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

Chaimae SADOUNE
CentraleSupélec
Gif-sur-Yvette, France
chaimae.sadoune@student-cs.fr

Manon LAGARDE
CentraleSupélec
Gif-sur-Yvette, France
manon.lagarde@student-cs.fr

Abstract

In this project, we investigate explainability techniques applied to image classification models trained on the PatchCamelyon (PCam) dataset, a benchmark of histopathologic scans used for binary classification of metastatic tissue presence. We explore both traditional and concept-based explanation approaches to interpret the decision-making process of two distinct deep learning models: a U-Net, originally trained for brain segmentation, and a Vision Transformer (ViT). The visual explanation methods—GradCAM, Saliency Maps, and Occlusion—provided heatmaps highlighting influential regions in the input patches. GradCAM effectively localized malignant tissue areas, while Occlusion captured critical regions by observing variations in prediction confidence. We further applied CRAFT (Concept Recursive Activation Factorization), a concept-based XAI technique, which extracts interpretable visual concepts from intermediate activation layers using Non-negative Matrix Factorization (NMF). CRAFT enabled a multi-level analysis of the internal representations and offered concept attribution maps back to the input space. Evaluation of the different explanation methods was carried out using three faithfulness metrics: Deletion, Insertion, and Fidelity. Among all methods, Occlusion achieved the best performance in terms of deletion and fidelity, while LIME showed high insertion score. Our results emphasize the strengths and limitations of each XAI technique and highlight the value of concept-based methods for domain-relevant insights in medical imaging.

1. Introduction

Deep learning models have achieved remarkable success in medical imaging classification tasks. However, the complexity and opaqueness of these models make them difficult to interpret and trust, especially in high-stakes domains such as oncology. As a result, Explainable Artificial Intelligence (XAI) has become an essential area of research, aiming to provide transparency and interpretability to black-box models. In this project, we focus on the PatchCamelyon (PCam) dataset. The dataset offers a valuable benchmark to study automated cancer detection at the patch level, a task with direct implications for clinical diagnostics and pathology workflows. Our main goal is to explore how different XAI methods help interpret and understand the predictions of deep learning models trained on PCam. The challenge lies in interpreting complex spatial features from models like UNeT, which uses convolutional encoder-decoder architecture, and ViT (Vision Transformers), which relies on self-attention mechanisms over image patches. These models are powerful but notoriously difficult to understand.
Through this work, we aim to answer the following research questions:

- How can we visualize and interpret the internal representations and decision logic of CNN- and Transformer-based 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)?

Our project provides both a qualitative and quantitative comparison of explanation techniques, with the goal of enhancing trust and transparency in medical image analysis.

2. Problem Definition

2.1. Task Description

The problem addressed in this project is a **binary image classification** task involving histopathological images. Each image $x \in \mathbb{R}^{96\times96\times3}$ is a colored patch extracted from a lymph node section, taken from the PatchCamelyon (PCam) dataset. The classification goal is to determine whether the image contains metastatic cancerous tissue (label 1) or benign (non-cancerous) tissue (label 0).
To perform this task, we train a machine learning model—

specifically deep neural networks like UNeT or ViT—that learns a function:

$$f_\theta : \mathbb{R}^{96\times96\times3} \to [0,1]$$

Here, $f_\theta(x)$ denotes the predicted probability that the image x contains cancerous tissue. The parameter θ represents the model's internal weights learned during training.

The final binary prediction is obtained by thresholding the output probability at a fixed value $\tau \in [0,1]$. That is:

$$\hat{y} = \mathbb{1}\{f_\theta(x) > \tau\}$$

where $\mathbb{1}\{\cdot\}$ is the indicator function that outputs 1 if the condition is true and 0 otherwise. In practice, $\tau = 0.5$ is commonly used, but it can be tuned depending on clinical constraints (e.g., to reduce false negatives).

The objective during training is to maximize the classification accuracy over a test distribution $\mathcal{D}_{\text{test}}$. The model seeks the optimal parameters θ^* that lead to the highest expected number of correct predictions:

$$\theta^* = \arg\max_\theta \ \mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{test}}} \ \mathbb{I}[\hat{y} = y]$$

This optimization objective ensures that the learned model generalizes well to unseen data, providing reliable predictions on new histopathologic samples. However, as will be discussed in the following sections, high accuracy alone is not sufficient in medical applications—transparency and explainability are also crucial.

2.2. Need for Explainability

Although achieving high classification accuracy is crucial, it is not a sufficient condition for deploying deep learning models in sensitive domains such as medical diagnostics. In clinical settings, decisions made by AI systems must be interpretable and justifiable to a wide range of stakeholders—including clinicians, patients, and regulatory bodies.

Stakeholders require clear answers to questions such as:

- Why was a particular image classified as malignant or benign?
- Which regions of the image contributed most to the model's decision?
- Is the model focusing on medically relevant features such as tumor architecture, cellular density, or chromatic patterns?

In many cases, deep neural networks operate as "black boxes", providing predictions without insight into their internal reasoning. This lack of transparency can hinder clinical trust, raise ethical concerns, and make model validation challenging.

For these reasons, our project goes beyond traditional performance metrics (e.g., accuracy) to also evaluate the quality and reliability of explanations provided by different XAI techniques. We focus on maximizing the faithfulness of an explanation, i.e., how well it reflects the model's true decision-making process.

This objective ensures that the produced explanations are not only human-readable but also aligned with the model's internal behavior, making them more trustworthy and clinically actionable.

3. Related Work

3.1. Pixel-Level Attribution Methods

One of the earliest and most intuitive approaches to explainability in image classification is based on pixel-level attribution. These methods aim to highlight which regions of the input image most influence the model's prediction.

Saliency Maps, introduced by Simonyan et al., use gradient information to compute a heatmap that indicates pixel importance. Although simple and widely used, saliency maps are known to be noisy and often lack clear localization, especially in medical images where interpretability is critical.

Grad-CAM (Selvaraju et al.) improved upon this by combining class-specific gradients with activation maps from convolutional layers, resulting in smoother and more interpretable explanations. Due to its practicality and visual clarity, Grad-CAM has been adopted in numerous medical imaging studies to localize regions associated with diseases.

Occlusion Sensitivity (Zeiler and Fergus) offers another perspective: it involves systematically masking out patches of the image and recording the impact on model predictions. While computationally expensive, this technique provides a faithful estimate of which image parts are most critical to the decision.

3.2. Model-Agnostic Interpretability

Beyond gradients, model-agnostic techniques such as **LIME** and **SHAP** provide flexible alternatives that can be applied to any black-box model.

LIME (Ribeiro et al.) works by generating perturbed samples around a given input and fitting an interpretable surrogate model (e.g., linear regression) to approximate the local decision boundary. Although LIME offers model independence, it can produce inconsistent explanations, especially in high-dimensional image data.

SHAP (Lundberg and Lee) is based on cooperative game theory and attributes importance scores to features via Shapley values. It has strong theoretical guarantees and produces consistent explanations, but also suffers from scalability issues when applied to complex models and large images.

While these methods are powerful in structured data settings, their application to medical image classification remains challenging due to the spatial and high-dimensional nature of the inputs.

3.3. Concept-Based Explanations

A more recent trend in XAI is the shift from low-level pixel importance to high-level semantic interpretability. In medical contexts, this means identifying abstract concepts such as "dense nuclei" or "irregular tissue architecture" that align with expert reasoning.

TCAV (Testing with Concept Activation Vectors) was among the first to formalize this idea. It computes directional derivatives in the model's activation space with respect to predefined concept vectors, allowing users to assess how much a concept influences the output. However, TCAV requires manual labeling of concept examples, limiting its automation and scalability.

To address this, **CRAFT** (Fel et al.) proposes an unsupervised framework to extract meaningful visual concepts directly from the activations of a network. Using Non-negative Matrix Factorization (NMF), CRAFT identifies shared components—interpreted as "concepts"—and ranks them via Sobol sensitivity indices. It then produces concept attribution maps, projecting their influence back onto the input image. This method allows explanations that are both structured and more aligned with human reasoning.

3.4. Positioning Our Work

Our work builds directly on these foundations. First, we replicate and compare three classical pixel-level attribution methods—**Saliency Maps**, **Grad-CAM**, and **Occlusion**—on two distinct architectures: UNeT and Vision Transformer (ViT). These methods serve as baselines for visual localization of model attention.

Second, we implement and analyze **CRAFT** on both models, investigating the internal concept structures and how they differ between convolutional and attention-based networks. This dual-model comparison under a unified concept-based lens is, to our knowledge, novel in the Patch-Camelyon setting.

Finally, to assess the faithfulness and trustworthiness of each explanation method, we adopt quantitative evaluation metrics such as **Deletion**, **Insertion**, and **Fidelity**, following the framework proposed by Samek et al.. This allows us to go beyond qualitative visuals and provide a rigorous, comparative benchmark of XAI techniques in the context of cancer detection from histopathologic images.

4. Methodology

4.1. Data and Preprocessing

The dataset used in this project is the **PatchCamelyon (PCam)** dataset, a large-scale benchmark designed for binary classification of metastatic cancer in histopathologic images. It consists of **327,680 RGB image patches**, each of size 96×96 pixels, extracted from whole-slide images of lymph node sections stained with hematoxylin and eosin (H&E). Each image is annotated with a binary label:

- **1** if metastatic tissue is present,
- **0** otherwise.

The data is derived from the Camelyon16 challenge, but the PCam dataset was curated and downsampled to focus specifically on patch-level classification. This fine granularity makes it well suited for evaluating both model performance and the interpretability of local predictions.

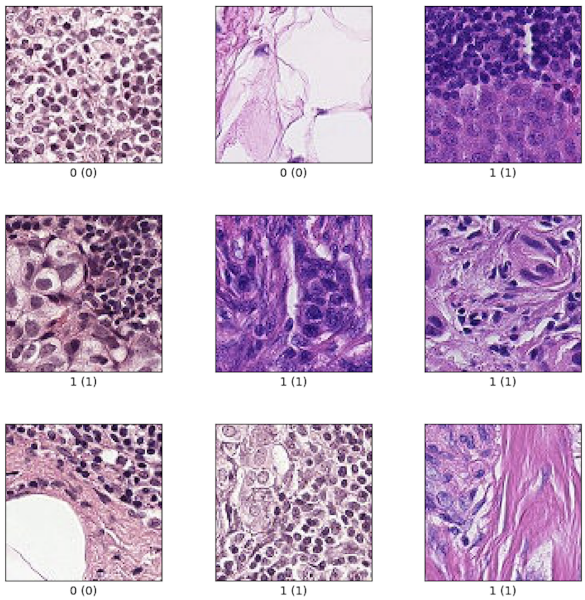

Figure 1. PCam dataset

To prepare the dataset for training and evaluation, the following preprocessing steps were applied:

- **Normalization** of pixel values to the $[0, 1]$ range.
- **Shuffling** and splitting into training (80%), validation (10%), and test (10%) sets.
- **Balancing the classes** to address the slight skew towards non-metastatic images.

In some cases (notably for the Vision Transformer), the input images were resized to fit the expected resolution of pretrained models (e.g., 224×224). We ensured that no image augmentation was applied during evaluation, in order to preserve the interpretability of attribution methods.

4.2. Model Architectures

In this study, we evaluate and interpret three deep learning models for binary image classification on the PCam dataset: a lightweight convolutional model **MobileNetV2**, a classical encoder-decoder **UNeT**, and a fully attention-based **Vision Transformer (ViT)**. These models offer complementary inductive biases and structural differences, allowing us to analyze how architecture influences both performance and the nature of visual explanations.

MobileNetV2

MobileNetV2 is a lightweight convolutional neural network architecture optimized for efficient inference on mobile and edge devices. It uses depthwise separable convolutions and inverted residual blocks with linear bottlenecks, significantly reducing computational complexity while retaining competitive accuracy.

In our work, we used a version of MobileNetV2 pretrained on ImageNet and fine-tuned its final classification layers on the PCam dataset. The low parameter count and modular architecture make MobileNetV2 ideal for testing XAI techniques, particularly in low-resource or embedded medical applications.

UNeT

The first model used is a pretrained UNeT, originally designed for medical image segmentation. UNeT features a symmetric encoder-decoder structure with skip connections between corresponding layers, allowing for precise localization and semantic context aggregation. While it was initially trained for brain tumor segmentation, we adapted its final layers to perform binary classification on the PCam patches by appending fully connected layers on top of the encoder's latent space.

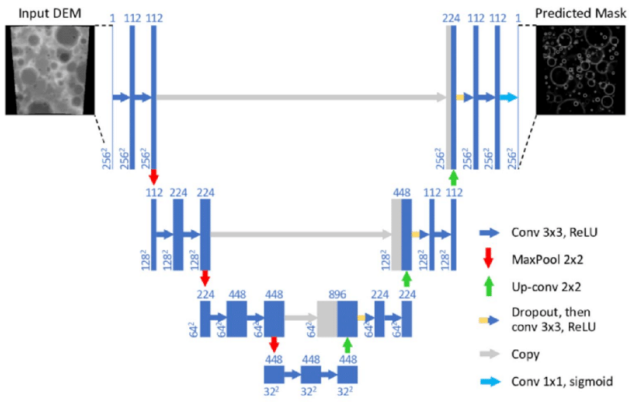

Figure 2. UNeT Architecture

The choice of UNeT is motivated by its widespread use in medical imaging, particularly due to its ability to preserve spatial resolution and anatomical consistency in predictions. These characteristics are especially relevant when applying visual attribution techniques such as Grad-CAM or CRAFT, which rely on structured feature maps.

Vision Transformer (ViT)

The second model is a DINO-pretrained Vision Transformer (ViT-S/16), which divide the input image into non-overlapping patches, linearly embed them, and process the resulting sequence using multi-head self-attention. Unlike convolutional models, ViTs learn global dependencies directly through attention, which may lead to more abstract or diffuse internal representations.

We selected a ViT pretrained using self-supervised DINO training, which has been shown to produce robust and semantically rich attention maps. The model was fine-tuned on the PCam dataset by replacing the classification head and retraining the last layers on the patch-level binary task.

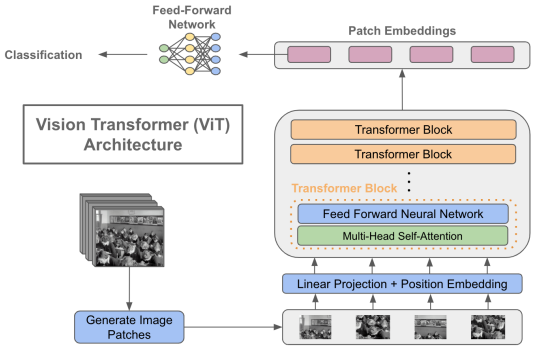

Figure 3. Structure of the Vision Transformer

This model offers an ideal counterpoint to UNeT, as it represents a fully attention-based approach with minimal architectural priors on locality.

4.3. Explanation Methods Implemented

4.3.1. Gradient-Based Attribution Methods

Gradient-based attribution methods explain a model's prediction by analyzing how the output changes with respect to small variations in the input features. These techniques assume differentiability and are most naturally suited to neural networks.

Saliency Maps

Saliency Maps are among the earliest and simplest visual explanation methods. They compute the gradient of the output class score with respect to each input pixel, yielding a sensitivity map:

$$S(x) = \left| \frac{\partial f_\theta^c(x)}{\partial x} \right|$$

where $f_\theta^c(x)$ is the score of class c. The absolute gradient magnitude at each pixel indicates how sensitive the prediction is to changes in that pixel.

Grad-CAM

Grad-CAM (Gradient-weighted Class Activation Mapping) improves upon raw gradients by combining them with the spatial activations from a convolutional layer. The method computes a class-specific localization map as follows:

1. Compute the gradients of the score $f_\theta^c(x)$ with respect to feature maps A^k of a chosen convolutional layer.

2. Average the gradients to obtain weights:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial f_\theta^c(x)}{\partial A_{i,j}^k}$$

3. Combine the feature maps:

$$L_{\text{Grad-CAM}}^c = \text{ReLU}\left(\sum_k \alpha_k^c A^k \right)$$

In our case, Saliency Maps and Grad-CAM were applied to a MobileNetV2 model pretrained on ImageNet, using malignant patches from the PCam dataset resized to 224×224 pixels. The model's final activation was set to linear to enable attribution methods based on logit outputs. Gradients and activation maps were computed using the `xplique` library.

4.3.2. Perturbation-Based Methods

Perturbation-based explanation methods evaluate the importance of input features by observing the impact of localized modifications (perturbations) on the model's output. Unlike gradient-based methods, they do not require backpropagation and can thus be applied to any black-box model.

Occlusion Sensitivity

Occlusion Sensitivity is a classical technique that involves systematically masking (or replacing) parts of the input image and measuring the variation in the predicted probability. The rationale is that if occluding a region significantly alters the output, then this region must be important for the prediction.
Let x be the original input image and $x_{[i,j]}^{\text{occluded}}$ denote the image with a patch centered at (i, j) replaced by a neutral value (e.g., zero or mean color). The importance of pixel (i, j) is then approximated by:

$$S(i,j) = f_\theta^c(x) - f_\theta^c(x_{[i,j]}^{\text{occluded}})$$

This process is repeated for all pixel regions by sliding a fixed-size window across the image.

4.3.3. Model-Agnostic Methods

Model-agnostic explanation methods are designed to work with any machine learning model, regardless of its internal architecture. They typically function by approximating the model's local behavior around a specific input using interpretable surrogate models.

LIME

LIME (Local Interpretable Model-agnostic Explanations) approximates the black-box model f_θ locally with an interpretable model g (e.g., linear regression or decision tree). The goal is to understand the decision of f_θ around a specific input x by sampling perturbed versions of x, generating predictions for each, and fitting g:

$$g = \arg\min_{g \in G} \ \mathcal{L}(f_\theta, g, \pi_x) + \Omega(g)$$

where $\mathcal{L}$ is a loss function measuring how well g approximates f_θ in the neighborhood of x, defined by the kernel π_x, and $\Omega(g)$ penalizes the complexity of g.

SHAP

SHAP (SHapley Additive exPlanations) is based on game theory and computes Shapley values to assign each feature an importance score that reflects its marginal contribution to the prediction.
For a given input x, SHAP estimates the prediction as a sum of feature contributions:

$$f_\theta(x) \approx \phi_0 + \sum_{i=1}^{n} \phi_i$$

where ϕ_i is the contribution of feature i, and ϕ_0 is the expected value of the model's output.

Other attribution methods—such as KernelSHAP, Integrated Gradients, SmoothGrad, and RISE—were also tested on MobileNetV2

4.3.4. Concept-Based Explanation

While pixel-based explanations help identify "where" a model is looking, they do not explain "what" the model has learned. Concept-based explanations address this gap by identifying human- or machine-discovered semantic patterns—called *concepts*—and linking them to model decisions.

CRAFT: Concept Recursive Activation Factorization

CRAFT (Concept Recursive Activation Factorization) is a recent concept-based explanation framework that extracts and ranks meaningful visual concepts from neural networks without requiring manual concept annotation. It is especially useful in domains like medical imaging, where the internal reasoning of models is difficult to interpret through low-level visualizations.
The pipeline involves five main steps:

1. Model Splitting. The original network f_θ is decomposed into two parts: a feature extractor g and a classifier h, such that:

$$f_\theta(x) = h(g(x))$$

The split is applied at a layer L selected for its semantic richness (e.g., bottleneck layer of UNeT or intermediate transformer block).

2. Activation Collection. A large number of image crops (e.g., 96×96 patches) are passed through g, and their corresponding activation maps at layer L are extracted. These are flattened and stacked into a matrix $A \in \mathbb{R}^{N \times D}$, where N is the number of samples and D the dimensionality of the activations.

3. Non-negative Matrix Factorization (NMF). CRAFT factorizes the activation matrix into K latent concepts via NMF:

$$A \approx UW \quad \text{with } U \in \mathbb{R}^{N \times K}, W \in \mathbb{R}^{K \times D}$$

Each row of W represents a learned visual concept in activation space. Each row of U gives the concept activation strength for a given sample.

4. Concept Ranking with Sobol Sensitivity. CRAFT ranks the discovered concepts using Sobol indices, which estimate the global importance of each concept on the model's output:

$$\mathcal{S}_k = \frac{\text{Var}_{u_k}[\mathbb{E}[f(u)|u_k]]}{\text{Var}[f(u)]}$$

This allows selecting the most influential concepts, i.e., those contributing most to model predictions.

5. Concept Attribution Maps. Finally, each concept is projected back into the input image space via activation maximization. This produces interpretable heatmaps indicating the spatial footprint of each concept.

We applied CRAFT to both UNeT and ViT. For UNeT, the split was performed at the bottleneck layer; for ViT, the output of the last attention block was used. In both cases, we extracted $K = 10$ concepts and computed their Sobol indices to identify the most influential ones.
This allowed us to visualize structured concepts such as cellular clusters, tissue texture, and nuclear regions. Unlike pixel-level explanations, these concepts were reusable across images and more interpretable for domain experts.

5. Evaluation and Results

5.1. Evaluation Metrics

To objectively assess the quality of the explanations generated by different XAI methods, we rely on three widely-used quantitative metrics: **Fidelity**, **Deletion**, and **Insertion**. These metrics are model-centric, meaning they evaluate how well an explanation reflects the model's actual decision process, rather than relying on human judgment.

5.1.1. Fidelity

Fidelity measures how accurately an explanation reflects the model's behavior. Given an explanation heatmap, pixels (or regions) identified as important should, when removed or perturbed, significantly affect the model's prediction. High fidelity indicates that the explanation correctly identifies the influential areas. Fidelity is often used as a generic criterion that underpins both deletion and insertion curves, as detailed below.

5.1.2. Deletion Metric

The deletion metric evaluates explanation faithfulness by progressively removing (zeroing out) the most important pixels according to the explanation map and observing the drop in the model's output score. Let x be the input, and $f_\theta(x)$ the original output score. The input is modified by masking top-$k\%$ pixels according to the explanation heatmap. A faithful explanation should result in a rapid drop in prediction confidence:

$$\text{Deletion curve: } D(k) = f_\theta(x_{\text{masked, top-}k\%})$$

We compute the Area Under the Curve (AUC); lower AUC indicates better deletion performance.

5.1.3. Insertion Metric

Conversely, the insertion metric starts with a fully masked image and gradually reintroduces the most important pixels (as indicated by the explanation). If the explanation is faithful, the model's confidence should increase quickly:

$$\text{Insertion curve: } I(k) = f_\theta(x_{\text{revealed, top-}k\%})$$

Here, higher AUC indicates better insertion performance.

Together, these metrics allow us to compare the explanatory power of different methods in a model-aligned and architecture-agnostic way. They were computed on a fixed test set of image patches for our models, and results are reported in the next section.

5.2. Results

MobileNetV2 Architecture

We evaluated twelve post-hoc explanation methods on a pretrained MobileNetV2 model using malignant image patches from the PCam dataset. The methods include gradient-based techniques (Saliency, GradientInput, GuidedBackprop, etc.), perturbation-based (Occlusion, RISE), and model-agnostic ones (LIME, KernelSHAP).
To quantitatively assess their performance, we computed tree metrics: **Deletion** (lower is better), **Insertion** and **Mu-Fidelity**. Each metric was applied to the same subset of resized inputs (224×224), using the `xplique` library with consistent hyperparameters.

Table 1. Deletion and Insertion scores for XAI on MobileNetV2 (best in bold).

Method	Deletion ↓	Insertion ↑
Saliency	0.576	0.638
GradientInput	0.163	0.483
GuidedBackprop	**-0.024**	0.369
IntegratedGradients	0.014	0.681
SmoothGrad	0.553	**1.004**
SquareGrad	0.696	0.200
VarGrad	0.709	0.255
GradCAM	0.059	0.399
GradCAM++	0.225	0.411
Occlusion	**-0.540**	**-0.163**
RISE	-0.108	0.518
LIME	0.086	**1.508**
KernelSHAP	0.766	0.948

Table 2. Fidelity scores for XAI on MobileNetV2 (higher is better).

Method	Fidelity ↑
Saliency	-0.022
GradientInput	0.065
GuidedBackprop	0.045
IntegratedGradients	**0.096**
SmoothGrad	-0.076
SquareGrad	-0.054
VarGrad	-0.058
GradCAM	0.022
GradCAM++	-0.057
Occlusion	-0.014
RISE	-0.011
LIME	0.088
KernelSHAP	0.008

Analysis. From the deletion metric, Occlusion and GuidedBackprop obtained the best scores, indicating strong sensitivity to important regions. However, Occlusion's insertion score was negative, suggesting instability when information is reintroduced progressively. In contrast, LIME and SmoothGrad achieved the highest insertion scores, though their deletion and fidelity values were weaker.

IntegratedGradients consistently ranked among the top performers across all metrics, combining good deletion, insertion, and fidelity scores. Among model-agnostic methods, LIME outperformed KernelSHAP in both fidelity and insertion.

Overall, while no single method dominated in all aspects, IntegratedGradients and GradCAM offered a good balance of faithfulness and interpretability in this setting.

UNeT Architecture: CRAFT

The concepts extracted by CRAFT reveal that the UNeT model strongly relies on low-level visual cues to make its predictions. The most influential concepts (e.g., Concept 0 and Concept 4) correspond to darker or more saturated tissue regions—often associated with **hyperchromatic nuclei**, a hallmark of malignancy in histopathology.
Additionally, other concepts highlight differences in cellular density: some patches correspond to **hypercellularity** (clusters of tightly packed cells), while others reflect **hypocellularity** (regions with sparse cell distribution). This suggests that the model is not merely responding to color, but also capturing meaningful structural variations that are diagnostically relevant. These insights confirm that the UNeT model is attending to pathology-relevant features, which enhances the trustworthiness of its predictions.

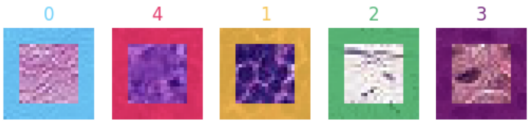

Figure 4. Concepts in images

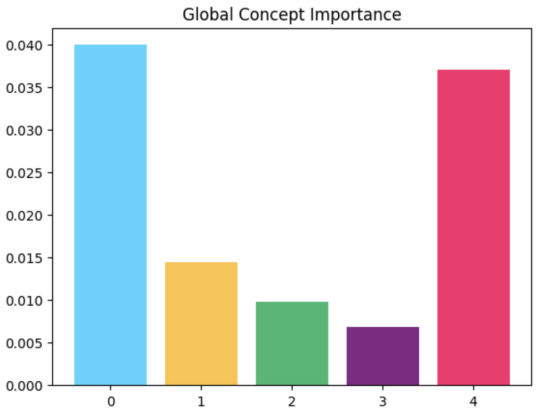

Figure 5. Top 5 most important CRAFT concepts extracted from UNeT

VIT Architecture : Craft

We also applied CRAFT to a fine-tuned Vision Transformer (ViT-S/16, pretrained with DINO), modified for binary classification. After training the classification head, we split the model into two components: g, the self-attention encoder, and f, the added linear classification head.

Observations. CRAFT results on ViT highlight the model's strong reliance on a limited set of abstract patterns.

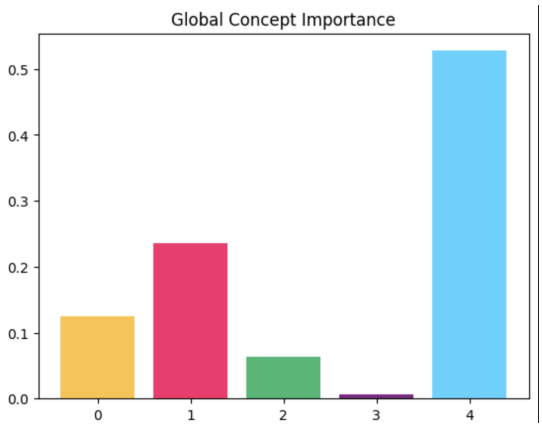

Figure 6. Global concept importance from CRAFT applied to ViT

This aligns with the ViT architecture, which encodes long-range dependencies via self-attention rather than focusing on fine-grained local features. The sparsity of important concepts suggests that ViT forms compact, possibly more generalized internal representations, in contrast to the more distributed and diverse features observed in UNeT.
While concept 4 may correspond to a global texture or color configuration associated with malignancy, further visual inspection would be needed to align it with medically interpretable structures.

ViT focuses more than UNeT on the cellular density and tissue architecture. Malignant tissues may show disorganized architecture, loss of tissue boundaries, and invasion into surrounding structures.

Conclusion

In this project, we investigated the interpretability of deep learning models for binary classification of histopathological images from the PatchCamelyon (PCam) dataset. We explored and compared a range of explanation techniques—including gradient-based, perturbation-based, model-agnostic, and concept-based methods—applied to three distinct architectures: MobileNetV2, UNeT, and Vision Transformer (ViT-S/16).
Our experiments demonstrated that while pixel-level attribution methods (such as Integrated Gradients and Grad-CAM) can provide useful insights, they often produce noisy or overly localized explanations, especially when applied to lightweight architectures like MobileNetV2. Perturbation and model-agnostic methods (e.g., Occlusion, LIME) offered more faithful but computationally costly alternatives.
Concept-based explanations using CRAFT proved particularly informative when applied to deeper models like UNeT and ViT. They revealed reusable, high-level visual patterns that aligned with medical concepts such as hyperchromasia, hypercellularity, and tissue texture. These insights not only enhance the transparency of the models but also support their integration into clinical decision-making pipelines.
Overall, this work highlights the importance of aligning explanation methods with model architectures, and the potential of concept-based approaches to bridge the gap between machine perception and human interpretability. Future work could extend this study by incorporating expert evaluation, exploring other concept learning frameworks, or integrating explanations into active learning or human-in-the-loop systems.

References

[1] Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Deep inside convolutional networks: Visualising image classification models and saliency maps.

[2] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[3] Zeiler, M. D., & Fergus, R. (2014). Visualizing and Understanding Convolutional Networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[4] Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 30.

[5] Fel, T., Picard, A., Bethune, L., Boissin, T., Vigouroux, D., Colin, J., Cadène, R., & Serre, T. (2023). CRAFT: Concept Recursive Activation Factorization for Explainability. *arXiv preprint arXiv:2301.11574*.

[6] Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., & Sayres, R. (2018). Interpretability Beyond Feature Attribution: Testing with Concept Activation Vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning (ICML)*.

[7] Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging Properties in Self-Supervised Vision Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 9650–9660.

[8] Veeling, B. S., Linmans, J., Winkens, J., Cohen, T., & Welling, M. (2018). Rotation Equivariant CNNs for Digital Pathology. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 210–218.