

- Image ***x*** - baseline ***x'*** - number of steps ***n***
- *Create* ***n*** images - ranging linearly from the ***baseline to the input image***
- *Calculate the gradient* for each image
- *Average* all calculated gradients, and multiply by (input-baseline)

$$\text{IntegratedGrads}_i(x) ::= (x_i - x'_i) \times \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i} d\alpha$$

04 Evaluation of Explainability Methods



03b COVID-19 CLASSIFICATION

Detected biases - Summary

Dataset	Samples	Classes	Biases
Covid-19: X-ray [30]	2,621	3	Letters; Artefacts; Camera Positioning; Arm Positioning; Patient Characteristic (children)
Covid-19: CT Scan [143]	646 / 556	2	Out of lung detection
ChestXRay2017 [64]	5,856	2	Letters; Artefacts; Camera Positioning; Arm Positioning; Patient Characteristic (children)
ChestX-ray14 [134]	112,120	14	Letters; Artefacts; Horizontal Lines; Vertical Lines; Rotation;

Table 4.2: Summary of biases detected in each dataset

- Models alone, without XAI **cannot** be trusted based on predictions