Interpretable Deep Learning for Cancer Detection: A Multi-Method XAI Analysis on the PCam Dataset

Academic Year: 2024–2025

Authors :

Chaimae SADOUNE
chaimae.sadoune@student-cs.fr

Manon LAGARDE
manon.lagarde@student-cs.fr

Supervised by:

Céline HUDELOT
celine.hudelot@centralesupelec.fr

Project Overview:

Context :

- Medical AI systems are increasingly used in diagnostics.
- Deep learning offers high accuracy but is often opaque (black-box models)

Problem :

- Clinicians, patients, and regulators need to understand and trust model decisions.

Solution :

- Use XAI techniques to interpret models trained on the PatchCamelyon dataset.

Project Objectives :

Our Objectives :

- To explore how explainability techniques can help **interpret deep learning models** used for cancer detection.
- To compare the **interpretability of convolutional neural networks (CNNs)** such as UNeT and MobileNetV2 with **attention-based architectures** like Vision Transformers (ViT).
- To evaluate multiple types of explanation methods: **gradient-based, perturbation-based, model-agnostic**, and **concept-based**.
- To assess the faithfulness and reliability of these methods using quantitative metrics including **Deletion, Insertion, and Fidelity.**

Our Problematics :

- How can we visualize and interpret the internal representations and decision logic of CNN- and Transformerbased models?
- What types of visual explanations are more trustworthy or faithful?
- Can concept-based explanations uncover higher-level patterns that correlate with clinical features (e.g., hyperchromaticity, tissue density, cellular disorganization)?

Dataset – PatchCamelyon (PCam) :

Dataset Overview :

- **Source :** Derived from Camelyon16 challenge.
- **Data :**
  - 327,680 RGB color image patches, extracted from histopathological scans of lymph node sections, each 96x96 pixels.
  - Binary labels: cancerous (1), benign (0).
- **Clinical Relevance :** Patch-level analysis can support faster pathology workflows.
- **Challenges :** Visual patterns are subtle and vary between samples.

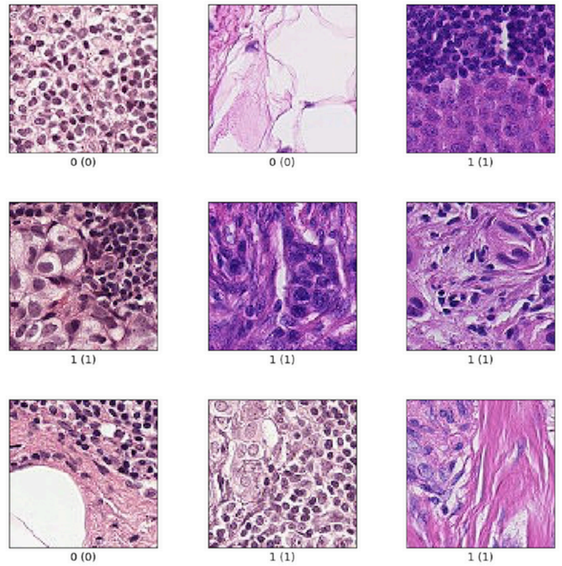
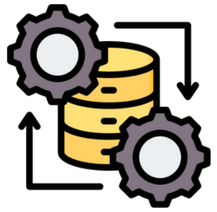


Figure 1. PCam dataset



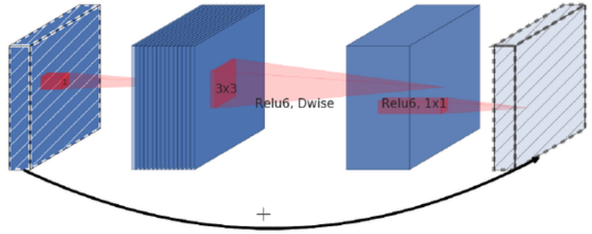
Data Preprocessing:

- Pixel values are normalized to the range [0, 1].
- The dataset is randomly split into 80% for training, 10% for validation, and 10% for testing.
- Images are resized to 224x224 pixels for compatibility with models like ViT and MobileNetV2.

Deep Learning Models :

MobileNetV2 :

- a **lightweight CNN architecture** optimized for performance on mobile devices.
- uses depthwise separable convolutions and inverted residuals with linear bottlenecks, to reduce computational cost while maintaining accuracy.
- In our work, it is used as a baseline for explainability evaluation.



UNeT:

- Encoder-decoder CNN architecture.
- Skip connections allow high spatial resolution.
- Used in segmentation; adapted here for classification by replacing its decoder with a classification head.

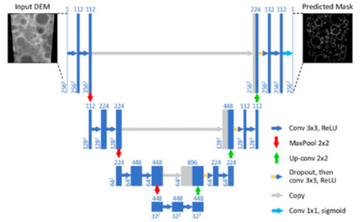


Figure 2. UNeT Architecture

ViT (Vision Transformer) :

- Treats input images as sequences of patches and processes them using self-attention.
- Learns global dependencies / attention patterns.
- Pretrained with DINO self-supervision.

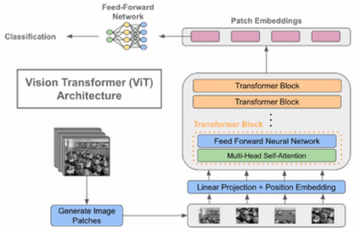


Figure 3. Structure of the Vision Transformer

Explainability Techniques Overview :

We investigate a wide range of explainability methods, categorized as follows :

Gradient-Based Methods : (Pixel-Level)
Saliency Maps : Calculate the gradient of the output with respect to input pixels.
Integrated Gradients : Accumulate gradients over interpolated inputs between a baseline and the actual input.
GradCAM : Generate class activation maps by combining gradient information with convolutional feature maps.

Perturbation-Based Methods :
Occlusion : Slide a blank mask over the image and measure prediction drop.
RISE : Apply random masking to estimate importance of regions.

Model-Agnostic Methods :
LIME : Train a local interpretable model (e.g., linear regression) on perturbed inputs.
SHAP : Use Shapley values from cooperative game theory to compute feature importances.

Concept-Based Methods :
CRAFT : Automatically extract and rank interpretable visual concepts from intermediate activations using Non-negative Matrix Factorization.

Faithfulness Evaluation Metrics :

To evaluate how accurately explanations reflect model decision-making, we use the following metrics :

Deletion : Important pixels, as indicated by the explanation, are progressively removed (e.g., set to zero). A faithful explanation leads to a rapid drop in the model's prediction confidence.

Insertion : Starting from a blank image, the most important pixels are progressively reintroduced. A faithful explanation causes a rapid recovery of the original prediction score.

Fidelity : Fidelity measures how closely the explanation aligns with the model's internal behavior, typically computed as a correlation or agreement between attribution scores and model sensitivity.

The metrics are computed using a fixed test set of malignant PCam samples across all methods and models.
Deletion captures how harmful removing important pixels is / Insertion captures how useful adding important pixels is.

MobileNetV2 – Quantitative XAI Results :

We evaluated twelve XAI methods on a pretrained MobileNetV2 model using malignant PCam patches. The evaluation used Deletion, Insertion, and Fidelity metrics.

Observations :

- **Occlusion** achieved the **best deletion score (-0.540)**, indicating high sensitivity to occluded regions.
- **LIME** recorded the **highest insertion score (1.508)**, showing that it successfully identifies regions that recover prediction when revealed.
- **Integrated Gradients** consistently performed well across all metrics, striking a good balance between faithfulness and interpretability.

Method	Deletion ↓	Insertion ↑	Fidelity ↑
Saliency	0.576	0.638	-0.022
GradientInput	0.163	0.483	0.065
GuidedBackprop	**-0.024**	0.369	0.045
IntegratedGradients	0.014	0.681	**0.096**
SmoothGrad	0.553	**1.004**	-0.076
SquareGrad	0.696	0.200	-0.054
VarGrad	0.709	0.255	-0.058
GradCAM	0.059	0.399	0.022
GradCAM++	0.225	0.411	-0.057
Occlusion	**-0.540**	**-0.163**	-0.014
RISE	-0.108	0.518	-0.011
LIME	0.086	**1.508**	0.088
KernelSHAP	0.766	0.948	0.008

Table 2. Deletion, Insertion and Fidelity scores for XAI on MobileNetV2 (best in bold).

↑ : Higher is better

↓ : Lower is better

Why does Occlusion perform best in the Deletion metric?

Occlusion performs best in the Deletion metric because it is fundamentally aligned with how Deletion is computed.

- Occlusion works by systematically masking regions of the input image and measuring how each mask affects the model's prediction.
- So when we evaluate Deletion (which also involves removing pixels), the logic is inherently similar.
- It directly tests the model's sensitivity to different parts of the input, leading to a very sharp drop in prediction confidence when important regions are removed.

MobileNetV2 – Quantitative XAI Results :

We evaluated twelve XAI methods on a pretrained MobileNetV2 model using malignant PCam patches. The evaluation used Deletion, Insertion, and Fidelity metrics.

Observations :

- **Occlusion** achieved the **best deletion score (-0.540)**, indicating high sensitivity to occluded regions.
- **LIME** recorded the **highest insertion score (1.508)**, showing that it successfully identifies regions that recover prediction when revealed.
- **Integrated Gradients** consistently performed well across all metrics, striking a good balance between faithfulness and interpretability.

Method	Deletion ↓	Insertion ↑	Fidelity ↑
Saliency	0.576	0.638	-0.022
GradientInput	0.163	0.483	0.065
GuidedBackprop	**-0.024**	0.369	0.045
IntegratedGradients	0.014	0.681	**0.096**
SmoothGrad	0.553	**1.004**	-0.076
SquareGrad	0.696	0.200	-0.054
VarGrad	0.709	0.255	-0.058
GradCAM	0.059	0.399	0.022
GradCAM++	0.225	0.411	-0.057
Occlusion	**-0.540**	**-0.163**	-0.014
RISE	-0.108	0.518	-0.011
LIME	0.086	**1.508**	0.088
KernelSHAP	0.766	0.948	0.008

Table 2. Deletion, Insertion and Fidelity scores for XAI on MobileNetV2 (best in bold).

↑ : Higher is better

↓ : Lower is better

Why does LIME get the best Insertion score despite being model-agnostic?

LIME performs well in the Insertion metric because it builds a local interpretable model (typically linear) around a specific prediction.

- Even though it is model-agnostic, LIME captures strong local correlations between input features and the output by generating many perturbations and fitting a simplified surrogate model.
- It identifies which regions contribute most to restoring the original prediction.
- This works particularly well in Insertion, where the task is to recover the prediction starting from a blank image.

➡ So while LIME doesn't rely on internal gradients, its data-driven approach allows it to effectively prioritize impactful pixels.

MobileNetV2 – Quantitative XAI Results :

We evaluated twelve XAI methods on a pretrained MobileNetV2 model using malignant PCam patches. The evaluation used Deletion, Insertion, and Fidelity metrics.

Observations :

- **Occlusion** achieved the **best deletion score (-0.540)**, indicating high sensitivity to occluded regions.
- **LIME** recorded the **highest insertion score (1.508)**, showing that it successfully identifies regions that recover prediction when revealed.
- **Integrated Gradients** consistently performed well across all metrics, striking a good balance between faithfulness and interpretability.

Method	Deletion ↓	Insertion ↑	Fidelity ↑
Saliency	0.576	0.638	-0.022
GradientInput	0.163	0.483	0.065
GuidedBackprop	**-0.024**	0.369	0.045
IntegratedGradients	0.014	0.681	**0.096**
SmoothGrad	0.553	**1.004**	-0.076
SquareGrad	0.696	0.200	-0.054
VarGrad	0.709	0.255	-0.058
GradCAM	0.059	0.399	0.022
GradCAM++	0.225	0.411	-0.057
Occlusion	**-0.540**	**-0.163**	-0.014
RISE	-0.108	0.518	-0.011
LIME	0.086	**1.508**	0.088
KernelSHAP	0.766	0.948	0.008

Table 2. Deletion, Insertion and Fidelity scores for XAI on MobileNetV2 (best in bold).

↑ : Higher is better

↓ : Lower is better

Integrated Gradients seems well-balanced ...

Integrated Gradients is a very good candidate as a **default XAI method**, especially for gradient-accessible models like CNNs.

- It offers a smooth approximation of feature attribution by integrating gradients over interpolated inputs from a baseline.
- In our results, it showed moderate deletion, good insertion, and the best fidelity, making it very stable and trustworthy.

It also avoids the noisiness of raw Saliency Maps and is less sensitive to gradient saturation

Concept-Based XAI with CRAFT - UNeT :

What Is CRAFT? A Shift from Pixels to Concepts :

CRAFT (Concept Recursive Activation Factorization for Explainability) is a **post-hoc explainability method**.

- It answers two critical questions:
  - Where does the model focus?
  - **What** is the model actually seeing?

- Instead of raw pixels, CRAFT works in the model's latent space—the internal representation learned by the network.
- It discovers **human-interpretable building blocks**, called **concepts**, from this space.

CRAFT Workflow :

- Crop patches from input images (e.g., 28×28 regions).
- Pass these through the model's intermediate layer (g(x)) to get latent activations.
- Apply **Non-Negative Matrix Factorization** (NMF) - decomposition is additive and interpretable—no negative or canceling features.
- Use **Sobol indices** to measure each concept's influence on the model's decision.
- Use implicit differentiation to compute **concept attribution maps**—visual heatmaps showing where each concept appears.

Concept-Based XAI with CRAFT - UNeT :

CRAFT finds and shows important visual patterns learned by the model.

Concepts :

- **Concept 0** : Hyperchromatic regions (dark nuclei) - commonly associated with malignancy.
- **Concept 4** : Hypercellularity (densely packed cells)

Clinical Alignment :

- Highlights areas linked to tumor characteristics.
- Enhances pathologist understanding of model behavior.

Figure 4. Concepts in images

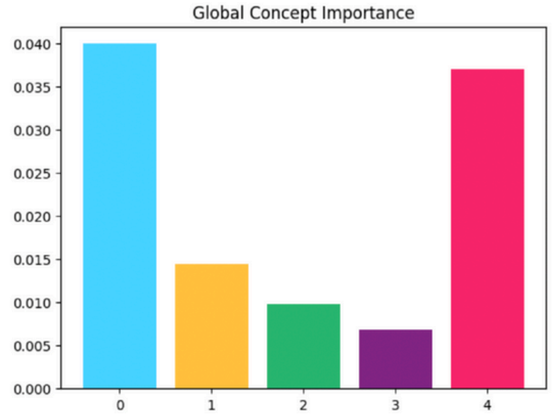
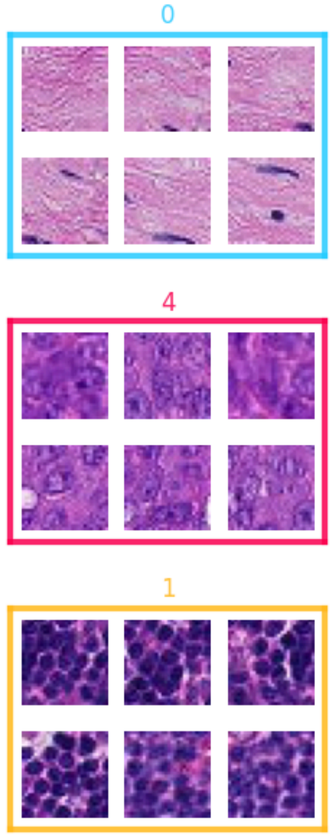
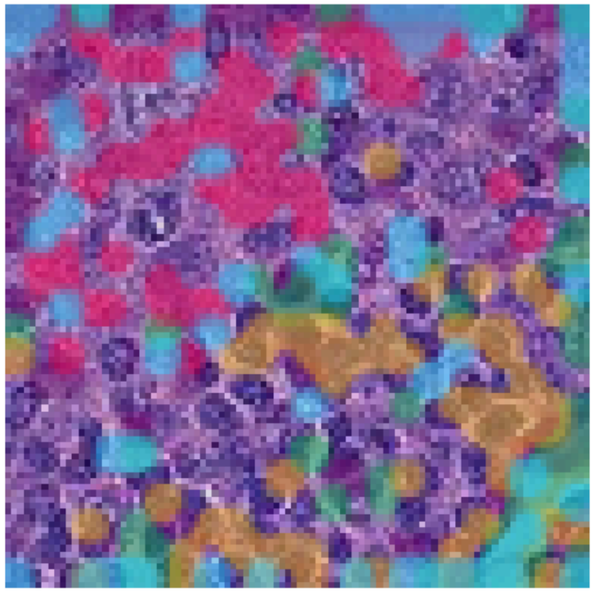


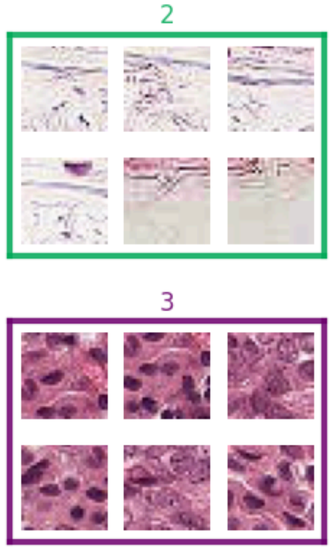
Figure 5. Top 5 most important CRAFT concepts extracted from UNeT

Concept-Based XAI with CRAFT – UNeT Visualization Tools:

Concept Crops :
Shows top image patches associated with each discovered concept.



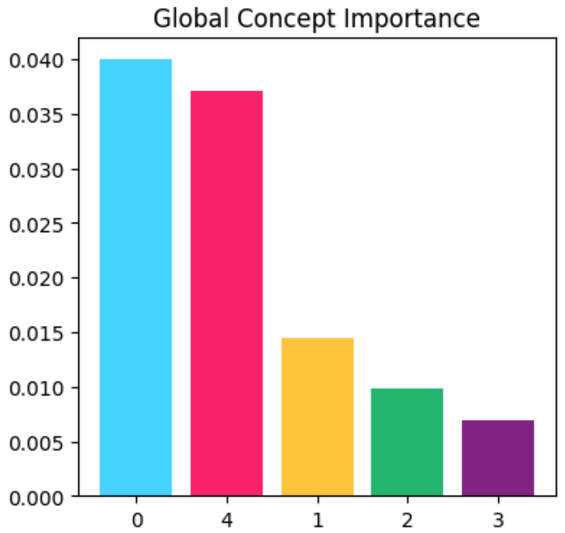




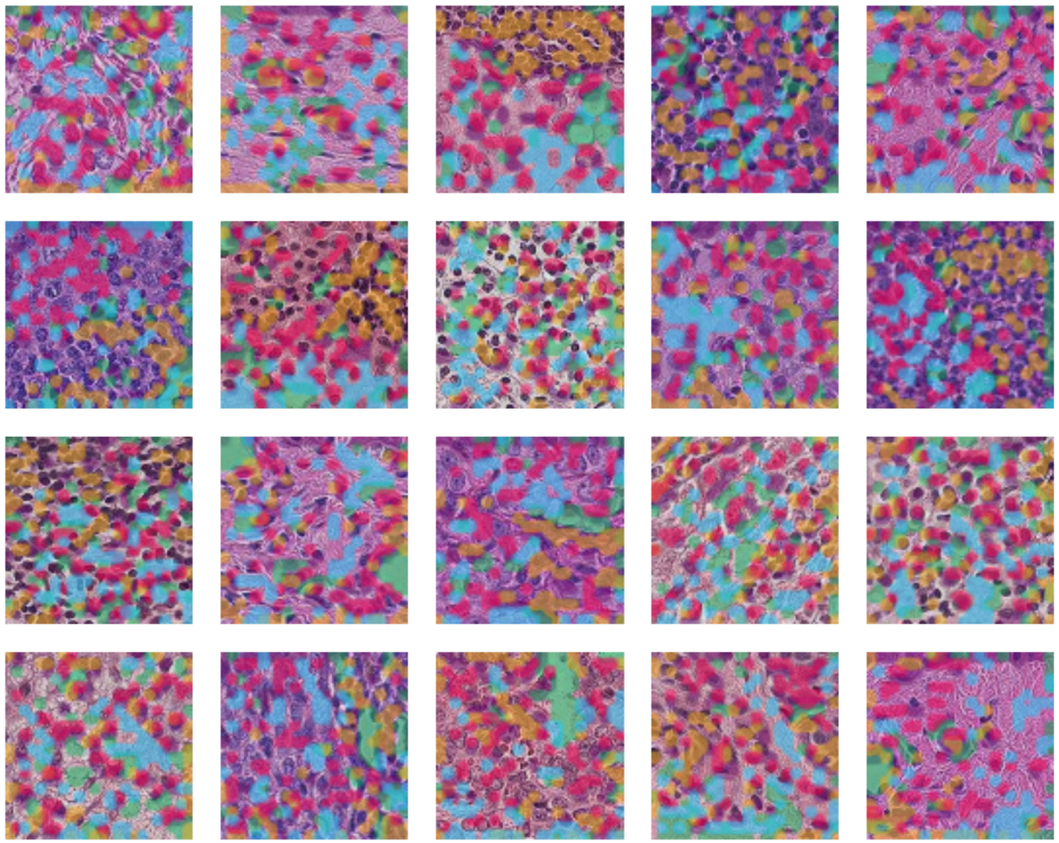
Attribution Maps :
Projects concepts back onto the image space, like heatmaps.

Concept Importances :
Visualizes the **Sobol sensitivity indice**s of each concept:

- Measures how much the model's output changes when you perturb that concept.
- High Sobol index → high influence on prediction.



Concept-Based XAI with CRAFT – UNeT Visualization Tools:



Per-Image Explanations

Shows which concepts are activated in a specific image and their strength.

CRAFT on ViT – Transformer Insights :

Key Insights :

- ViT produced fewer but more abstract and globally distributed concepts.
- For example, Concept 4 highlighted irregular tissue architecture and disorganized cellular regions—both are indicators of malignancy.

Figure 4. Concepts in images

Comparison with UNeT :

- **UNeT** : Captures **spatially detailed, localized features**.
- **ViT** : Captures **broader, texture-level** or **architectural patterns**.

These results show how model architecture influences the type and granularity of interpretable features.

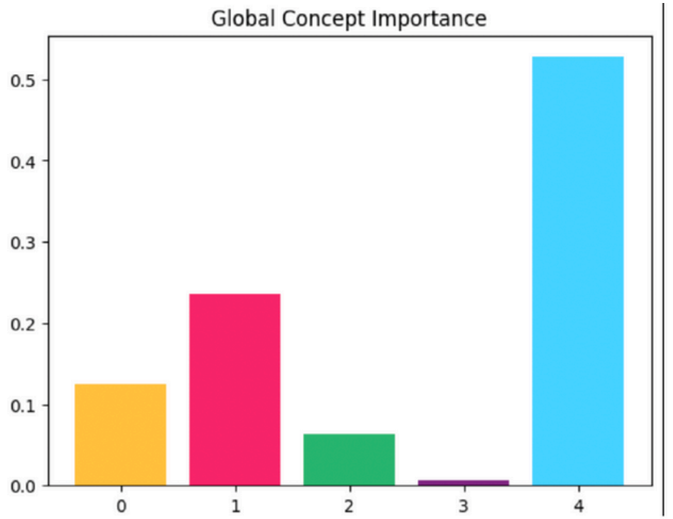


Figure 6. Global concept importance from CRAFT applied to ViT

Clinical Interpretability of CRAFT Concepts :

CRAFT enhances clinical interpretability by mapping high-level medical concepts like:

- **Hyperchromaticity** (darker nuclei)
- **Tissue architecture** (disorganization, structure)
- **Cell density** (hypercellularity)

Figure 4. Concepts in images

These concept maps are **reusable** across patients and samples, making them more informative than pixel-level heatmaps. The ability to interpret and reuse these concepts makes CRAFT particularly valuable in clinical decision-support tools.

- CRAFT works **post-hoc**, so no need to retrain the model.
- Concepts are extracted directly from activations, not manually defined.
- NMF ensures non-negative, additive combinations (Each part adds meaning.)—more interpretable.
- Sobol indices give quantitative faithfulness

XAI Techniques – Strengths & Weaknesses :

Method	Strenghts	Limitations
Saliency Maps	Fast, simple to cpmpute and Intuitive	Often noisy and unstable
GradCAM	Visually clear Heatmaps	Limited to CNNS with conv layers
Occlusion	High fidelity and robusteness	Slow, depends on patch size and computationally expensive
LIME	Model -agnostic, interpretable	Sensitive to input sampling, unstable
CRAFT	Reveals high-level concepts	More complex to set up, no pixel-wise

Project Takeaways :

Key Takeways :

- Pixel-based methods are quick but often noisy
- Perturbation methods are more faithful but costly
- Concept-based methods (CRAFT) provide the richest semantic insight
- Model architecture influences what can be explained (ViT vs. UNeT).

Future Directions :

- Collaborate with domain experts (e.g., pathologists) to annotate and validate learned concepts.
- Extend this framework to other medical domains like radiology or dermatology.
- Combine concept-based explanations with attention mechanisms for multi-layered interpretability.
- Explore how explanations can guide model refinement through human-in-the-loop systems or active learning.

CentraleSupélec | université PARIS-SACLAY

MScAI
Master of Science in Artificial Intelligence
Applied to society

Thank You