Weill Cornell Medicine

ISBI 2024
ATHENS, GREECE

Introduction to Explainable Artificial Intelligence (XAI) in biomedical imaging and radiology

Tutorial 5: Explainable AI in Biomedical Imaging
21st International Symposium on Biomedical Imaging (ISBI)

Nicolas A. Karakatsanis, PhD, DABSNM, SM'IEEE
Assistant Professor of Biomedical Engineering
Department of Radiology, Weill Cornell Medical College
Adjunct Assistant Professor in Radiology
Biomedical Engineering & Imaging Institute,
Icahn School of Medicine at Mount Sinai

Disclosures Statement

Nothing to disclose

Lecture Scope/Objective

- **Taxonomy on Explainable AI (XAI) methods**
  - for biomedical imaging including basic and clinical research in radiology and nuclear medicine

- **Challenges and Opportunities of XAI applications**
  - for wide clinical adoption and impactful outcome in global healthcare

XAI = transparency & trustworthiness for AI models

XAI biomedical imaging technologies can be a major factor for enabling the translation of AI benefits in radiology and global healthcare

Introduction to Explainable Artificial Intelligence (XAI) in biomedical imaging: taxonomy, clinical applications & future perspectives

Deep learning (DL)

- has demonstrated a remarkable performance in diagnostic imaging for various diseases and modalities
- therefore, has a high potential to be used as a clinical tool.

AI models are usually evaluated in terms of predictive performance, e.g., classification accuracy.

- yet, performance metrics do not capture whether the evaluated model is right for the right reasons
- ML models can replicate biases and other confounding patterns from the input data when these are discriminative
  - example: COVID19 detection found to rely on markers, edges, arrows and other clinically irrelevant annotations
  - “Clever Hans” behavior results in a classifier’s right decision but for the *wrong* reasons (*shortcut learning*)

Moreover, current practice shows low deployment of these algorithms in clinical practice,

- because DL algorithms lack transparency and trust due to their underlying black-box mechanism.

For successful wide deployment, explainable artificial intelligence (XAI) could be introduced

- to close the gap between the medical professionals and the DL algorithms.

In this lecture,

- introduce XAI in the context of biomedical imaging and XAI methods taxonomy
- present example application for magnetic resonance (MR), computed tomography (CT), and positron emission tomography (PET) imaging
- Discuss challenges, opportunities and future suggestions to maximize clinical benefit in radiology

Introduction to Explainable Artificial Intelligence (XAI) in biomedical imaging: taxonomy, clinical applications & future perspectives

Computer-aided diagnostics (CAD) using deep learning (DL)

- widely used in diagnostic imaging for various diseases and modalities
- similar or superior in comparison to medical professional aided diagnostics
- great potential to be introduced in clinical workflow

However, despite promising results, DL not widely deployed in clinical practice yet.

- unlike simpler machine learning (ML), DL not require manual extraction of features depending on VOI
- DL extract features in unsupervised way, without a priori defined assumptions/regulations

Explainability = understanding of DL model decisions and reasoning

DL adoptability requires efficient learning & explainability to work in synergy

Despite DL's superior learning capabilities, they lack transparency due to underlying black-box mechanism.

- DL are difficult to validate, i.e., determine which features trigger model decision,
- lack trustworthiness which severely limits their (clinical) deployment

Introduction to Explainable Artificial Intelligence (XAI) in biomedical imaging: taxonomy, clinical applications & future perspectives

DL transparency should be improved to close this gap of adoptability due to AI opacity (black-box)

- provide medical professionals/stakeholders with a pragmatic explanation of the model's decisions

Explainable artificial intelligence (XAI) can mitigate AI opacity

- XAI's attribution (i.e., feature importance) methods provide information on why a decision is made
- thus, allowing to back-propagate the models decision to target specific attributions present in the image.

XAI in biomedical imaging may be potentially used as a new imaging biomarker (IB) in clinic

- function as an indicator of normal and/or pathogenic biological processes,
- complement medical professionals in medical decision-making.
- monitor and assess responses to an exposure or (therapeutic) intervention.

XAI should provide transparency about quality/legibility of its decision, explanation, and associated errors.

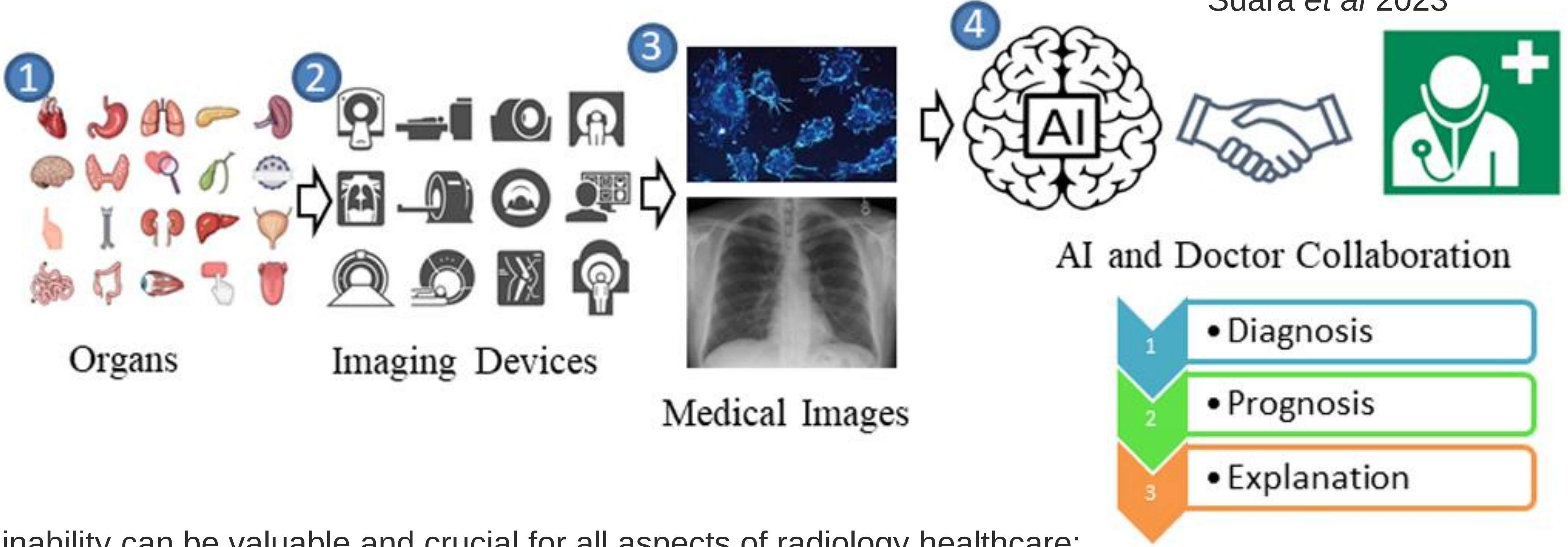
- XAI must therefore cross "translational gaps: technical and clinical validation and cost-effectiveness
- the new European Medical Device Regulation (EU MDR) endorses strict transparency regulations

XAI crucial for the more transparent, ethical (unbiased) safe and trustworthy deployment of DL algorithms in clinical practice,

but better understanding of current practice is first required

Explainability in biomedical imaging and clinical radiology/nuclear medicine practice

Suara *et al* 2023



Explainability can be valuable and crucial for all aspects of radiology healthcare:

- physicians/scientists/technologists
- patients’ community
- healthcare system/providers administration
- developers of AI models and their XAI component

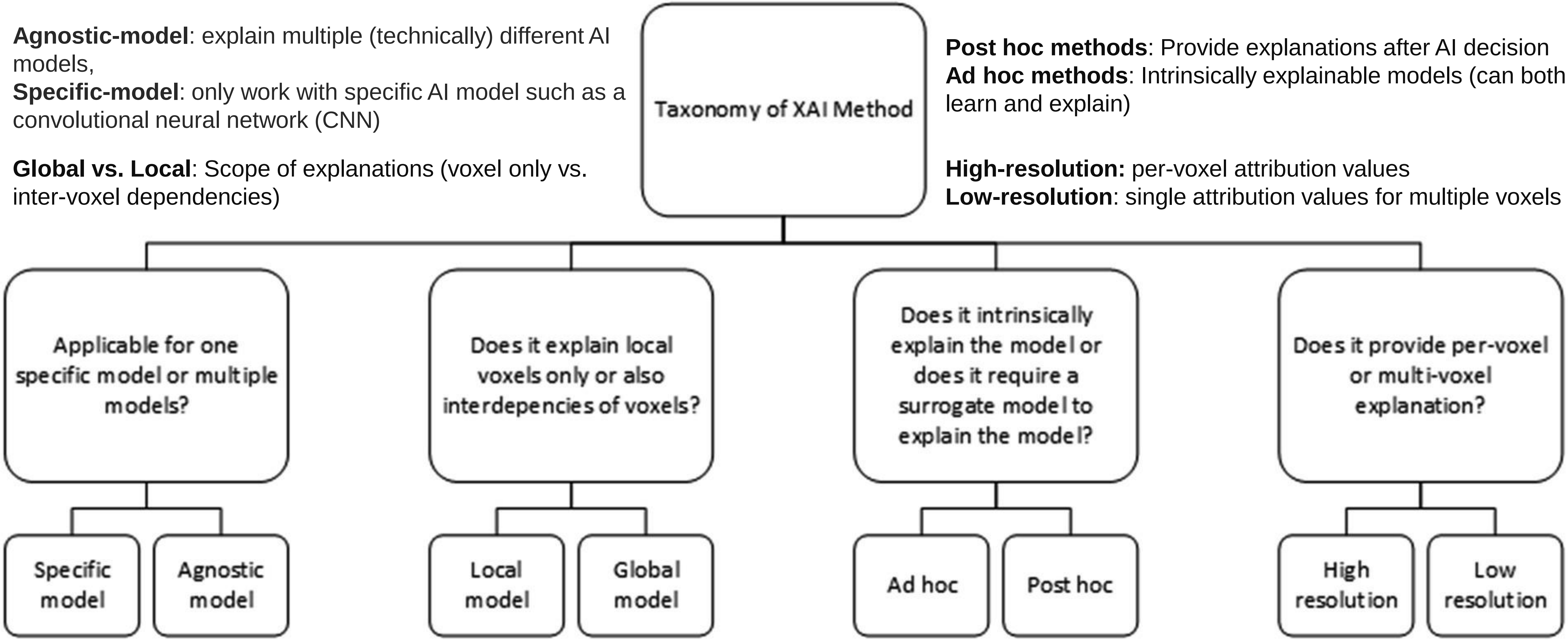
XAI Methods Taxonomy

Agnostic-model: explain multiple (technically) different AI models,
Specific-model: only work with specific AI model such as a convolutional neural network (CNN)

Global vs. Local: Scope of explanations (voxel only vs. inter-voxel dependencies)

Post hoc methods: Provide explanations after AI decision
Ad hoc methods: Intrinsically explainable models (can both learn and explain)

High-resolution: per-voxel attribution values
Low-resolution: single attribution values for multiple voxels


Taxonomy of XAI Method
Applicable for one specific model or multiple models?
Does it explain local voxels only or also interdepencies of voxels?
Does it intrinsically explain the model or does it require a surrogate model to explain the model?
Does it provide per-voxel or multi-voxel explanation?
Specific model
Agnostic model
Local model
Global model
Ad hoc
Post hoc
High resolution
Low resolution

de Vries *et al* 2023

Post-hoc XAI methods

Post hoc analysis

- provides model explanation after the decision (e.g. classification)
  - an AI model that is able to learn,
  - but requires an additional model to provide an explanation.

The majority (~75%) of DL algorithms supporting XAI in biomedical imaging employ *post hoc* XAI methods

- due to its wide availability and
- its plug-and-play deployment

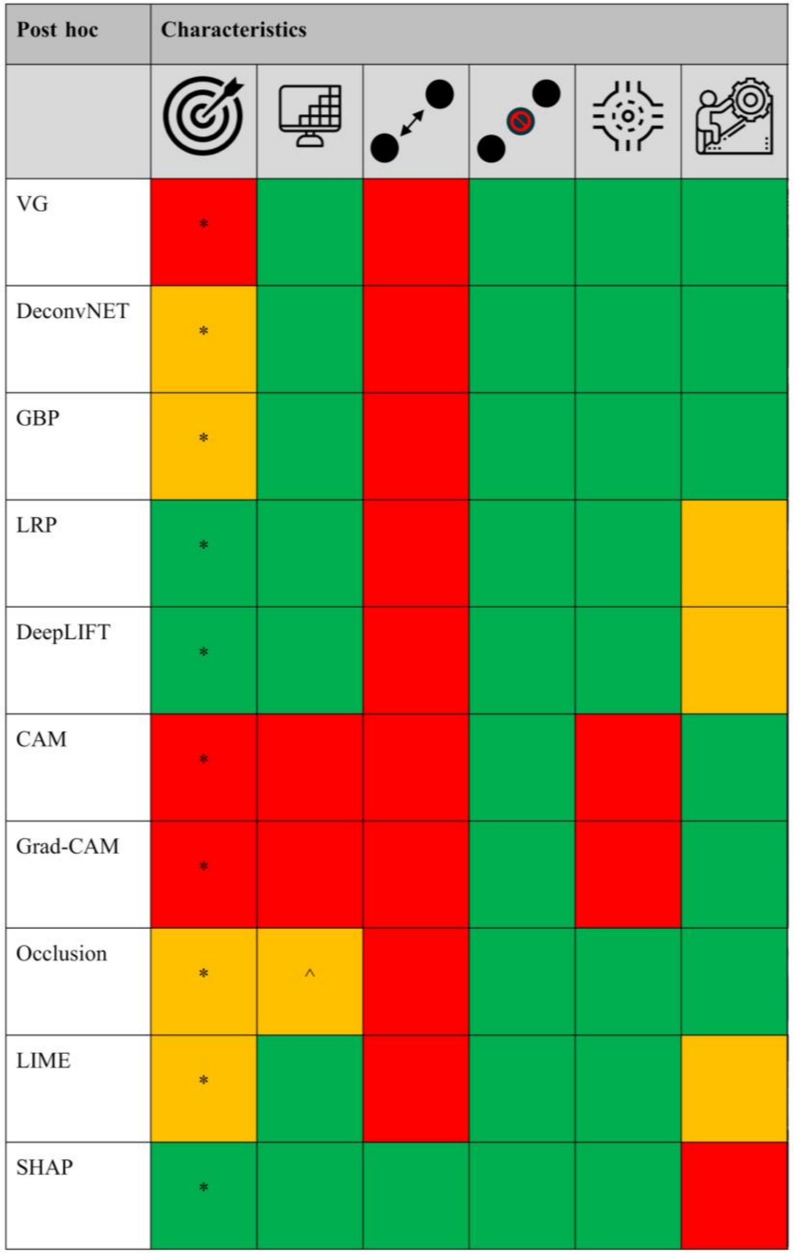
P*ost-hoc* XAI methods can be divided here into

- gradient-propagation methods,
- perturbation methods
- also, segmentation & radiomic methods will be briefly discussed

Table at left: overview of post hoc XAI methods performance

- scores [low/no (red), average (orange), and high/yes (green) performance] based on
- target specificity, spatial-resolution and local/global voxel dependency capability, model agnosticism, and technical simplicity

Overview of the *post hoc* attribution methods



Post hoc	Characteristics					
	🎯	🖥	↔	⊘	⌖	⚙
VG	🟥 *	🟩	🟥	🟩	🟩	🟩
DeconvNET	🟧 *	🟩	🟥	🟩	🟩	🟩
GBP	🟧 *	🟩	🟥	🟩	🟩	🟩
LRP	🟩 *	🟩	🟥	🟩	🟩	🟧
DeepLIFT	🟩 *	🟩	🟥	🟩	🟩	🟧
CAM	🟥 *	🟥	🟥	🟩	🟥	🟩
Grad-CAM	🟥 *	🟥	🟥	🟩	🟥	🟩
Occlusion	🟧 *	🟧 ^	🟥	🟩	🟩	🟩
LIME	🟧 *	🟩	🟥	🟩	🟩	🟧
SHAP	🟩 *	🟩	🟩	🟩	🟩	🟥

de Vries *et al* 2023

Post-hoc XAI: Gradient-propagation approaches

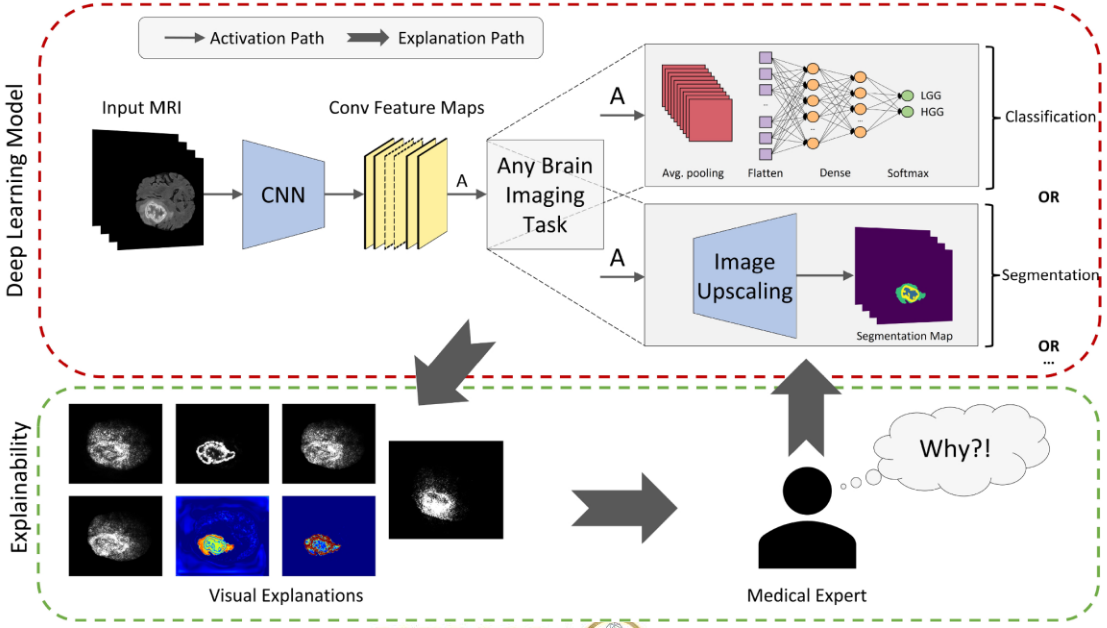
Vanilla gradient (VG)

post-hoc XAI method that create an attribution map by calculating gradients over the layers using a single forward and backward propagation, i.e.:

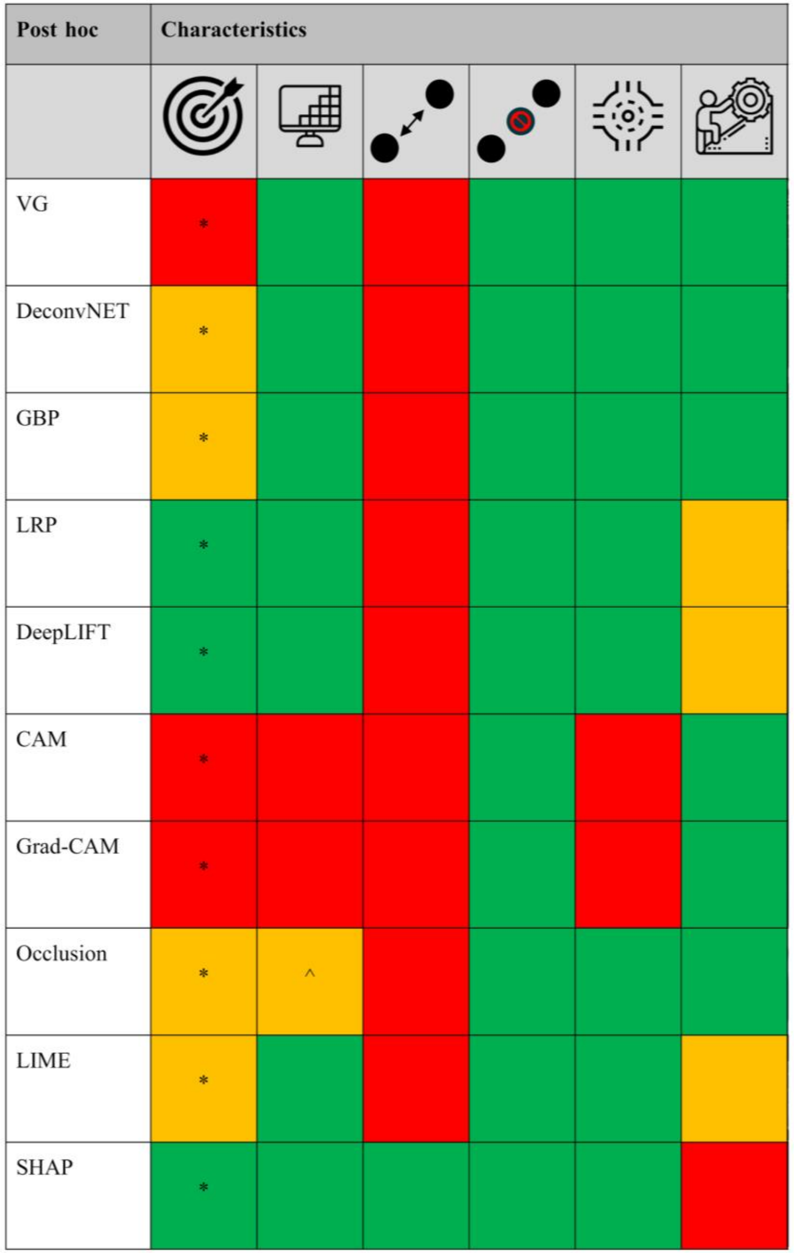
- the input image is fed into the AI model calculating an output score (forward),
- subsequently the dependence (gradient) between the neurons/convolution layers and the output is calculated (backward) to create an attribution map.

$P_c(X^I)$: prediction of class c, computed by the classification CNN layer for input image X^I.
VG objective: find the L_2-regularized image, exhibiting maximum P_c, (λ = regularization term):

$$VG = argmax_c P_c\left(X^I\right) - \lambda \|X^I\|_2^2$$



Overview of the *post hoc* attribution methods

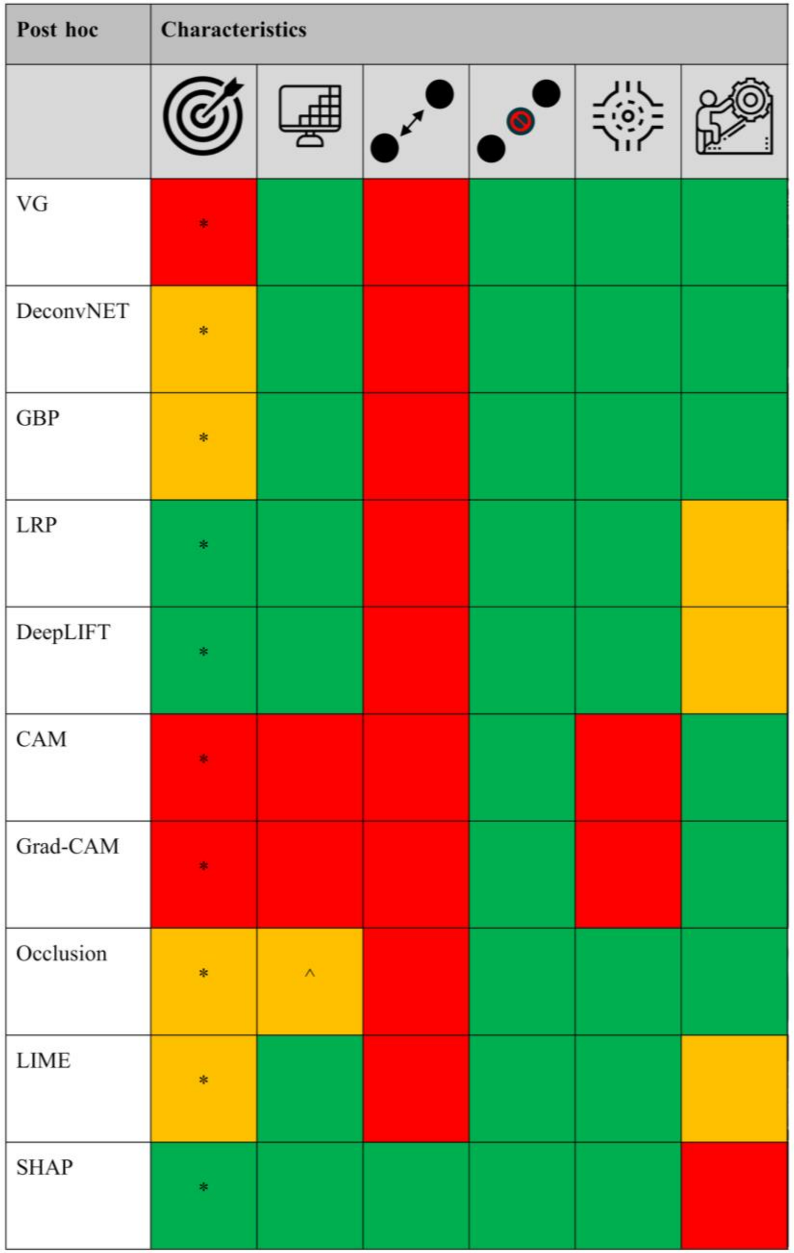


Post hoc	Characteristics					
	🎯	🖥	↔	⊘	⊙	⚙
VG	🟥 *	🟩	🟥	🟩	🟩	🟩
DeconvNET	🟨 *	🟩	🟥	🟩	🟩	🟩
GBP	🟨 *	🟩	🟥	🟩	🟩	🟩
LRP	🟩 *	🟩	🟥	🟩	🟩	🟨
DeepLIFT	🟩 *	🟩	🟥	🟩	🟩	🟨
CAM	🟥 *	🟥	🟥	🟩	🟥	🟩
Grad-CAM	🟥 *	🟥	🟥	🟩	🟥	🟩
Occlusion	🟨 *	🟨 ^	🟥	🟩	🟩	🟩
LIME	🟨 *	🟩	🟥	🟩	🟩	🟨
SHAP	🟩 *	🟩	🟩	🟩	🟩	🟥

de Vries *et al* 2023

Post-hoc XAI: Gradient-propagation approaches

Overview of the *post hoc* attribution methods



Post hoc	Characteristics					
	🎯	🖥	↔	🚫	⌖	⚙
VG	🟥 *	🟩	🟥	🟩	🟩	🟩
DeconvNET	🟨 *	🟩	🟥	🟩	🟩	🟩
GBP	🟨 *	🟩	🟥	🟩	🟩	🟩
LRP	🟩 *	🟩	🟥	🟩	🟩	🟨
DeepLIFT	🟩 *	🟩	🟥	🟩	🟩	🟨
CAM	🟥 *	🟥	🟥	🟩	🟥	🟩
Grad-CAM	🟥 *	🟥	🟥	🟩	🟥	🟩
Occlusion	🟨 *	🟨 ^	🟥	🟩	🟩	🟩
LIME	🟨 *	🟩	🟥	🟩	🟩	🟨
SHAP	🟩 *	🟩	🟩	🟩	🟩	🟥

de Vries *et al* 2023

Vanilla gradient (VG)

post-hoc XAI method that create an attribution map by calculating gradients over the layers using a single forward and backward propagation, i.e.:

- the input image is fed into the AI model calculating an output score (forward),
- subsequently the dependence (gradient) between the neurons/convolution layers and the output is calculated (backward) to create an attribution map.

VG pros & cons

- simple intuitive attribution method with low computational power
- generates noisy attribution maps
- suffers from gradient saturation, i.e., change in a neuron does not affect the output of the network and therefore cannot be measured
- lacks the ability to differentiate between classes (e.g., healthy vs. disease) suggesting inability to generate clear and class-discriminative attribution maps

Example 1: NeuroXAI attribution-based framework compared VG and other attribution-based visualization methods for MRI analysis of brain tumors

- VG exhibited noisy attribution maps and gradient saturation both for classification and segmentation feature visualization.

Example 2: Predict contrast enhancement phase from CT images

- similar results were seen using VG for feature visualization

Post-hoc XAI: Gradient-propagation approaches

DeconvNET

effectively equivalent of VG apart from the way it calculates the gradient over a Rectified Linear Unit (ReLU) function i.e.,

- a linear function that will output only positive input values

designed to work similar to convolutional networks but reverse

- reversing pooling component, reversing filter component etc.

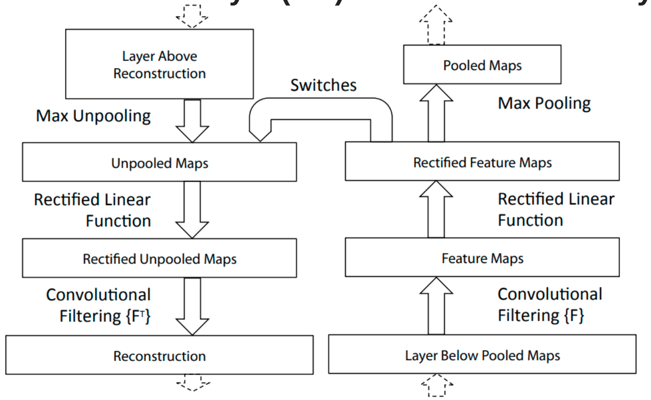
deconvolutional approach to explaining a model:

- not training a DeconvNet but rather probe our CNN with it

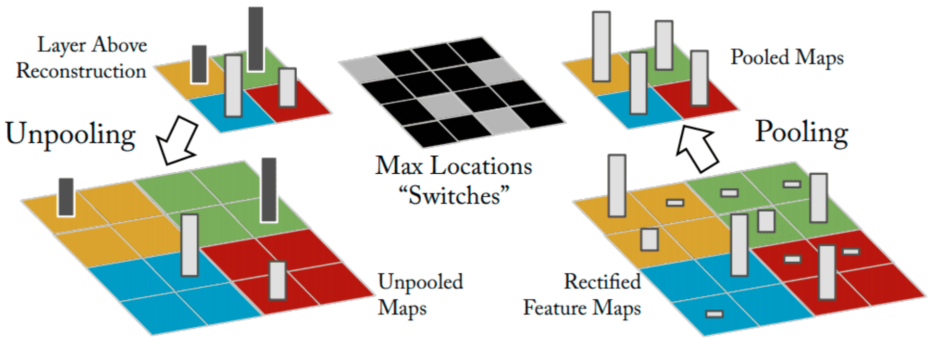
To reconstruct the activation on a specific layer:

- attach **Deconv layers** to corresponding **CNN layers**
- then an image is passed through the CNN, and the network computes the output.
- to examine a reconstruction for a given class *c*:
    - set all activations except the one responsible for predicting class *cc* to zero.
    - then propagate through DeconvNet layers & pass all feature maps as inputs to corresponding layers

A DeconvNet layer (left) attached to a CNN layer (right)

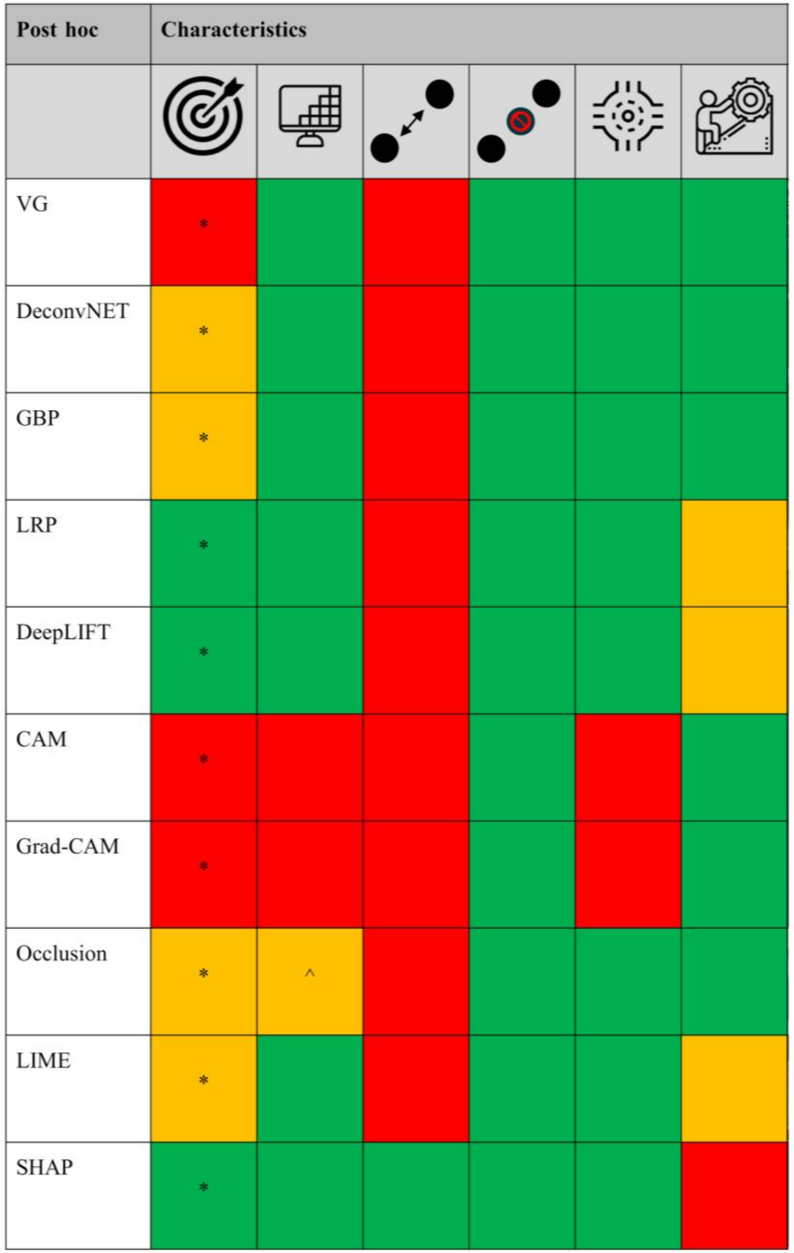


the un-pooling layer



Zeiler *et al* 2013

Overview of the *post hoc* attribution methods



Post hoc	Characteristics					
VG	🟥 *	🟩	🟥	🟩	🟩	🟩
DeconvNET	🟨 *	🟩	🟥	🟩	🟩	🟩
GBP	🟨 *	🟩	🟥	🟩	🟩	🟩
LRP	🟩 *	🟩	🟥	🟩	🟩	🟨
DeepLIFT	🟩 *	🟩	🟥	🟩	🟩	🟨
CAM	🟥 *	🟥	🟥	🟩	🟥	🟩
Grad-CAM	🟥 *	🟥	🟥	🟩	🟥	🟩
Occlusion	🟨 *	🟨 ^	🟥	🟩	🟩	🟩
LIME	🟨 *	🟩	🟥	🟩	🟩	🟨
SHAP	🟩 *	🟩	🟩	🟩	🟩	🟥

de Vries *et al* 2023

Post-hoc XAI: Gradient-propagation approaches

DeconvNET

effectively equivalent of VG apart from the way it calculates the gradient over a Rectified Linear Unit (ReLU) function i.e.,

- a linear function that will output only positive input values

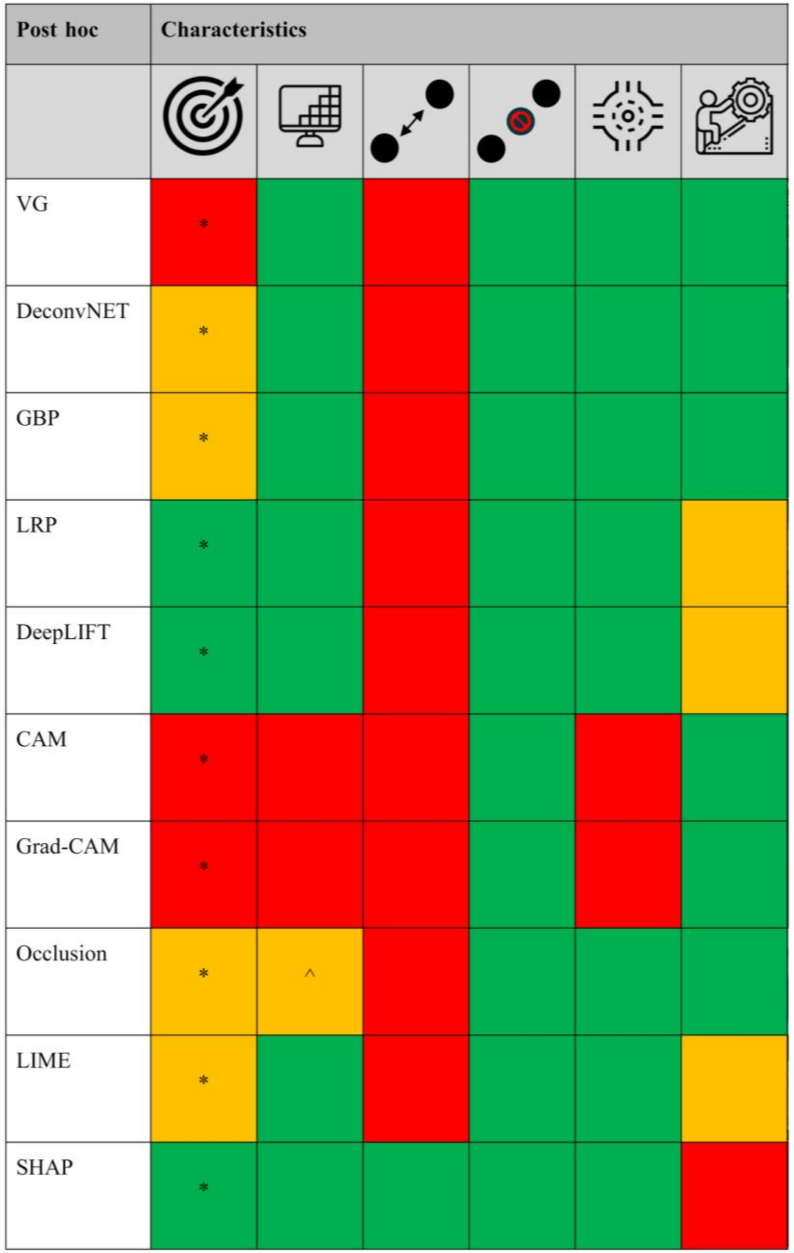
DeconvNET pros & cons

- helps with improving model convergence during model training
- more human-interpretable results by focusing on more specific regions
- absence of pooling layers results in non-image-specific explanations
- less informative attribution maps (heatmaps) due to rejection of negative pieces of evidence during backpropagation

Example: TorchEsegeta, a framework for interpretable and explainable image-based DL algorithms, compared multiple attribution methods for

- interlayer CNN visualization in blood vessel segmentation in human brain
- VG and DeconvNET provided more human-interpretable results than the other attribution methods (e.g., DeepLIFT and GradCAM++),
    - as they mainly focused on the vessels, while other methods also showed non-vessel activation.

Overview of the *post hoc* attribution methods



Post hoc	Characteristics					
VG	🟥 *	🟩	🟥	🟩	🟩	🟩
DeconvNET	🟨 *	🟩	🟥	🟩	🟩	🟩
GBP	🟨 *	🟩	🟥	🟩	🟩	🟩
LRP	🟩 *	🟩	🟥	🟩	🟩	🟨
DeepLIFT	🟩 *	🟩	🟥	🟩	🟩	🟨
CAM	🟥 *	🟥	🟥	🟩	🟥	🟩
Grad-CAM	🟥 *	🟥	🟥	🟩	🟥	🟩
Occlusion	🟨 *	🟨 ^	🟥	🟩	🟩	🟩
LIME	🟨 *	🟩	🟥	🟩	🟩	🟨
SHAP	🟩 *	🟩	🟩	🟩	🟩	🟥

de Vries *et al* 2023

Post-hoc XAI: Gradient-propagation approaches

Guided back-propagation (GBP): incorporates both VG and DeconvNet

VG accuracy is limited due to the flow of negative gradients

- they decrease the accuracy of the visualization of the higher layers

GBP approach: combine VG and DeconvNet

- by adding “guide” to the VG with the help of deconvolution

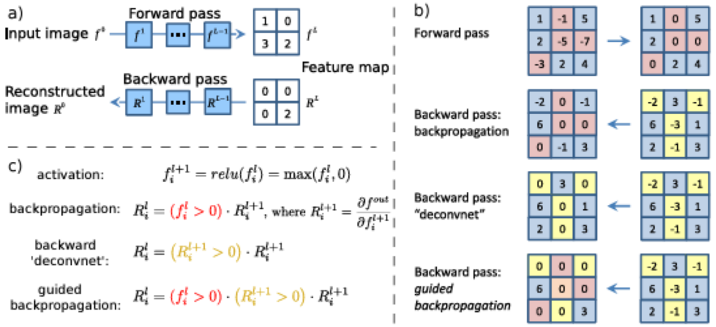
DeconvNet: when computing values at the *Rectification* component

- all non-positive values are masked with the *ReLU*.
- in that layer, the computed values are calculated only base on the top signal (reconstruction from the upper layer), and the input is ignored.

VG: focuses instead on the gradient values computed based on the input image.

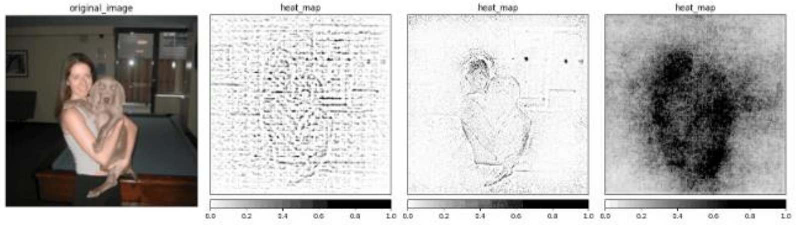
GBP: take DeconvNet masking of the *Rectification* layer and apply it on VG gradient values

- thus remove noise caused by the negative gradient values.
- this noise removal is the reason GBP’s prefix “guided”.
- Deconvolution guides backpropagation of VG to produce sharper attribution maps



a)
Forward pass
Input image f^0
f^1
f^{L-1}
1 0 / 3 2
f^L
Feature map
Backward pass
Reconstructed image R^0
R^1
R^{L-1}
0 0 / 0 2
R^L
b)
Forward pass
1 -1 5 / 2 -5 -7 / -3 2 4
1 0 5 / 2 0 0 / 0 2 4
Backward pass: backpropagation
-2 0 -1 / 6 0 0 / 0 -1 3
-2 3 -1 / 6 -3 1 / 2 -1 3
Backward pass: "deconvnet"
0 3 0 / 6 0 1 / 2 0 3
-2 3 -1 / 6 -3 1 / 2 -1 3
Backward pass: guided backpropagation
0 0 0 / 6 0 0 / 0 0 3
-2 3 -1 / 6 -3 1 / 2 -1 3
c)
activation: $f_i^{l+1} = relu(f_i^l) = \max(f_i^l, 0)$
backpropagation: $R_i^l = (f_i^l > 0) \cdot R_i^{l+1}$, where $R_i^{l+1} = \frac{\partial f^{out}}{\partial f_i^{l+1}}$
backward 'deconvnet': $R_i^l = (R_i^{l+1} > 0) \cdot R_i^{l+1}$
guided backpropagation: $R_i^l = (f_i^l > 0) \cdot (R_i^{l+1} > 0) \cdot R_i^{l+1}$

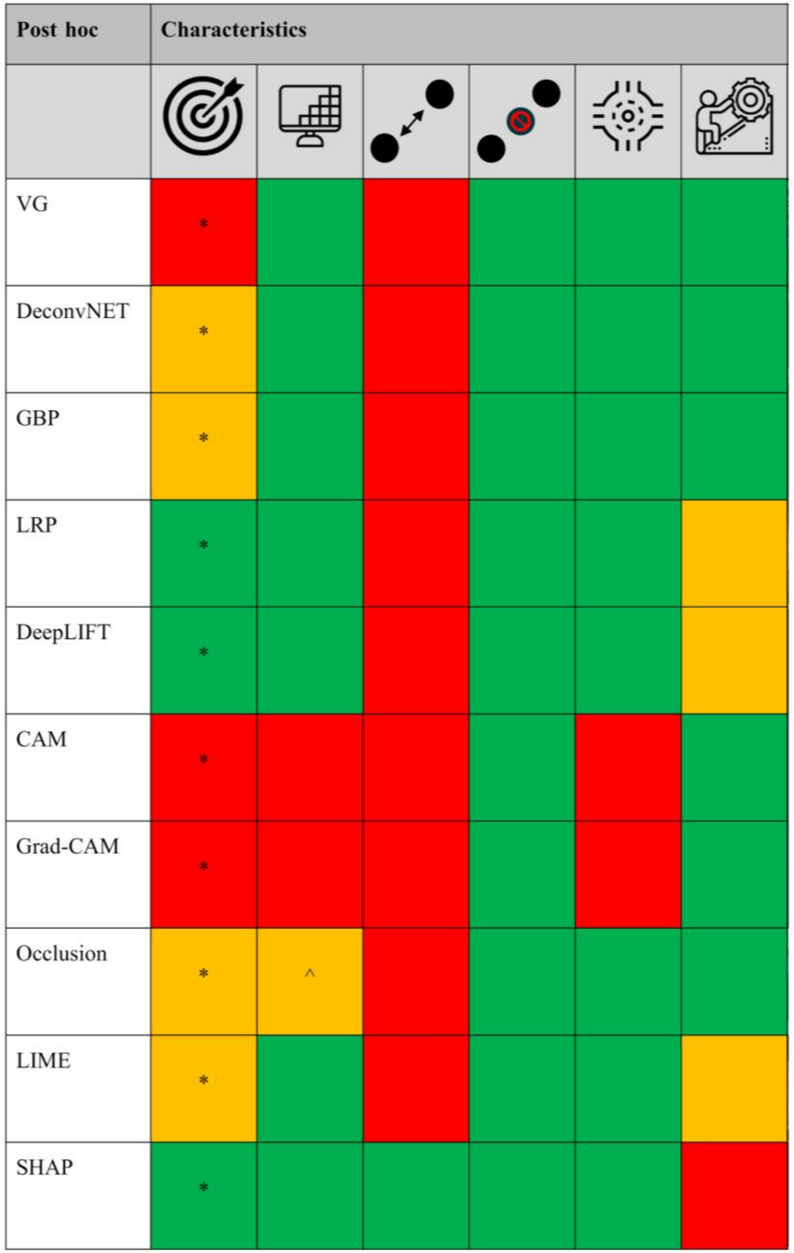
Springengberg et al. 2014,



(a) Original Image. (b) Deconvolution results. (c) GBP results. (d) Saliency results.

Erdem, (Feb 2022).
"XAI Methods - Guided Backpropagation"

Overview of the *post hoc* attribution methods



Post hoc	Characteristics					
	🎯	🖥	↔	⊘	⌖	⚙
VG	🟥 *	🟩	🟥	🟩	🟩	🟩
DeconvNET	🟨 *	🟩	🟥	🟩	🟩	🟩
GBP	🟨 *	🟩	🟥	🟩	🟩	🟩
LRP	🟩 *	🟩	🟥	🟩	🟩	🟨
DeepLIFT	🟩 *	🟩	🟥	🟩	🟩	🟨
CAM	🟥 *	🟥	🟥	🟩	🟥	🟩
Grad-CAM	🟥 *	🟥	🟥	🟩	🟥	🟩
Occlusion	🟨 *	🟨 ^	🟥	🟩	🟩	🟩
LIME	🟨 *	🟩	🟥	🟩	🟩	🟨
SHAP	🟩 *	🟩	🟩	🟩	🟩	🟥

de Vries *et al* 2023

Post-hoc XAI: Gradient-propagation approaches

Guided back-propagation (GBP): incorporates both VG and DeconvNet

VG accuracy is limited due to the flow of negative gradients

- they decrease the accuracy of the visualization of the higher layers

GBP approach: combine VG and DeconvNet

- by adding "guide" to the VG with the help of deconvolution

GBP pros & cons

- fewer activated voxels and therefore less noisy attribution maps than VG or DeconvNET alone
- overly sparse attribution maps, not useful for complete image characterization

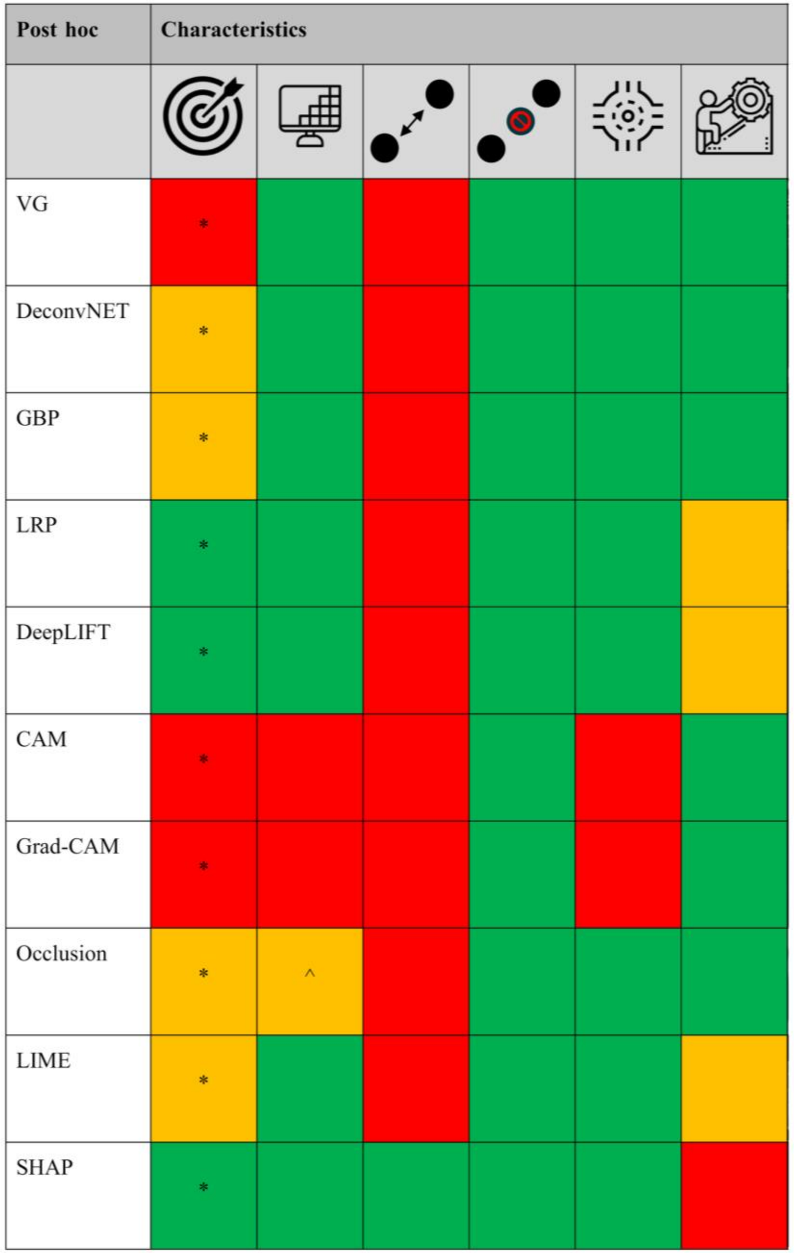
Example 1: In the NeuroXAI framework:

- GBP showed target specific attribution maps with indeed less noise than VG

Example 2: study for predicting brain abnormalities using MRI:

- an additional smoothing function to the GBP proposed to suppress the amount of noise and the effect of non-target specific attributions even more
- GBP attribution maps showed low noise and accurate localization of a range of morphological distinct abnormalities.
- However, GBP may also result in overly sparse attribution maps, not useful for complete image characterization

Overview of the *post hoc* attribution methods



Post hoc	Characteristics					
VG	🟥 *	🟩	🟥	🟩	🟩	🟩
DeconvNET	🟨 *	🟩	🟥	🟩	🟩	🟩
GBP	🟨 *	🟩	🟥	🟩	🟩	🟩
LRP	🟩 *	🟩	🟥	🟩	🟩	🟨
DeepLIFT	🟩 *	🟩	🟥	🟩	🟩	🟨
CAM	🟥 *	🟥	🟥	🟩	🟥	🟩
Grad-CAM	🟥 *	🟥	🟥	🟩	🟥	🟩
Occlusion	🟨 *	🟨 ^	🟥	🟩	🟩	🟩
LIME	🟨 *	🟩	🟥	🟩	🟩	🟨
SHAP	🟩 *	🟩	🟩	🟩	🟩	🟥

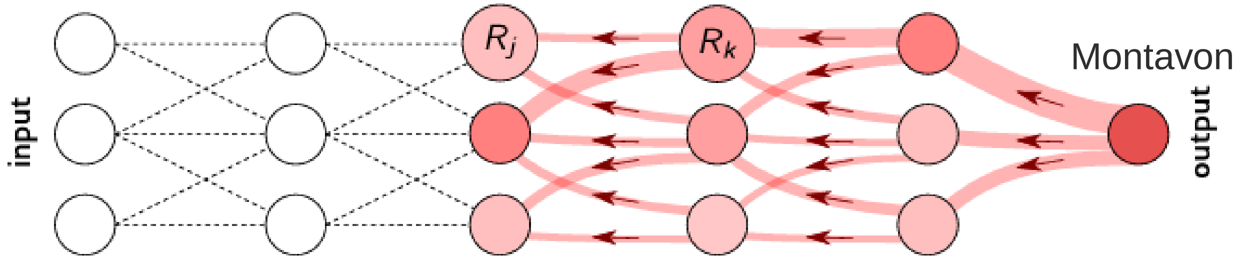
de Vries *et al* 2023

Post-hoc XAI: Scores propagation approaches

Layer-wise relevance propagation (LRP)

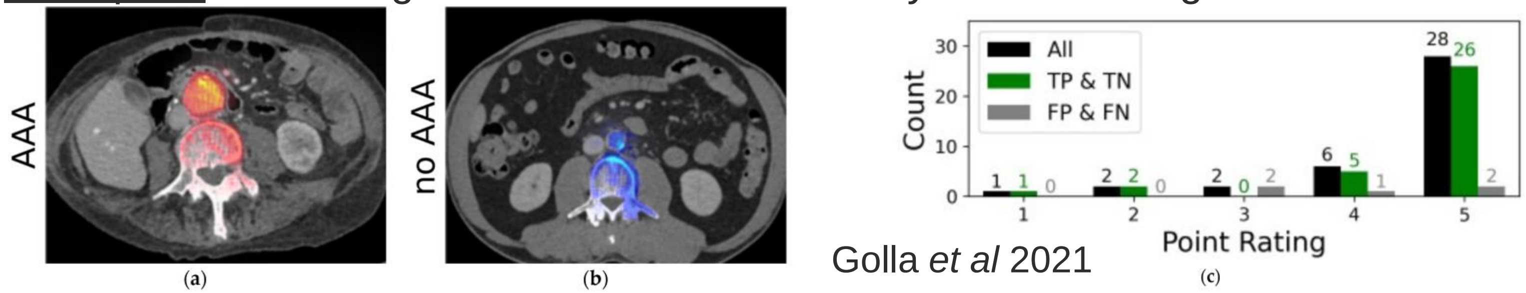
XAI method propagating the relevance (e.g. class) score backward over the neural layers to the input image using LRP specific rules/concept:

- LRP concept: conserve inter-neuron dependency, i.e., what has been received by a neuron layer will be redistributed to the following lower layer in equal quantity.



Montavon *et al* 2019

Example 1: screening of abdominal aortic aneurysm in CT images



(a) (b) (c)

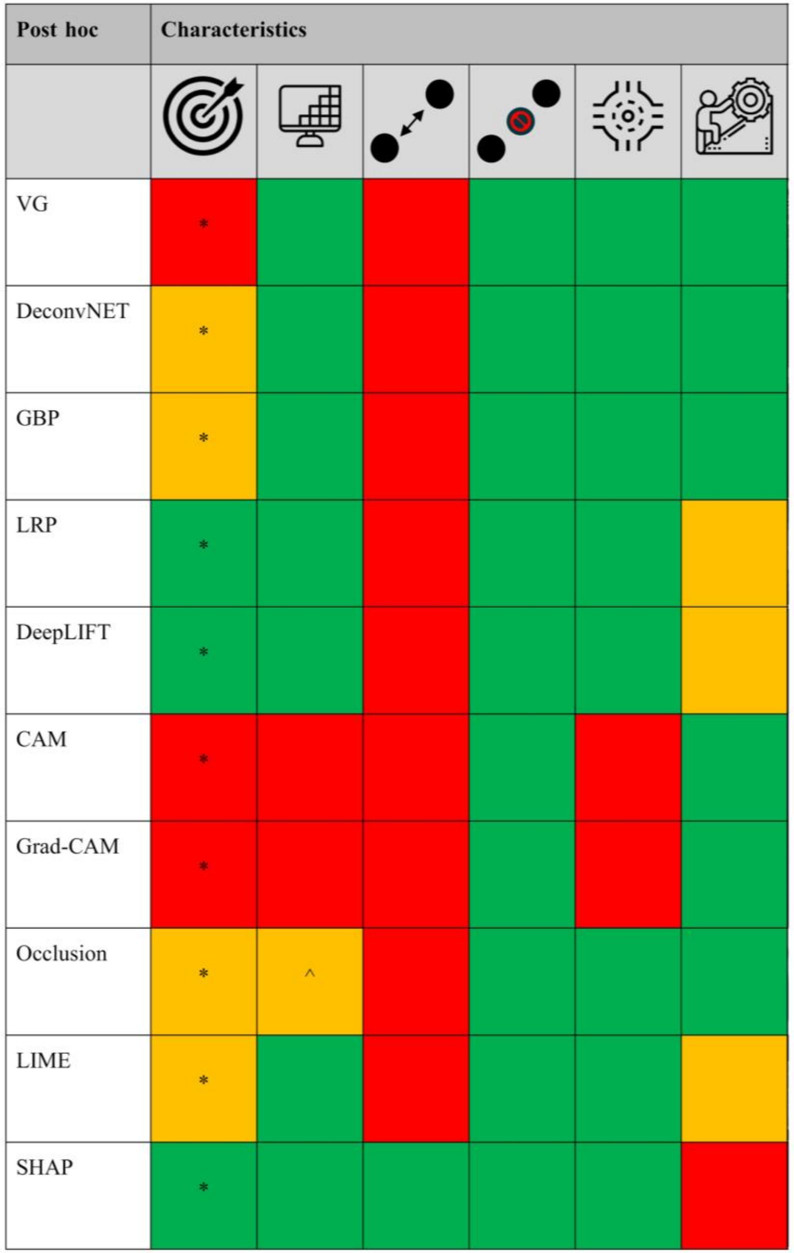
Golla *et al* 2021

Left image: relevance maps for predicting the AAA super-imposed on the CT images for two example cases.

- high relevance around the aorta confirms correctly learned classification based on the ROI.

Right image: score distribution of the assessment by an experienced radiologist

Overview of the *post hoc* attribution methods



Post hoc	Characteristics					
	🎯	🖥	↔	🚫	⊕	⚙
VG	🟥 *	🟩	🟥	🟩	🟩	🟩
DeconvNET	🟨 *	🟩	🟥	🟩	🟩	🟩
GBP	🟨 *	🟩	🟥	🟩	🟩	🟩
LRP	🟩 *	🟩	🟥	🟩	🟩	🟨
DeepLIFT	🟩 *	🟩	🟥	🟩	🟩	🟨
CAM	🟥 *	🟥	🟥	🟩	🟥	🟩
Grad-CAM	🟥 *	🟥	🟥	🟩	🟥	🟩
Occlusion	🟨 *	🟨 ^	🟥	🟩	🟩	🟩
LIME	🟨 *	🟩	🟥	🟩	🟩	🟨
SHAP	🟩 *	🟩	🟩	🟩	🟩	🟥

de Vries *et al* 2023

Post-hoc XAI:
Scores propagation approaches

Layer-wise relevance propagation (LRP)

XAI method propagating the relevance (e.g. class) score backward over the neural layers to the input image using LRP specific rules/concept:

- LRP concept: conserve inter-neuron dependency, i.e., what has been received by a neuron layer will be redistributed to the following lower layer in equal quantity.

LRP pros & cons

- decomposition is based on propagating relevance scores between the neurons instead of gradients thus, addressing the gradient saturation of VG/DeconvNET
- low specificity in attributions may be observed due to low DL model performance, biased input data or non-optimal LRP configuration (possibly >1 absolute reason)

Example 1: screening of abdominal aortic aneurysm in CT images

- LRP showed clear class difference based on activation difference in aortic lumen
- however, high activation for both classes seen in vertebra, indicating either:
  - DL model is biased,
  - DL model did not converge, the vertebra is a confounder, or
  - LRP also incorporates non-target specific features in its attribution map.

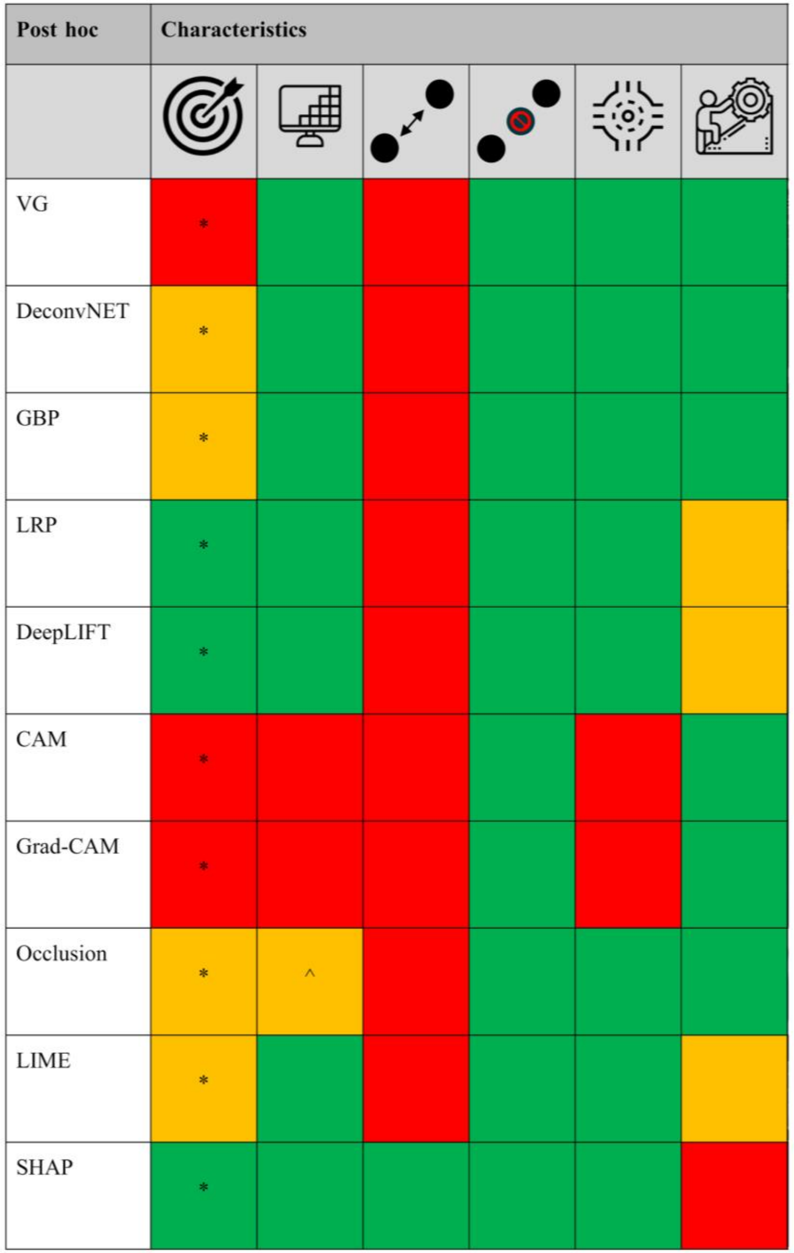
Example 2: COVID-19 classification,

- similar result: LRP was not able to visualize target-specific features

Example 4: however, other studies showed

- class-discriminative regions and precise localization of lesions using LRP

Overview of the *post hoc* attribution methods



Post hoc	Characteristics					
VG	🟥 *	🟩	🟥	🟩	🟩	🟩
DeconvNET	🟨 *	🟩	🟥	🟩	🟩	🟩
GBP	🟨 *	🟩	🟥	🟩	🟩	🟩
LRP	🟩 *	🟩	🟥	🟩	🟩	🟨
DeepLIFT	🟩 *	🟩	🟥	🟩	🟩	🟨
CAM	🟥 *	🟥	🟥	🟩	🟥	🟩
Grad-CAM	🟥 *	🟥	🟥	🟩	🟥	🟩
Occlusion	🟨 *	🟨 ^	🟥	🟩	🟩	🟩
LIME	🟨 *	🟩	🟥	🟩	🟩	🟨
SHAP	🟩 *	🟩	🟩	🟩	🟩	🟥

de Vries *et al* 2023

Post-hoc XAI:
Scores propagation approaches

Deep Learning Important Features (DeepLIFT)

XAI method adopts an “improved” LRP approach by employing a *neutral reference* activation (e.g., neuron activation of normal/healthy CT scan without pathology/ disease) to describe the change of a new neuron activation relative to this reference

- based on these differences, contribution scores are calculated for each neuron to compute an attribution map

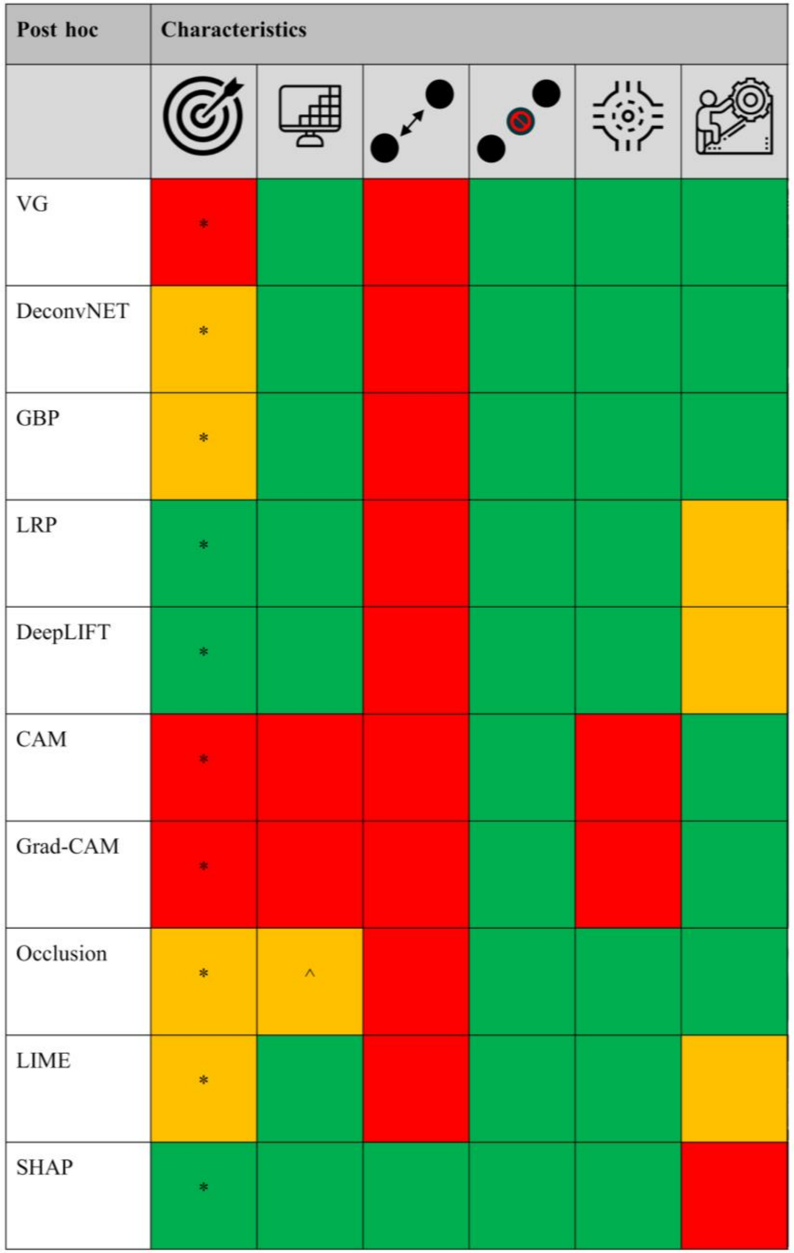
DeepLIFT pros & cons

- addresses the saturation problem of other XAI propagation approaches
- backpropagates the importance signal based on the network’s internal wiring; thus, two identical models with different internal wiring could, in principle, produce different attributions (i.e. not “implementation invariant”)

Example 1: identification of Multiple Sclerosis (MS) patients on MRI

- DeepLIFT was compared to LRP and VG by perturbating their three attribution maps for three VOIs.
- DeepLIFT performed slightly better than LRP and much better than VG in extracting target-specific features.
- Both LRP and DeepLIFT were able to tackle gradient saturation

Overview of the *post hoc* attribution methods



Post hoc	Characteristics					
VG	🟥 *	🟩	🟥	🟩	🟩	🟩
DeconvNET	🟨 *	🟩	🟥	🟩	🟩	🟩
GBP	🟨 *	🟩	🟥	🟩	🟩	🟩
LRP	🟩 *	🟩	🟥	🟩	🟩	🟨
DeepLIFT	🟩 *	🟩	🟥	🟩	🟩	🟨
CAM	🟥 *	🟥	🟥	🟩	🟥	🟩
Grad-CAM	🟥 *	🟥	🟥	🟩	🟥	🟩
Occlusion	🟨 *	🟨 ^	🟥	🟩	🟩	🟩
LIME	🟨 *	🟩	🟥	🟩	🟩	🟨
SHAP	🟩 *	🟩	🟩	🟩	🟩	🟥

de Vries *et al* 2023

Post-hoc XAI:
Class Activation Mapping

Class Activation Mapping (CAM)

XAI method adopting one of the most well-known model-specific attribution methods

- employs a Global Average Pooling (GAP) layer instead of multiple dense layers,
- GAP layer introduces linearity btw last convolution layer and final dense layer
- as CAM only uses features from the last convolution layer, low-dimension attribution maps are generated.

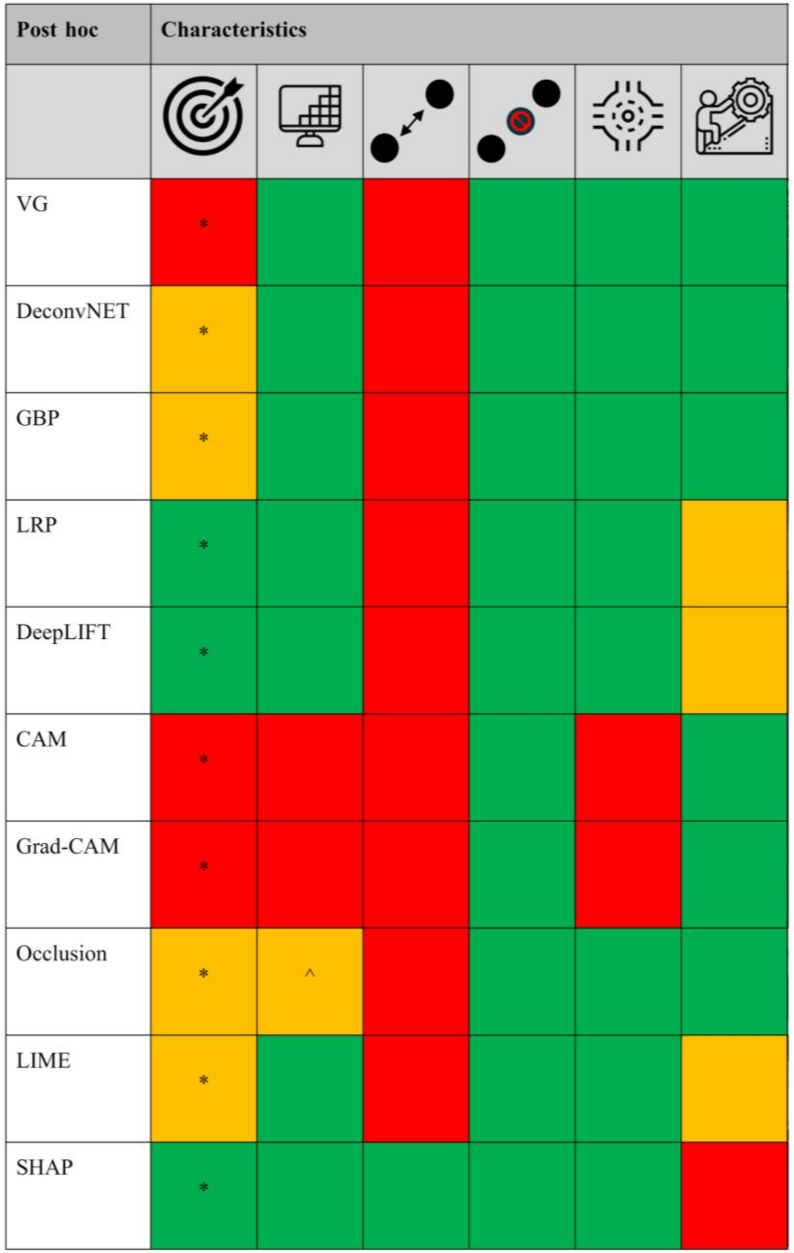
CAM pros & cons

- can visualize whether a model is able to roughly focus on specific targets,
- low specificity, thus it lacks discriminative power to accurately characterize class based features
- only applicable to CNNs with a GAP layer immediately prior to prediction which
  - may exhibit inferior performance (e.g. classification) vs. general networks
  - or be inapplicable top other common imaging tasks (e.g. captioning)

Example: Perturbation analysis of multiple attribution methods showed that

- gradient based methods have higher specificity than CAM
- yet, CAM can be discriminative
  - in classification tasks in which the classes have clear visual differences, e.g., healthy brain vs. Alzheimer's brain or
  - by performing patch based (more focused) tumor analysis instead of whole image tumor analysis

Overview of the *post hoc* attribution methods

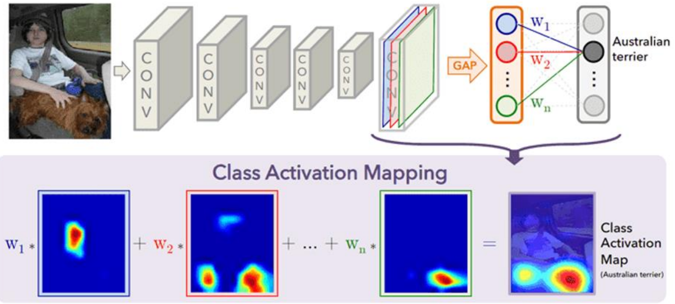


Post hoc	Characteristics					
	🎯	🖥	↔	🚫	⌖	⚙
VG	🟥 *	🟩	🟥	🟩	🟩	🟩
DeconvNET	🟨 *	🟩	🟥	🟩	🟩	🟩
GBP	🟨 *	🟩	🟥	🟩	🟩	🟩
LRP	🟩 *	🟩	🟥	🟩	🟩	🟨
DeepLIFT	🟩 *	🟩	🟥	🟩	🟩	🟨
CAM	🟥 *	🟥	🟥	🟩	🟥	🟩
Grad-CAM	🟥 *	🟥	🟥	🟩	🟥	🟩
Occlusion	🟨 *	🟨 ^	🟥	🟩	🟩	🟩
LIME	🟨 *	🟩	🟥	🟩	🟩	🟨
SHAP	🟩 *	🟩	🟩	🟩	🟩	🟥

de Vries *et al* 2023

Post-hoc XAI:
Class Activation Mapping

Suara *et al* 2023



Gradient-weighted CAM (Grad-CAM)

XAI method adopting a generalization of the CAM approach by employing the gradient of the output of the network with respect to the activations of the feature maps to generate the attribution map (heatmap)

- class-specific producing a separate visualization for every class
- employs gradient-based weighted average of feature maps to create heatmaps

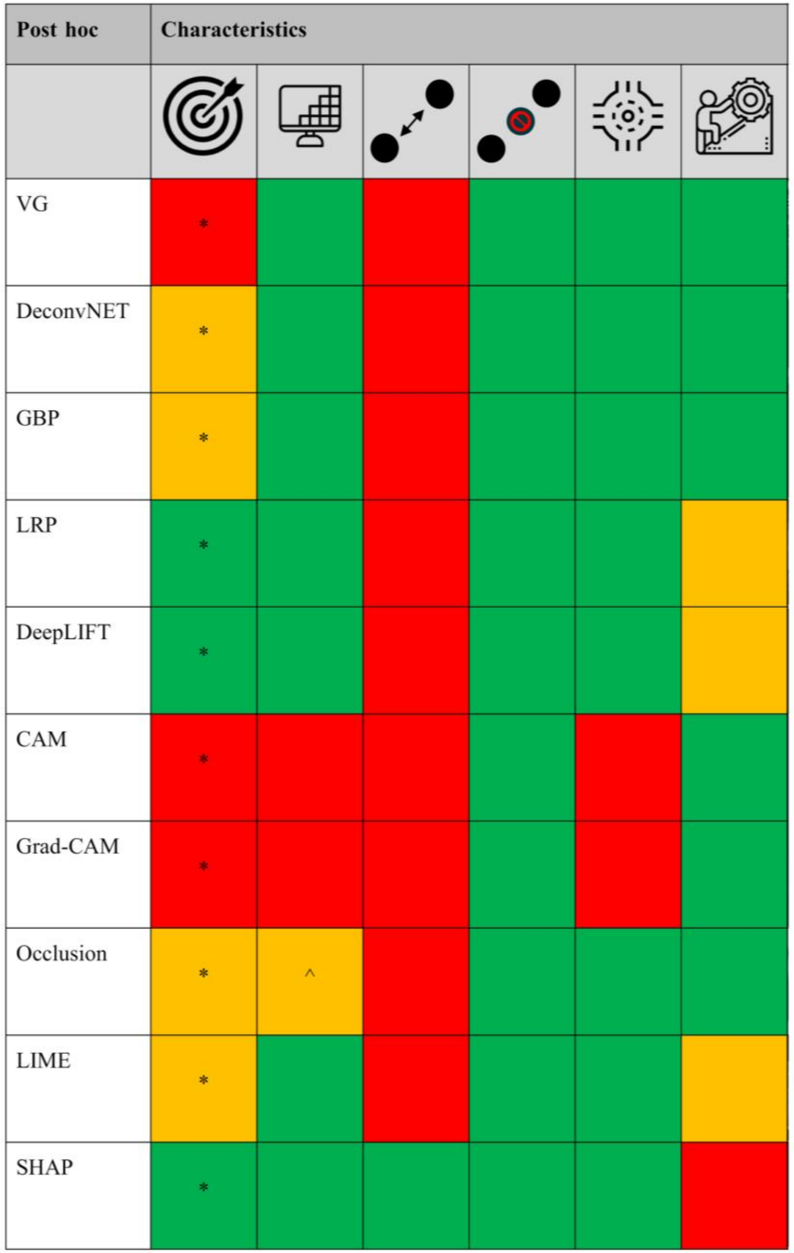
Grad-CAM pros & cons

- improves accuracy and interpretability of results over VG/DeconvNET/GBP in studies aiming at cancer lesion detectability
- unlike CAM, no requirement for specific CNN architecture thus applicable for captioning and visual question answering:
  - for fully-connected CNNs GradCAM reduces to CAM
- limitations include lack of robustness to changes in input image and unclear explanation of the basis for prediction in complex images
- low specificity & resolution due to the low dimensionality of its attribution maps

Grad-CAM current/future developments:

- use more sophisticated methods for computing the gradients and weights to improve robustness to input image variations
- incorporating prior knowledge into the method to reduce misdiagnosis risk

Overview of the *post hoc* attribution methods



Post hoc	Characteristics					
VG	🟥 *	🟩	🟥	🟩	🟩	🟩
DeconvNET	🟨 *	🟩	🟥	🟩	🟩	🟩
GBP	🟨 *	🟩	🟥	🟩	🟩	🟩
LRP	🟩 *	🟩	🟥	🟩	🟩	🟨
DeepLIFT	🟩 *	🟩	🟥	🟩	🟩	🟨
CAM	🟥 *	🟥	🟥	🟩	🟥	🟩
Grad-CAM	🟥 *	🟥	🟥	🟩	🟥	🟩
Occlusion	🟨 *	🟨 ^	🟥	🟩	🟩	🟩
LIME	🟨 *	🟩	🟥	🟩	🟩	🟨
SHAP	🟩 *	🟩	🟩	🟩	🟩	🟥

de Vries *et al* 2023

Post-hoc XAI:
Class Activation Mapping

Overview of the *post hoc* attribution methods

Gradient-weighted CAM (Grad-CAM)

Example 1: Whole-image or segmented lung CT-based COVID19 detection

- most used XAI method producing both very specific and non-specific attributions
- able to roughly locate the potential COVID-19 lesions with accurate predictions
- both DL model's and Grad-CAM's specificity significantly improved when lungs were a-priori segmented as opposed to using whole CT image input

> DL and XAI methods can benefit from medical based data minimization
> - i.e. by reducing the trainable features or non-informative features from image

Example 2: Automated grading of enlarged perivascular spaces in acute stroke and cerebral hemorrhage detection using the whole image (without data minimization)

- similar non-target specific attribution maps

Example 3: A-priori anatomical segmentation to classify and visualize

- mortality risks on myocardial PET, Alzheimer's disease & schizophrenia on MRI
- although data manipulation suppresses non-specific features, Grad-CAM still suffers from low specificity due to its low-dimensional attribution maps

Example 4.1: class-discriminative for classification tasks with clear radiological difference between the classes

Example 4.2: lacks fine-grained details for tasks with less obvious radiological differences, e.g., predicting survival from tumor characteristics,

- complementary attribution methods should then be used: e.g. VG and GBP

Post hoc	Characteristics					
	🎯	🖥	↔	🚫	⌖	⚙
VG	🟥 *	🟩	🟥	🟩	🟩	🟩
DeconvNET	🟨 *	🟩	🟥	🟩	🟩	🟩
GBP	🟨 *	🟩	🟥	🟩	🟩	🟩
LRP	🟩 *	🟩	🟥	🟩	🟩	🟨
DeepLIFT	🟩 *	🟩	🟥	🟩	🟩	🟨
CAM	🟥 *	🟥	🟥	🟩	🟥	🟩
Grad-CAM	🟥 *	🟥	🟥	🟩	🟥	🟩
Occlusion	🟨 *	🟨 ^	🟥	🟩	🟩	🟩
LIME	🟨 *	🟩	🟥	🟩	🟩	🟨
SHAP	🟩 *	🟩	🟩	🟩	🟩	🟥

de Vries *et al* 2023

Post-hoc XAI:
Class Activation Mapping

Feature importance-weighted Grad-Cam (Grad-CAM++)

Grad-CAM averages the gradients of different feature maps to produce heatmaps
Grad-CAM++ takes the weighted average of those gradients measuring the importance of every unit of a feature map

- Grad-CAM++ pros: better localization of target-specific features than Grad-CAM

Example 1: prediction of knee osteoarthritis using MRI

- Grad-CAM++ attained more target-specific attribution maps than Grad-CAM

Guided Grad-CAM (gGrad-CAM)

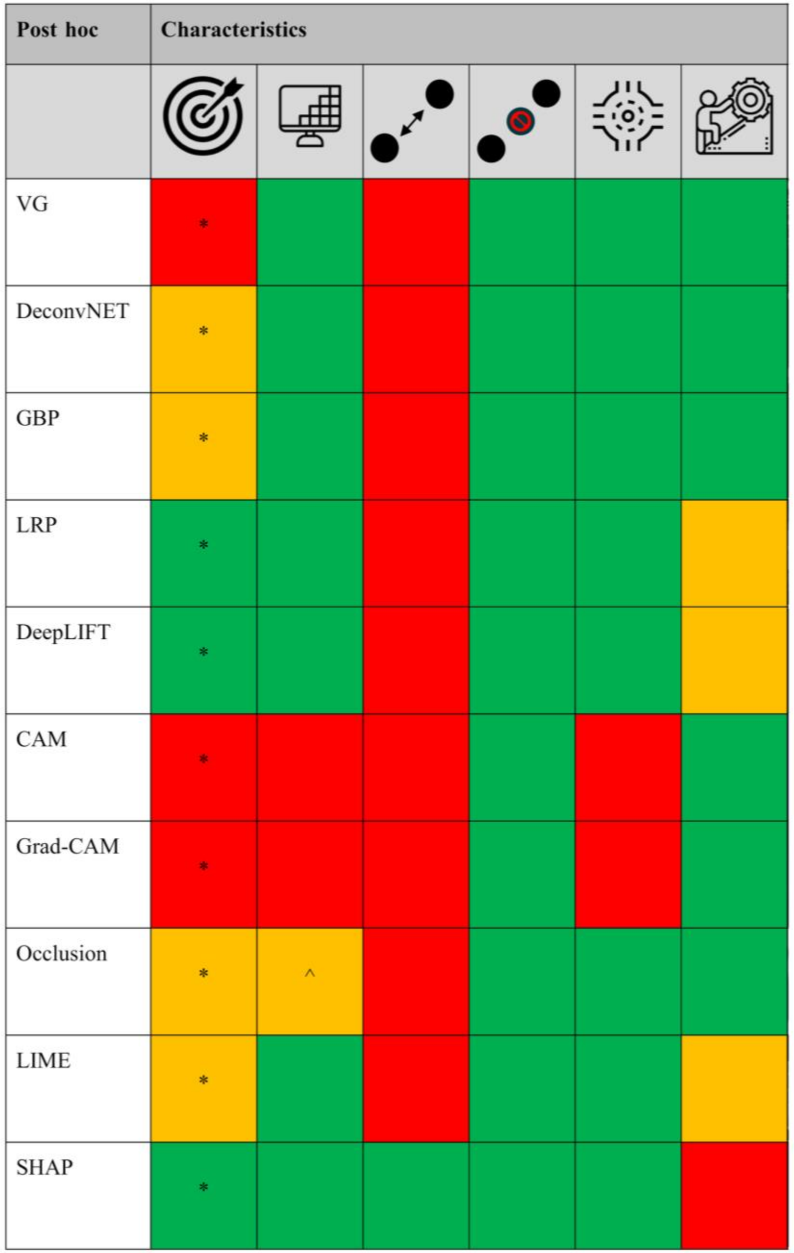
Combination of GBP and grad-CAM XAI methods

- point-wise multiplication of the final GBP (high-resolution) and grad-CAM (less noisy, class-discriminative) attribution maps
- gGrad-CAM pros: combine the advantages of attribution methods for human-interpretable and precise model visualization (best of two worlds)

Example 1: Comparison of different XAI methods for brain glioma classification

- Grad-CAM provided the least noisy attribution maps and
- GBP provided attribution maps with high resolution but not class-discriminative.
- gGrad-CAM provided both class-discriminative as high-resolution maps
  - the edges of the tumor are highlighted instead of the whole tumor

Overview of the *post hoc* attribution methods



Post hoc	Characteristics					
VG	🟥 *	🟩	🟥	🟩	🟩	🟩
DeconvNET	🟨 *	🟩	🟥	🟩	🟩	🟩
GBP	🟨 *	🟩	🟥	🟩	🟩	🟩
LRP	🟩 *	🟩	🟥	🟩	🟩	🟨
DeepLIFT	🟩 *	🟩	🟥	🟩	🟩	🟨
CAM	🟥 *	🟥	🟥	🟩	🟥	🟩
Grad-CAM	🟥 *	🟥	🟥	🟩	🟥	🟩
Occlusion	🟨 *	🟨 ^	🟥	🟩	🟩	🟩
LIME	🟨 *	🟩	🟥	🟩	🟩	🟨
SHAP	🟩 *	🟩	🟩	🟩	🟩	🟥

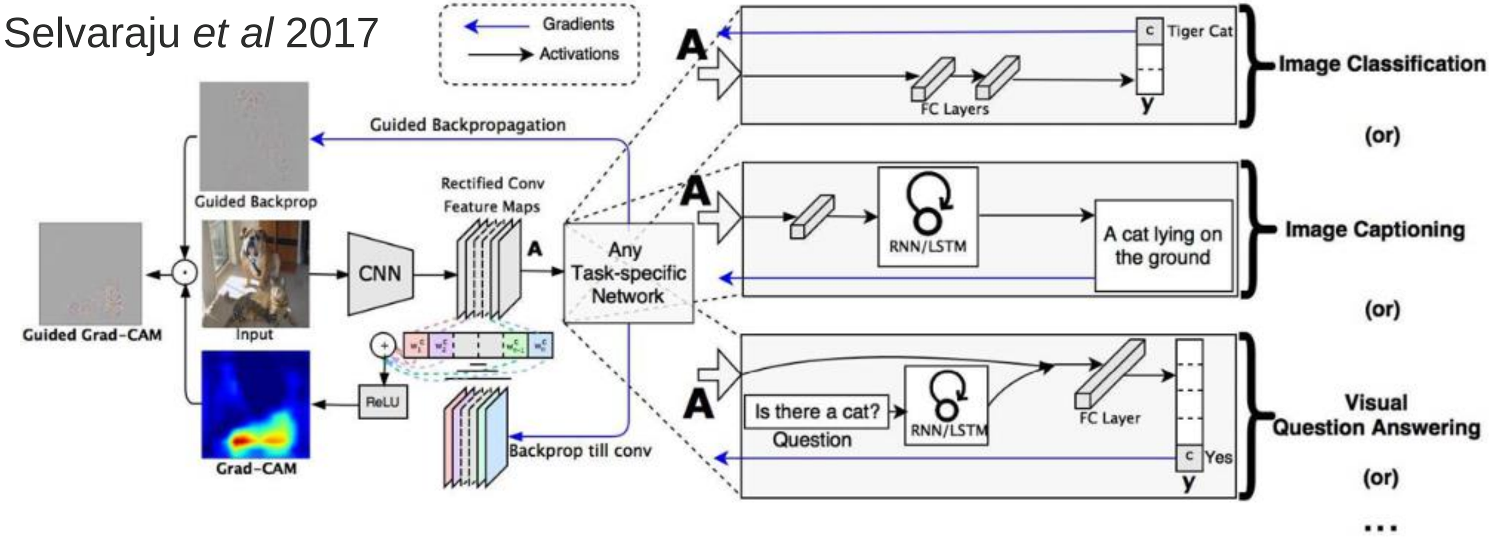
de Vries *et al* 2023

Post-hoc XAI:
Class Activation Mapping

Guided Grad-CAM (gGrad-CAM)

Combination of GBP and grad-CAM XAI methods (best of 2 worlds)

Selvaraju *et al* 2017



Non-target specific features in attribution maps arise because of

- underperformance in DL algorithms and/or attribution methods
- lack of anatomical specificity in input data (e.g. unsegmented whole images)
- input image artifacts tricking DL algorithms and attribution methods

Therefore, to ensure high specificity in clinical radiology workflow:

- use high-quality input image data,
- apply (medical based) data minimization
- employ a priori (DL-based) quality control to detect bias in the input images

Overview of the *post hoc* attribution methods

Post hoc	Characteristics					
	🎯	🖥	⤢	⦸	⌖	⚙
VG	🟥 *	🟩	🟥	🟩	🟩	🟩
DeconvNET	🟨 *	🟩	🟥	🟩	🟩	🟩
GBP	🟨 *	🟩	🟥	🟩	🟩	🟩
LRP	🟩 *	🟩	🟥	🟩	🟩	🟨
DeepLIFT	🟩 *	🟩	🟥	🟩	🟩	🟨
CAM	🟥 *	🟥	🟥	🟩	🟥	🟩
Grad-CAM	🟥 *	🟥	🟥	🟩	🟥	🟩
Occlusion	🟨 *	🟨 ^	🟥	🟩	🟩	🟩
LIME	🟨 *	🟩	🟥	🟩	🟩	🟨
SHAP	🟩 *	🟩	🟩	🟩	🟩	🟥

de Vries *et al* 2023

Post-hoc XAI: Perturbation-based methods

Occlusion mapping

as the name implies, this approach attempts to reveal the feature importance of a model using systematic or random perturbation/conditioning over the image

- by occluding different voxels/patches/regions (e.g., setting input pixels to zero)
- in contrast to previous methods, occlusion maps do not take the feature maps into account, but only the different patches (grid- or atlas-wise combination of multiple pixels) of the input image

variant 1: Randomized Input Sampling for Explanation (RISE)

- generates multiple random perturbation maps, pointwise-multiplied with input

variant 2: Perturbations using square-grid divisions of the input image

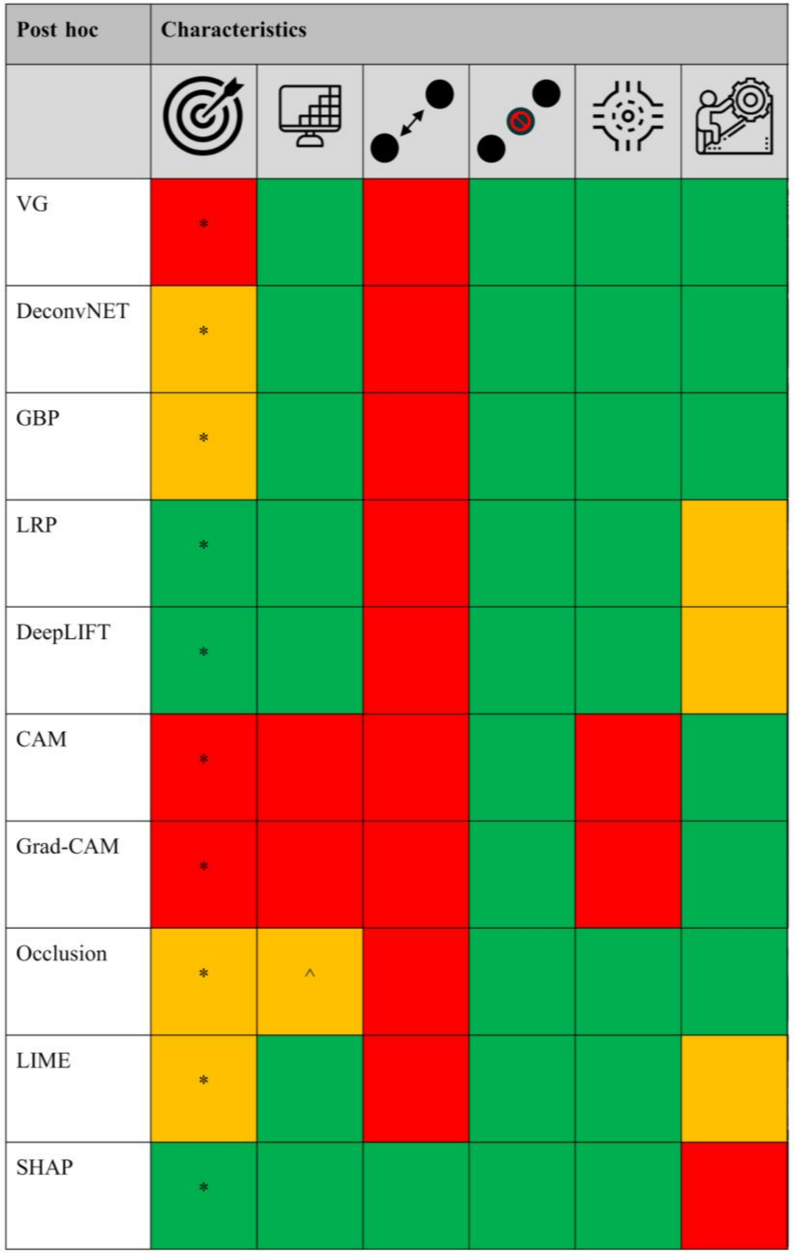
Occlusion maps pros & cons

- simple and intuitive method to perform and interpret
- can easily be adapted to specific occlusion analysis (e.g. clinical atlas-based)
- can easily be compared with more traditional clinical/anatomical atlas-based analysis thus providing a medical based, transparent and intuitive visualization of the DL algorithm
- too rigid to follow anatomical/pathological structures present in the images
- large computational requirements due to many forward & backward propagations

Example 1: predicting and visualizing Alzheimer’s Disease

- instead of rectangles, the Harvard-Oxford cortical and subcortical structural atlas was used for occlusion mapping

Overview of the *post hoc* attribution methods



Post hoc	Characteristics					
	🎯	🖥	●↔●	●⊘●	⊙	⚙
VG	🟥 *	🟩	🟥	🟩	🟩	🟩
DeconvNET	🟨 *	🟩	🟥	🟩	🟩	🟩
GBP	🟨 *	🟩	🟥	🟩	🟩	🟩
LRP	🟩 *	🟩	🟥	🟩	🟩	🟨
DeepLIFT	🟩 *	🟩	🟥	🟩	🟩	🟨
CAM	🟥 *	🟥	🟥	🟩	🟥	🟩
Grad-CAM	🟥 *	🟥	🟥	🟩	🟥	🟩
Occlusion	🟨 *	🟨 ^	🟥	🟩	🟩	🟩
LIME	🟨 *	🟩	🟥	🟩	🟩	🟨
SHAP	🟩 *	🟩	🟩	🟩	🟩	🟥

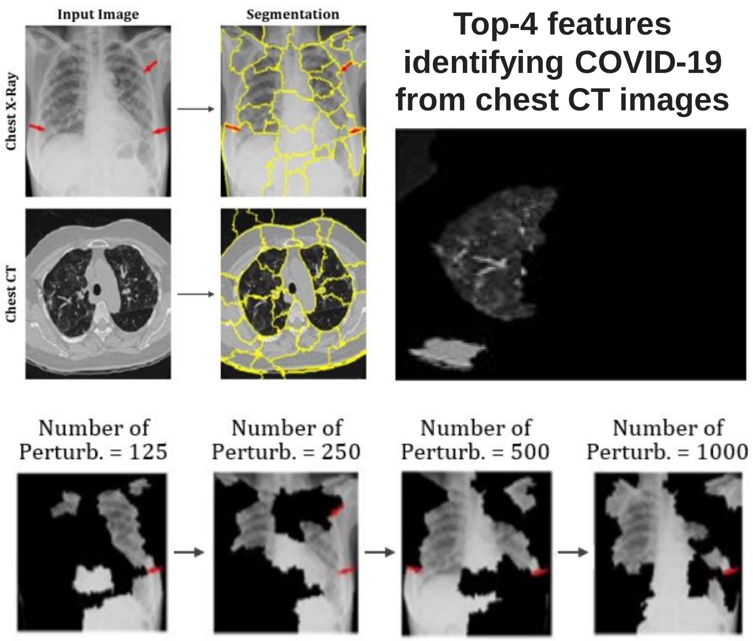
de Vries *et al* 2023

Post-hoc XAI:
Perturbation-based methods

Local interpretable model-agnostic explanations (LIME)

XAI method perturbating super-pixels (group of pixels that share common pixel/voxel characteristics) instead of predefined or random occlusion functions

- super-pixels: follow anatomical/pathological structures/characteristic of the image



Top-4 features identifying COVID-19 from chest CT images

Ahsan *et al* 2021

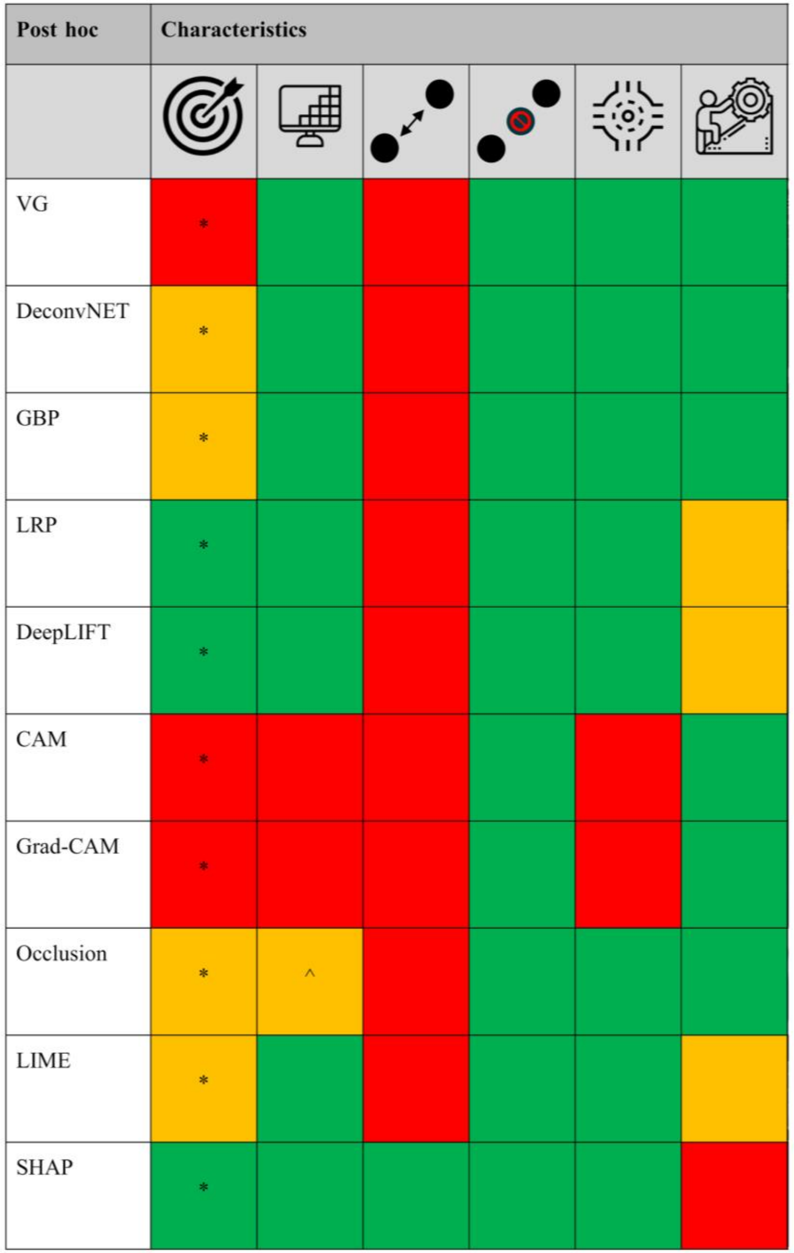
LIME pros & cons

- provide a better representation of the image than predefined occlusion mapping
- as LIME uses super-pixels as a whole, its occlusion maps have relatively low specificity thus suffering from non-target specific activation.
- requires initialization parameters (kernel size, maximum distance, etc.) to compute super-pixels, which can be difficult to optimize

Example: CT- image-based COVID19 detection

- indeed, the selected super-pixels followed anatomical/pathological structures/ characteristics of the chest CT image
- non-target specific features (e.g., chest wall) showed high activation suggesting low specificity and non-target specific activation

Overview of the *post hoc* attribution methods



Post hoc	Characteristics					
VG	🟥 *	🟩	🟥	🟩	🟩	🟩
DeconvNET	🟨 *	🟩	🟥	🟩	🟩	🟩
GBP	🟨 *	🟩	🟥	🟩	🟩	🟩
LRP	🟩 *	🟩	🟥	🟩	🟩	🟨
DeepLIFT	🟩 *	🟩	🟥	🟩	🟩	🟨
CAM	🟥 *	🟥	🟥	🟩	🟥	🟩
Grad-CAM	🟥 *	🟥	🟥	🟩	🟥	🟩
Occlusion	🟨 *	🟨 ^	🟥	🟩	🟩	🟩
LIME	🟨 *	🟩	🟥	🟩	🟩	🟨
SHAP	🟩 *	🟩	🟩	🟩	🟩	🟥

de Vries *et al* 2023

Post-hoc XAI: Perturbation-based methods

SHapley additive exPlanations (SHAP)

Advanced XAI algorithm calculating *shapley values:*

- mathematical formulas based on game theory to measure the marginal contributions of all factors that could affect a prize in a competition
- For SHAP, they represent each voxel's attribution to the change of the expected model prediction when conditioning on that voxel using reference samples

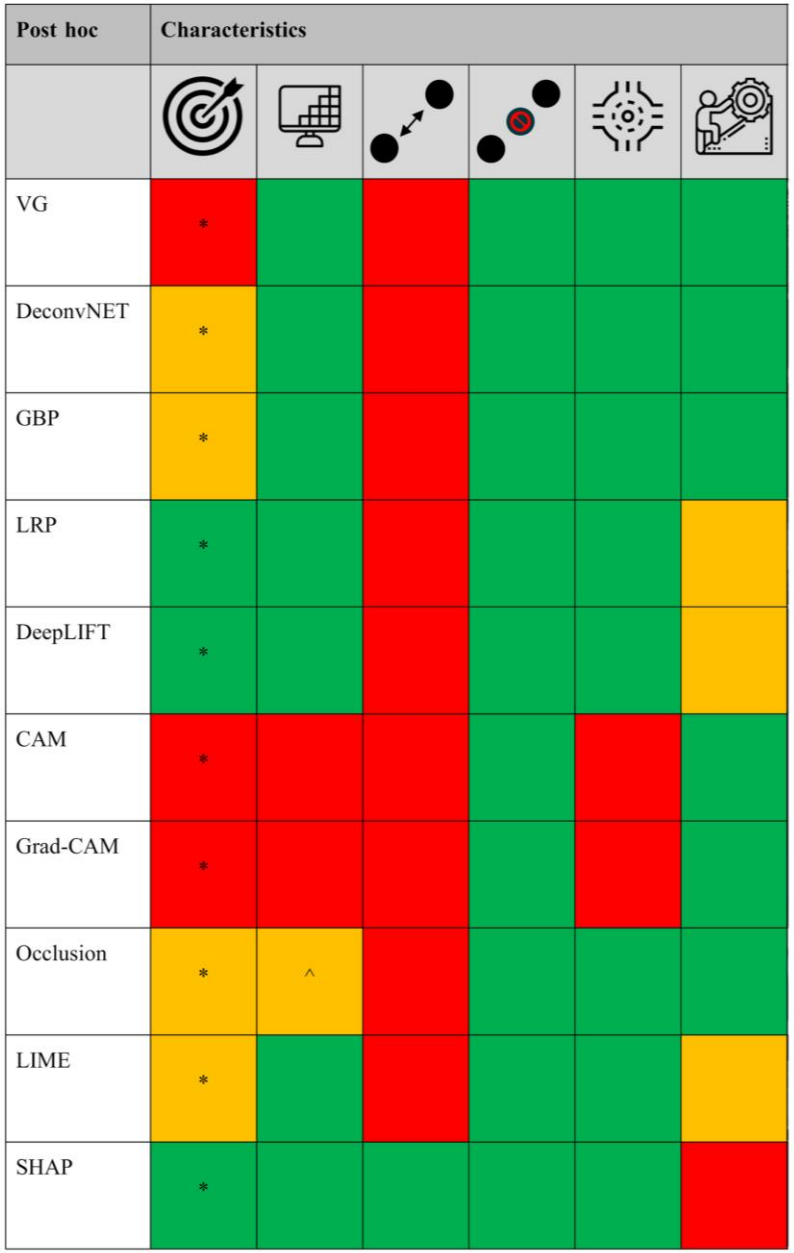
DeepSHAP: an extension of SHAP (works in an almost similar way as DeepLIFT)

- can provide both local and global explanation based on individual pixels/voxels,
- can also inform whether a pixel/voxel is negatively associated or positively associated with the predictive class.

SHAP pros & cons

- powerful in providing both local and global interpretability
- DeepSHAP may be difficult to interpret due to presence of both positive and negative associations with the model's decision
- due to the reference samples, DeepSHAP may not work optimal in classification of substantial (non-)rigid anatomical/pathological variations in the images.
  - feature explanation may thus be negatively impacted by anatomical differences between reference and input image resulting in low specificity
- computationally expensive for large datasets or complex models
- additional data and analysis needed to explain the causal nature of SHAP's feature relationships results

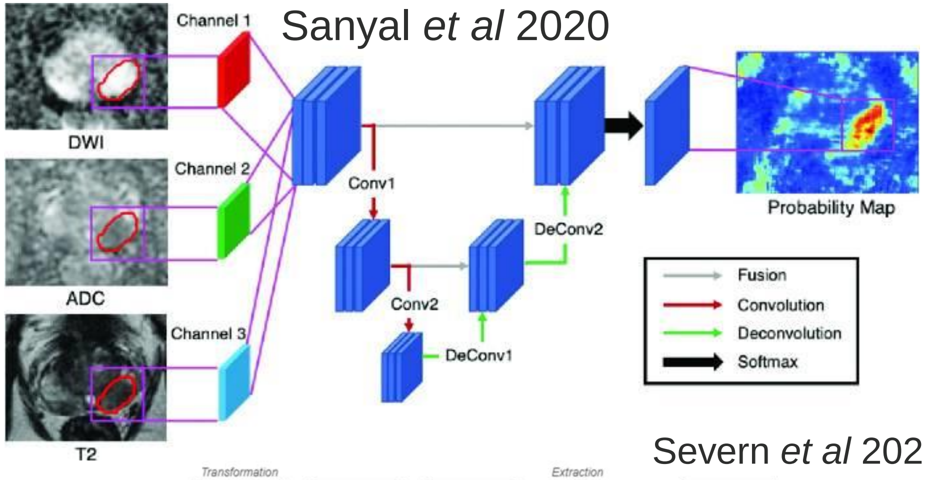
Overview of the *post hoc* attribution methods



Post hoc	Characteristics					
VG	🟥 *	🟩	🟥	🟩	🟩	🟩
DeconvNET	🟨 *	🟩	🟥	🟩	🟩	🟩
GBP	🟨 *	🟩	🟥	🟩	🟩	🟩
LRP	🟩 *	🟩	🟥	🟩	🟩	🟨
DeepLIFT	🟩 *	🟩	🟥	🟩	🟩	🟨
CAM	🟥 *	🟥	🟥	🟩	🟥	🟩
Grad-CAM	🟥 *	🟥	🟥	🟩	🟥	🟩
Occlusion	🟨 *	🟨 ^	🟥	🟩	🟩	🟩
LIME	🟨 *	🟩	🟥	🟩	🟩	🟨
SHAP	🟩 *	🟩	🟩	🟩	🟩	🟥

de Vries *et al* 2023

Post-hoc XAI: Probability maps, Deep Feature maps, radiomics and physics/clinical data



Sanyal *et al* 2020

<u>Probability Maps as Attribution Maps (voxel-level annotation)</u>

previous post hoc XAI methods predominantly focus on weak training labels

- i.e., one label for the whole image (e.g. classes in classification tasks).

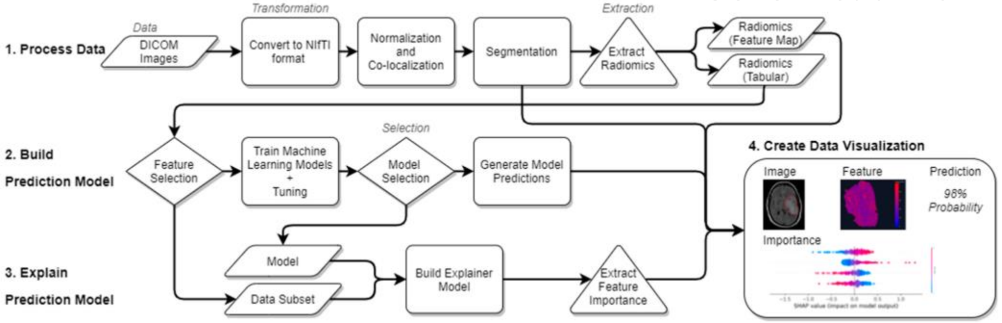
In contrast, segmentation DL algorithms use voxel-level annotations and compute <u>voxel-level probability maps</u> less complex to understand and interpret

<u>example</u>: used for detecting prostate lesion from multi-parametric MR

<u>Radiomics as explainable features thank to their strict math definitions</u>

segmentations can be used to explore radiomic-based differences btw classes
Radiomics involve the automated extraction of clinically relevant information
<u>example</u>: a joint detection & radiomics classification showed clear difference btw COVID-19 & community acquired pneumonia using explainable radiomics

Severn *et al* 2022



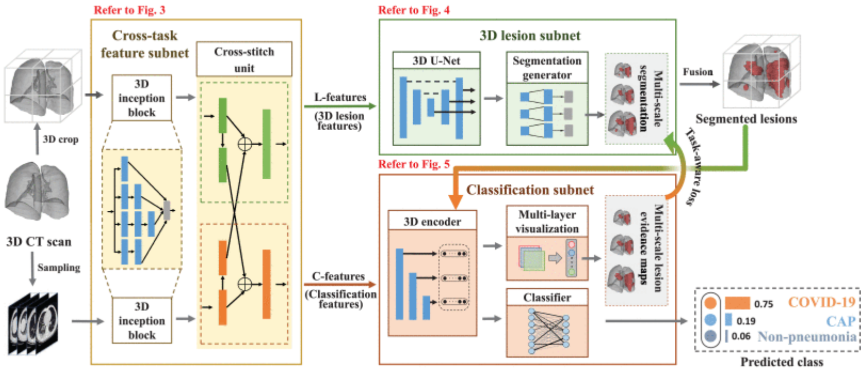
Probability maps and radiomics <span style="color:red">limitations</span> impacting DL potential & performance:

- voxel-level annotation is very cumbersome
- radiomics depends on accurate VOI annotations and *a priori* assumptions

<u>Deep Feature Maps:</u> attributions maps of the intermediate layers of DL models

- directly visualize the underlying feature extraction mechanism
- provide understanding of which features and how they are used

Wang *et al* 2021



<u>Physics-based AI models & clinical data (e.g., patient history)</u>

- both may aid in better performance and transparency of AI algorithms

Ad-hoc XAI: Inherently explainable AI models in biomedical imaging

Post-hoc XAI limitations:

- reverse-engineer an already trained predictive model.
- as the explanation is detached from the predictive model, the explanation is not guaranteed to truthfully mimic the internal calculations of the black box
- post-hoc heatmaps do not explain the full reasoning process
  - only give intuition, making them irrelevant to tasks in realistic scenario’s
  - it has been shown that feature attribution maps do not fulfil clinical requirements to correctly explain a model’s decision process

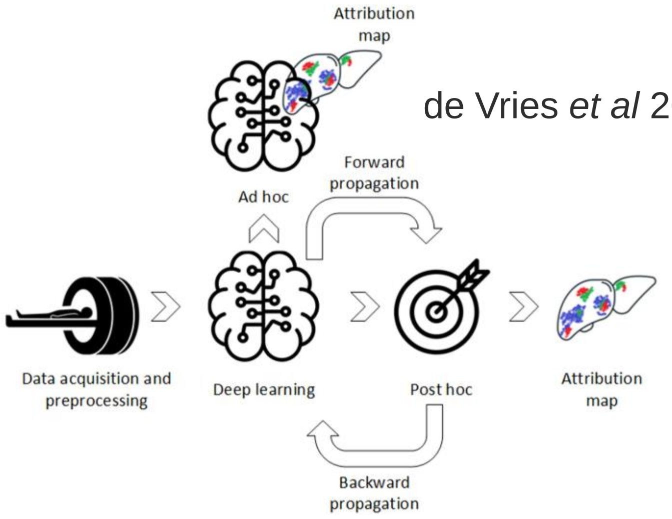
Promising alternative: Ad-hoc XAI methods

Ad hoc XAI models are intrinsically able to learn and explain,

- different to the DL models that predominantly focusses on learning to achieve high performance (learning) and require a post hoc XAI algorithm to explain model behavior

- intrinsically interpretable models, examples below and in the Table:
  - based on reference prototypical parts (xDNN, PIP-Net, etc.)
  - capsule networks: linked subcollections of neurons in subsequent CNN layers, forming a CNN network
  - spatial attention mapping employs attention estimators trained during model training (ad-hoc) to compute attention masks from a convolution layer as a goal to extract important local feature vectors

Conceptual difference between *post hoc* and *ad hoc* XAI methods



de Vries *et al* 2023

Ad *hoc* attribution methods

de Vries *et al* 2023

Ad hoc	Characteristics					
	🎯	🖥	↔	⊘	⌖	⚙
xDNN	🟨 *	🟩	🟥	🟩	🟩	🟩
Attention estimator	🟩 *	🟩	🟥	🟩	🟩	🟩
Capsule network	🟩 *	🟩	🟩	🟩	🟩	🟨

Ad-hoc XAI: Inherently explainable AI models in biomedical imaging

Explainable deep neural network (xDNN):
XAI method employing a prototype identification layer in the network

- to identify new data samples based on similarity to predefined data samples (prototypes)
- representative prototypes need to be selected for each class.
  - they are used to assess new images based on their similarity.

xDNN provides the user with transparent and intuitive model explanations, we humans extract features based on previous experience

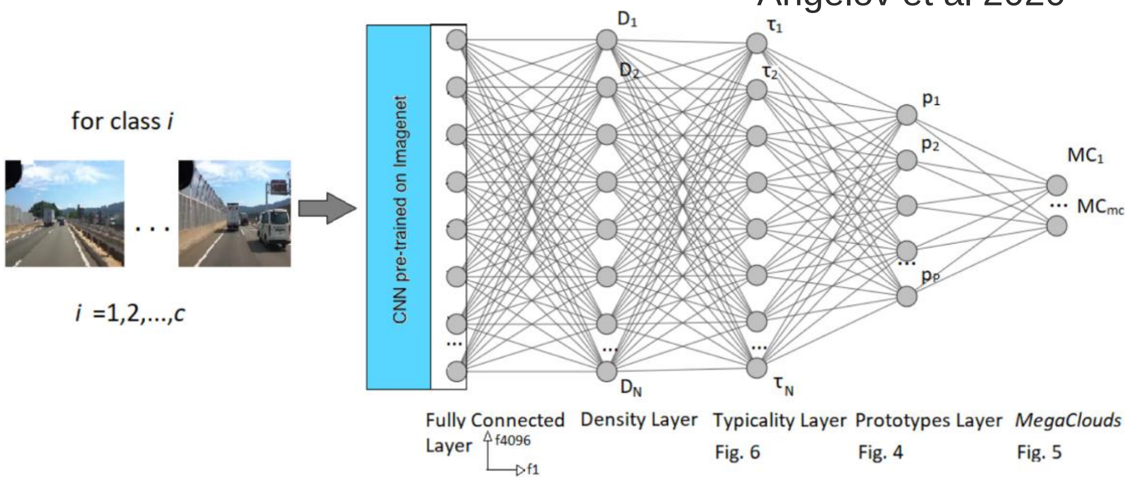
xDNN pros and cons

- xDNN can be very powerful in tasks where there is known difference between classes
- selecting representative prototypes per class can be a difficult task,
  - especially in case of a cohort with a wide variety in disease morphology.
- difference in class morphology is not always trivial
  - therefore, obtaining representative prototypes can be difficult.

Examples: COVID-19 screening and artifact detection

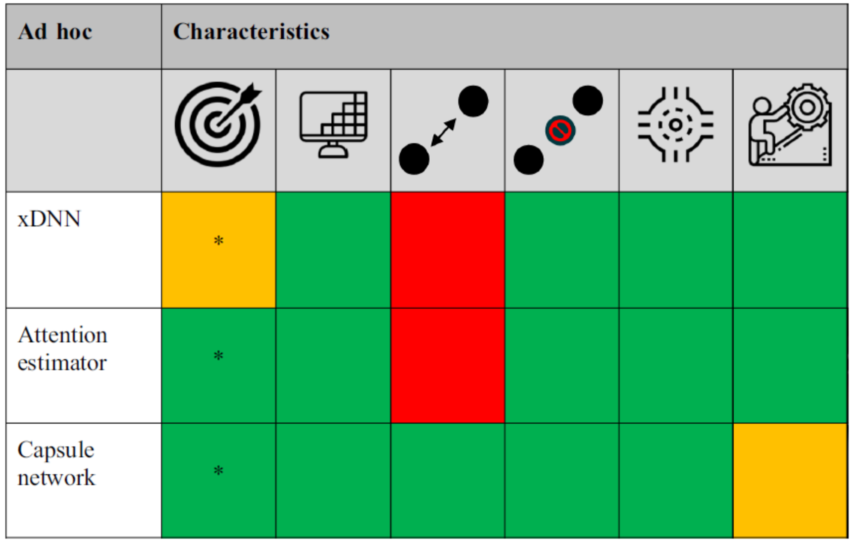
- in these studies, representative prototypes were used to assess new images based on their similarity

Angelov et al 2020



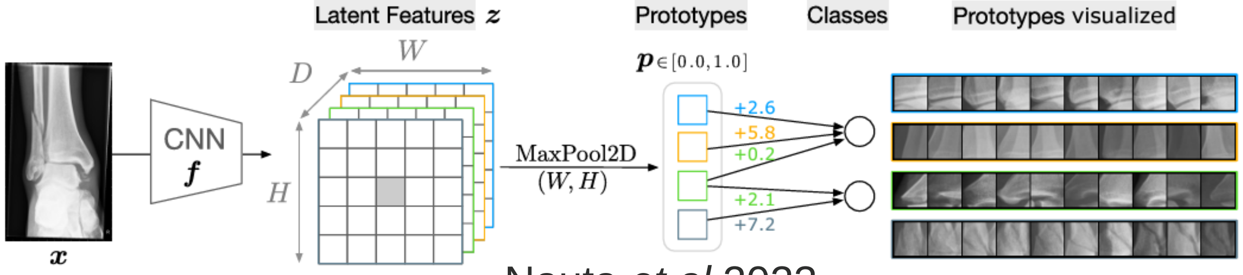
Ad hoc attribution methods

de Vries *et al* 2023



Ad hoc	Characteristics					
xDNN	🟨 *	🟩	🟥	🟩	🟩	🟩
Attention estimator	🟩 *	🟩	🟥	🟩	🟩	🟩
Capsule network	🟩 *	🟩	🟩	🟩	🟩	🟨

Ad-hoc XAI methods: Part-Prototype explainable-by-design AI image classifier



Nauta *et al* 2023

Patch-based Intuitive Prototypes Network (PIP-Net)

a next generation inherently interpretable (ad-hoc XAI) ML model

- explainable-by-design image classifier,
- reasoning adopts the "*recognition-by-components*" concept
  - analyzing whether patches in an input image are similar to learned human-understandable prototypical image parts
- permits understanding of its reasoning with prototypical parts
- abstain from a decision for out-of-distribution inputs

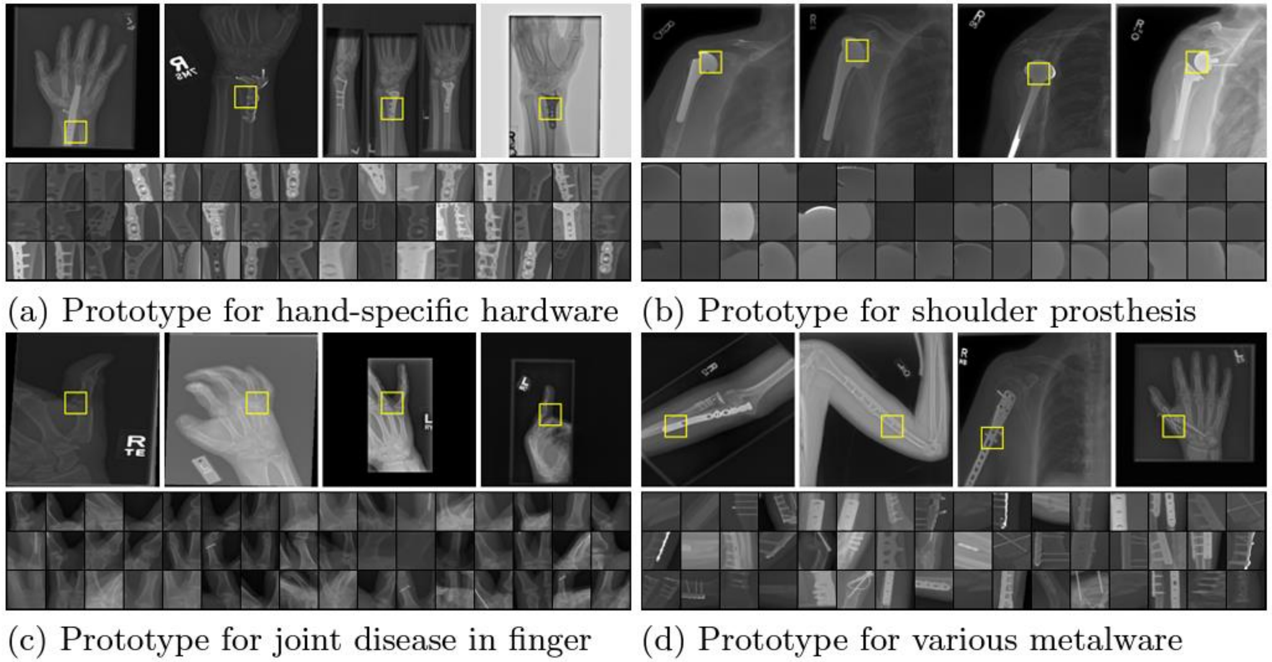
PIP-Net's potential

- its decision-making process is in line with medical classification standards
- as a part-prototype model, does not require any part annotations and only rely on image level class labels.
- because of its unsupervised pretraining of prototypes, data quality problems such as undesired text in an X-ray or labelling errors can be easily identified
- humans can manually correct the reasoning of PIP-Net by directly disabling undesired prototypes

Current challenge:

- Improve alignment of PIP-Net with domain (medical/clinical) knowledge

Interpreting and Correcting Medical Image Classification with PIP-Net



(a) Prototype for hand-specific hardware

(b) Prototype for shoulder prosthesis

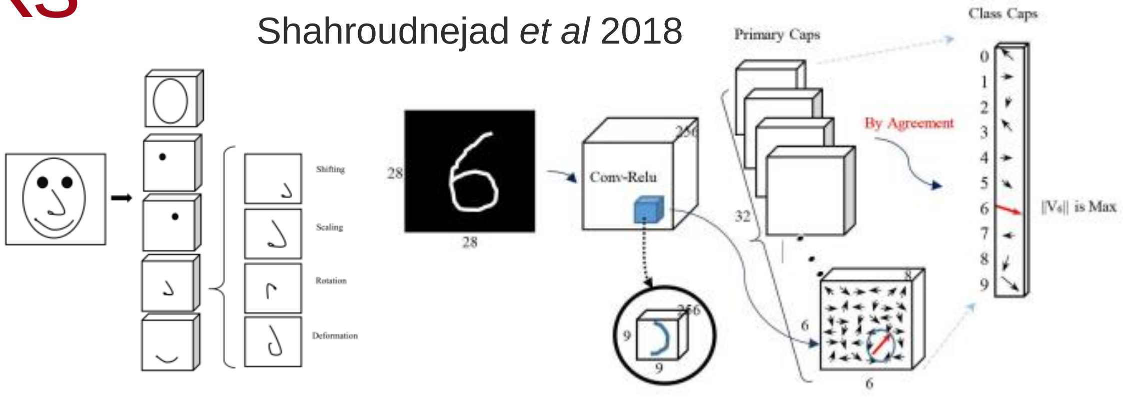
(c) Prototype for joint disease in finger

(d) Prototype for various metalware

Figure: Prototypes relevant to class "abnormal", visualized with a set of image patches and four representative images indicating where the prototype is detected

Ad-hoc XAI: Capsule Networks

Shahroudnejad *et al* 2018



Capsule networks (CapsNets):

XAI method replacing the scalar feature maps from convolution neural networks by vectorized representations (i.e., capsules),

a network of capsules (building block of network) at multiple layers

- capsule: neurons subcollections at a layer describing the presence and instantiation parameters (orientation, thickness, skewed, position, etc.) of a particular object (e.g., tumor or lung) at a given location as a vector

CapsNets encapsulate (possibly large) number of pose information with other instantiation parameters for different object parts

- capsule networks are deep in width instead of deep in height

Capsules from a lower layer try to predict the output for the higher layer based on the instantiation parameters.

- lower layer vectors with high agreement are routed to the following layer
- the other vectors are suppressed, ideally resulting in target specific attribution maps

MIXCAPS: an extension of traditional capsule network,

- instead of a single CNN, a mixture of (expert) CNNs specialized on a subset of the data and an ensemble of capsule networks is used

Capsule networks pros and cons

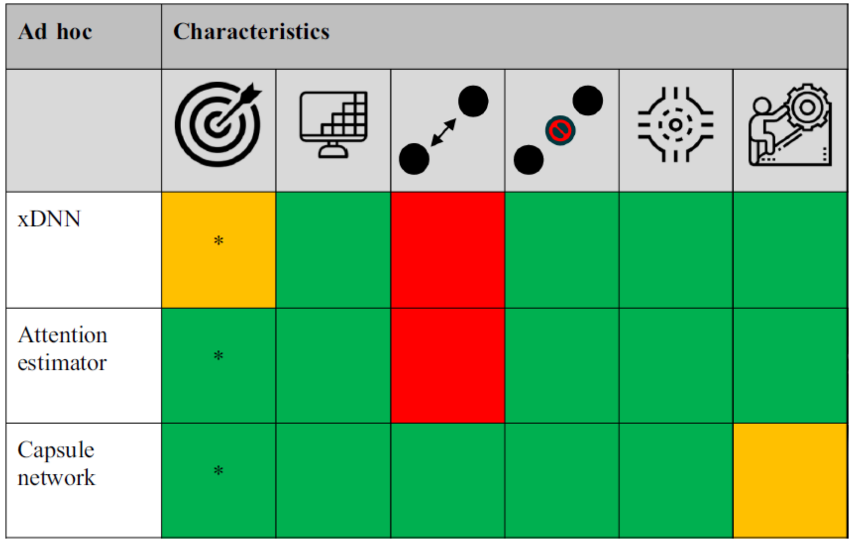
- the new sensation in DL, as they can eliminate the pose and deformation challenges faced by CNNs, require less data and less computational power
- better handling of different visual stimulus & better understanding of proportional changes compared to pooling methods (max/average pooling) of CNNs
- full potential not shown yet, further understanding required before standardized

Examples: detection and visualization of lung cancer nodules in CT images

- MIXCAPS outperformed single CapsNet, CNN and CNN mixture

Ad hoc attribution methods

de Vries *et al* 2023

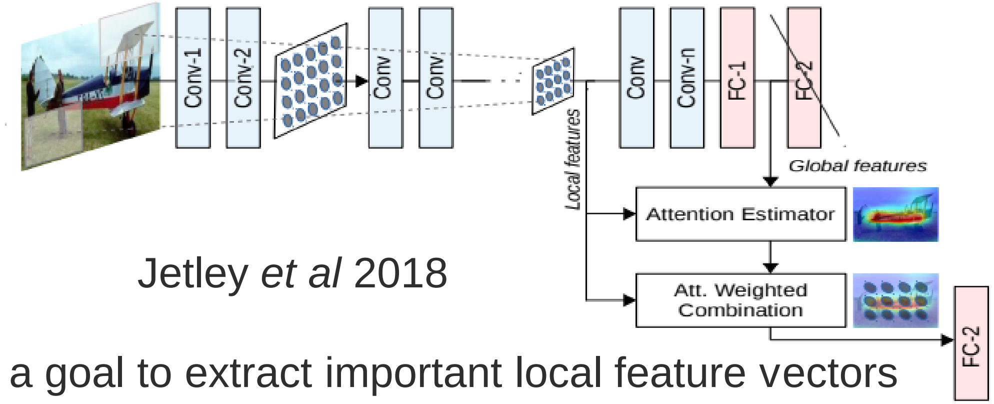


Ad hoc	Characteristics					
	🎯	🖥	↔	🚫	⌖	⚙
xDNN	🟨 *	🟩	🟥	🟩	🟩	🟩
Attention estimator	🟩 *	🟩	🟥	🟩	🟩	🟩
Capsule network	🟩 *	🟩	🟩	🟩	🟩	🟨

Ad-hoc XAI: Attention Mapping

Attention Mapping:
XAI method employing an end-to-end trainable spatial self-attention mechanism

- trained during model training (not after, as in in post hoc attention mechanisms) to support (important) feature extraction
- replaces traditional non-learnable pooling operations (e.g., max-pooling).



Jetley *et al* 2018

It uses attention estimators to compute attention mask from convolution layers as a goal to extract important local feature vectors

- classify input image using only weighted combination of local features, weights represented by attention map.
- network is thus forced to learn a pattern of attention relevant to solving the task at hand

Attention mapping also investigated combined with Multi Instance Learning (MIL). MIL tackles the downsides of weak labels & labor intensive per-voxel annotation by using a set of labeled bags, each consisting of multiple instances (slices).

- bag annotated as positive: if all instances in bag negative (e.g., no disease)
- bag annotated as negative: at least one instance in bag is positive

MIL intrinsically provides a more interpretable decision and in combination with attention mapping it gives insight into every voxel's contribution to the bag label

Example 1: assessment of interpretable MRI biomarkers for Alzheimer's disease

- high correspondence btw attention scores of specific regions & classification score

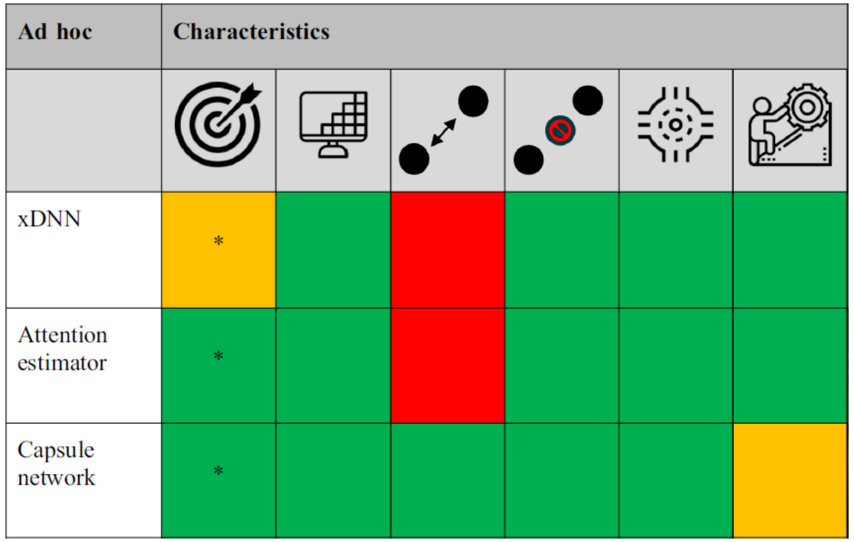
Example 2: COVID19 detection

- better capabilities to extract more complex and scattered regions

Example 3: inverted papilloma, CT nasal polyp classification, adenocarcinoma screening using CT, and segmentation of multiple organs from MRI

- superior target-specific feature extraction

Ad *hoc* attribution methods

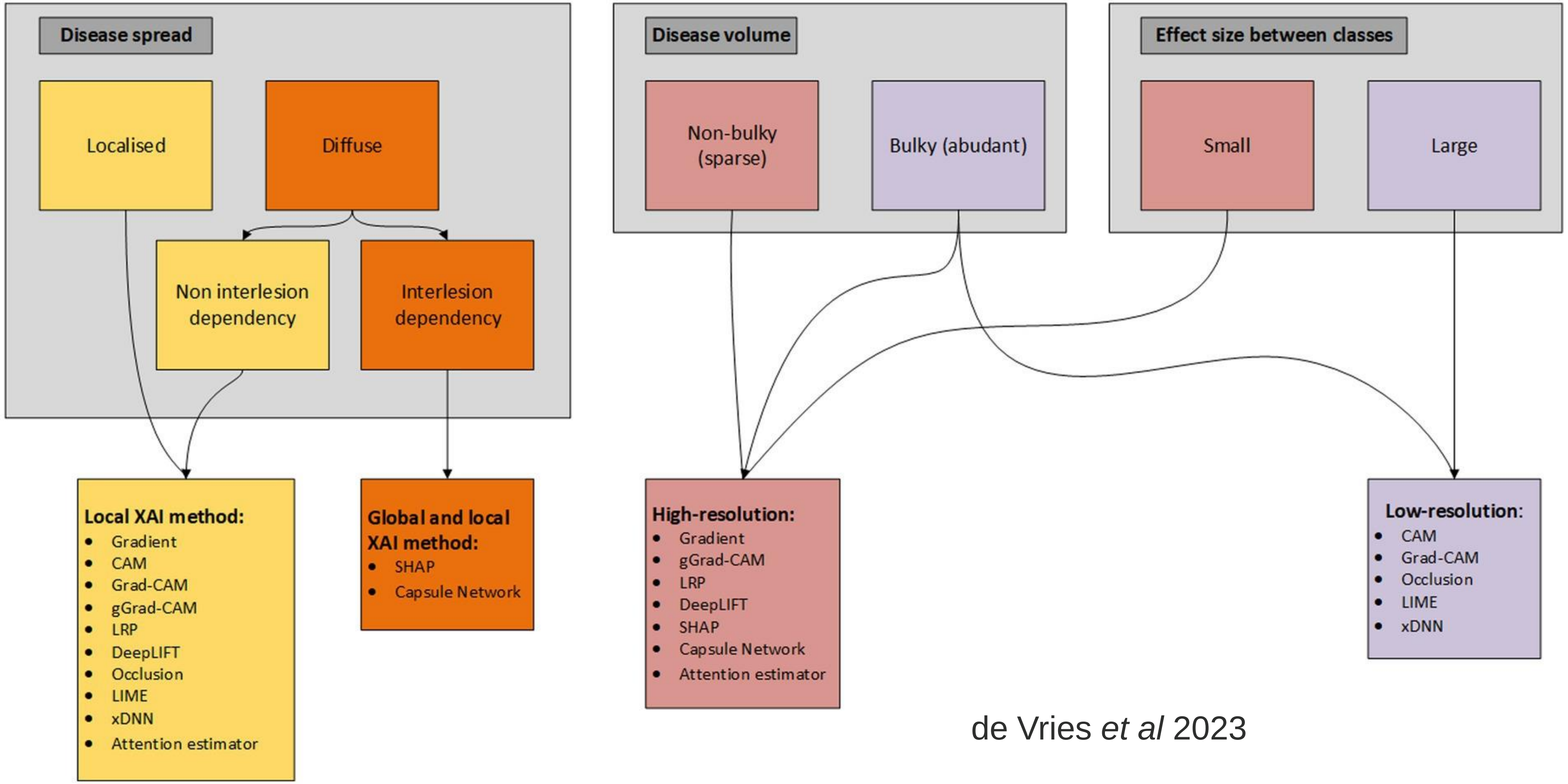


Ad hoc	Characteristics					
	🎯	🖥	● ↔ ●	● ⊘ ●	⌖	⚙
xDNN	🟨 *	🟩	🟥	🟩	🟩	🟩
Attention estimator	🟩 *	🟩	🟥	🟩	🟩	🟩
Capsule network	🟩 *	🟩	🟩	🟩	🟩	🟨

de Vries *et al* 2023

Disease-specific XAI

Flowchart of XAI methods most probably applicable for disease specific characteristics.



de Vries *et al* 2023

Important Note: this flowchart can be helpful for researchers

- to determine *a priori* what XAI methods currently present in literature can aid in explaining their DL model.

However, in the end researchers should determine how the complexity of the AI task compares with the complexity of the XAI method

- therefore, the flowchart should only be seen as an additional tool for XAI application

Evaluating explainable AI: from anecdotal evidence to quantitative evaluation and the Co-12 categorization scheme

XAI rising popularity to understand high-performing black boxes raised the question:

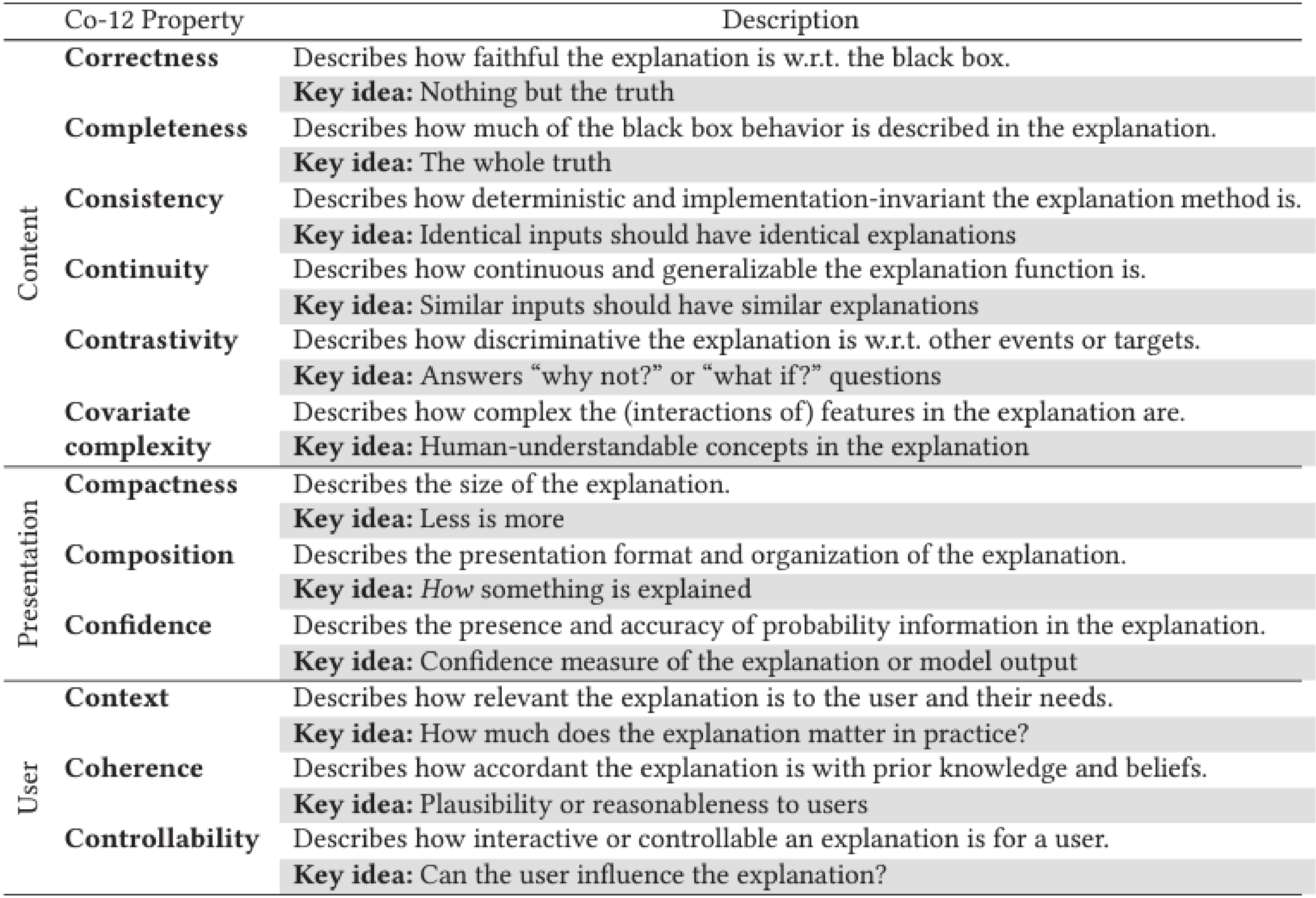
- *how to evaluate explanations of machine learning (ML) models?*

AI Interpretability & explainability are non-binary multi-faceted concepts

Co-12: 12 conceptual properties serving as categorization scheme for systematically reviewing the evaluation practices of XAI

Co-12 scheme provides researchers and practitioners with concrete tools

- thoroughly validate, benchmark, and compare new and existing XAI methods
- open up opportunities to include quantitative metrics as optimization criteria during model training to *optimize for both accuracy and interpretability simultaneously*

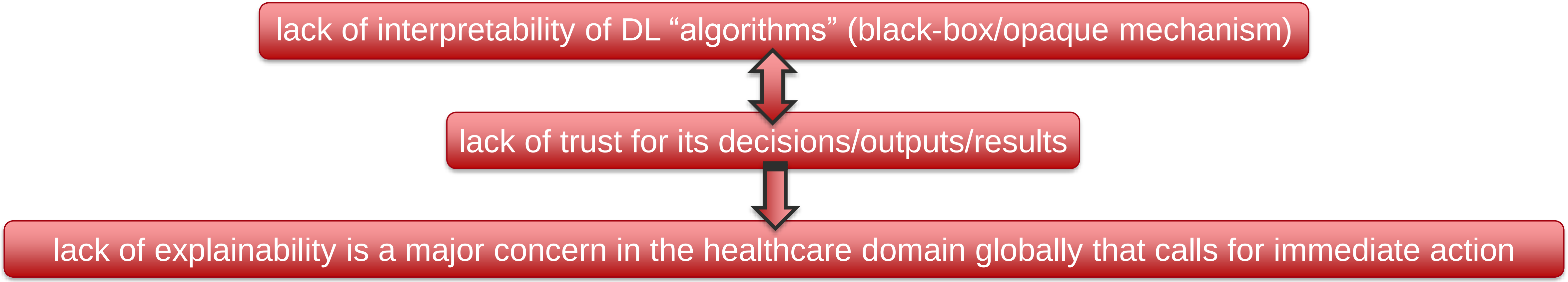


	Co-12 Property	Description
Content	**Correctness**	Describes how faithful the explanation is w.r.t. the black box.
		Key idea: Nothing but the truth
	Completeness	Describes how much of the black box behavior is described in the explanation.
		Key idea: The whole truth
	Consistency	Describes how deterministic and implementation-invariant the explanation method is.
		Key idea: Identical inputs should have identical explanations
	Continuity	Describes how continuous and generalizable the explanation function is.
		Key idea: Similar inputs should have similar explanations
	Contrastivity	Describes how discriminative the explanation is w.r.t. other events or targets.
		Key idea: Answers "why not?" or "what if?" questions
	Covariate complexity	Describes how complex the (interactions of) features in the explanation are.
		Key idea: Human-understandable concepts in the explanation
Presentation	**Compactness**	Describes the size of the explanation.
		Key idea: Less is more
	Composition	Describes the presentation format and organization of the explanation.
		Key idea: *How* something is explained
	Confidence	Describes the presence and accuracy of probability information in the explanation.
		Key idea: Confidence measure of the explanation or model output
User	**Context**	Describes how relevant the explanation is to the user and their needs.
		Key idea: How much does the explanation matter in practice?
	Coherence	Describes how accordant the explanation is with prior knowledge and beliefs.
		Key idea: Plausibility or reasonableness to users
	Controllability	Describes how interactive or controllable an explanation is for a user.
		Key idea: Can the user influence the explanation?

Nauta *et al* 2023

Introduction to Explainable Artificial Intelligence (XAI) in biomedical imaging: taxonomy, clinical applications & future perspectives

One important challenge with deep learning (DL) models is lack of explainability:



lack of interpretability of DL “algorithms” (black-box/opaque mechanism)

lack of trust for its decisions/outputs/results

lack of explainability is a major concern in the healthcare domain globally that calls for immediate action

Explainable Artificial Intelligence (XAI): an emerging area of research seeking to address this issue by

- providing interpretable models
- without sacrificing accuracy.

XAI DL techniques aim to explain how a DL-based AI (DL-AI) model arrived at its decision by

- providing saliency/attribution maps or
- highlighting the most relevant features in the input image that contributed to the model’s output.
- *Important note*: **explanations should align with medical knowledge or be supported by clinical evidence**

By improving the interpretability of DL-AI models,

- clinicians can better understand and trust the decision-making process of these models, to
- improve patient safety and lead to more efficient and effective diagnosis and treatment of various diseases.

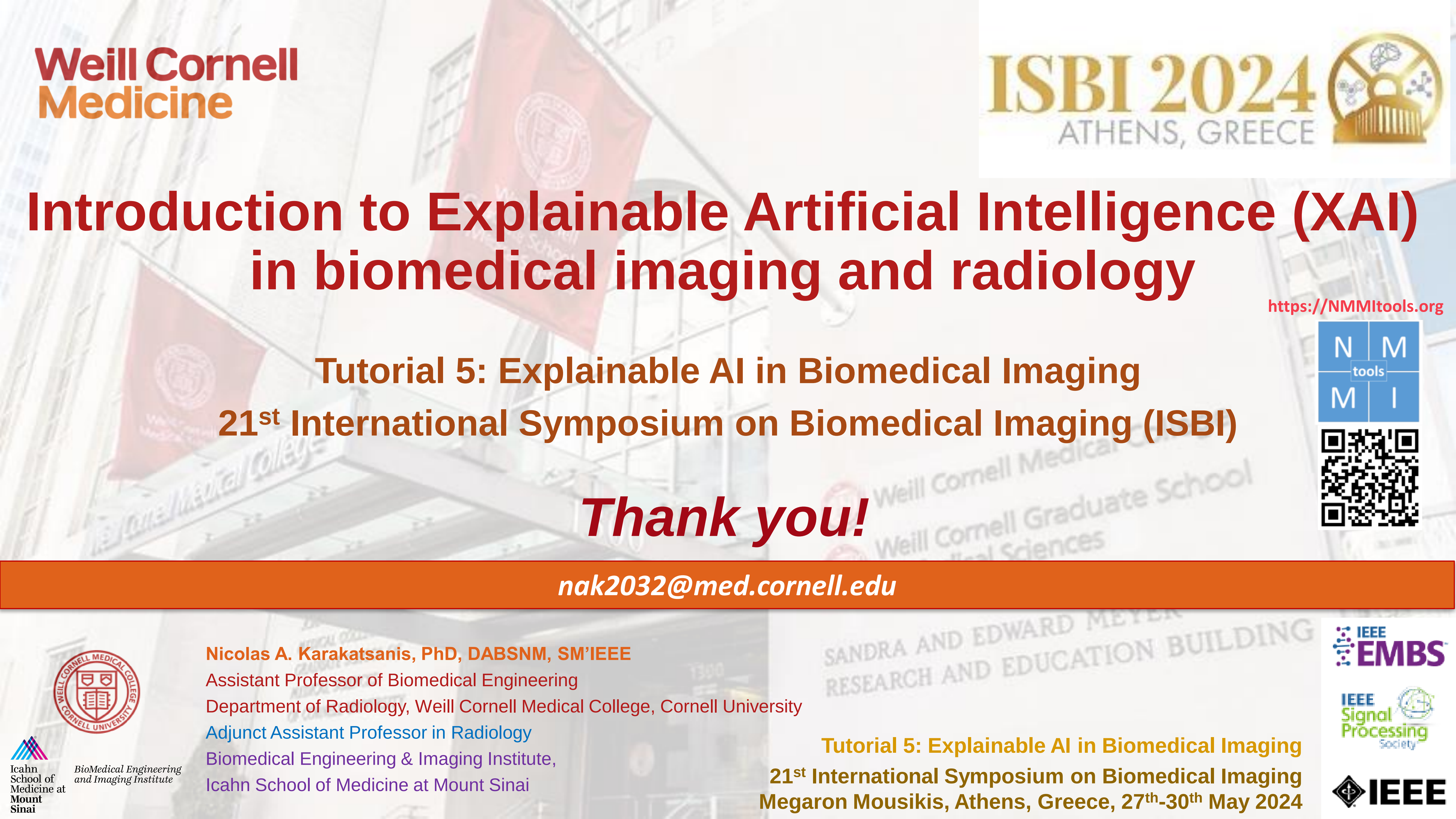
XAI’s crucial role towards the widespread clinical adoption of AI and DL in clinical radiology/nuclear medicine

The acceptance, proliferation, and widespread adoption of a system are directly proportional to the trust in that system.

Objectives of XAI in biomedical imaging and clinical radiology/nuclear medicine

- Increasing trust and reliability
- Encouraging collaboration between humans and artificial intelligence
- Reducing bias and discrimination
- Compliance with regulatory requirements

While exploring Explainable AI (XAI) can be beneficial, it is also a matter of being human. We want to know how things work in our subconscious, and the urge to understand that neural networks are ‘our’ creation is incredibly frustrating


Weill Cornell Medical College
Weill Cornell Medical College
Weill Cornell Graduate School of Medical Sciences
SANDRA AND EDWARD MEYER
RESEARCH AND EDUCATION BUILDING
1300

Weill Cornell Medicine

ISBI 2024
ATHENS, GREECE

Introduction to Explainable Artificial Intelligence (XAI) in biomedical imaging and radiology

https://NMMItools.org

N M tools M I

Tutorial 5: Explainable AI in Biomedical Imaging

21st International Symposium on Biomedical Imaging (ISBI)

Thank you!

nak2032@med.cornell.edu

Nicolas A. Karakatsanis, PhD, DABSNM, SM’IEEE
Assistant Professor of Biomedical Engineering
Department of Radiology, Weill Cornell Medical College, Cornell University
Adjunct Assistant Professor in Radiology
Biomedical Engineering & Imaging Institute,
Icahn School of Medicine at Mount Sinai

Icahn School of Medicine at Mount Sinai

BioMedical Engineering and Imaging Institute

Tutorial 5: Explainable AI in Biomedical Imaging
21st International Symposium on Biomedical Imaging
Megaron Mousikis, Athens, Greece, 27th-30th May 2024

IEEE EMBS

IEEE Signal Processing Society

IEEE