Interpretable and Explainable Machine Learning

Author: Bc. Šimon Šmída (xsmida03@vutbr.cz)
Supervisor: prof. Ing. Lukáš Sekanina, PhD.

Interpretability and Explainability

- **Interpretability:** *"How the model functions internally to make decisions?"*
- **Explainability:** *"Why the model produced a particular decision?"*

Tradeoff: Model Performance vs. Interpretability

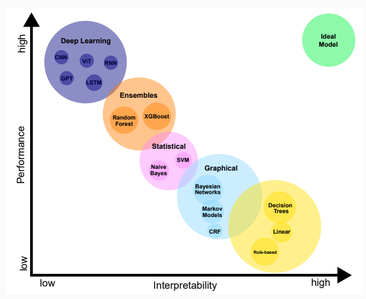


Figure 1: Trade-off between interpretability and performance across ML families [2]: rule-based/linear models are highly transparent but less accurate; graphical, statistical, and ensemble methods balance both; deep learning maximizes performance at the cost of interpretability; the "Ideal Model" remains theoretical.

Taxonomy of Existing XAI Methods

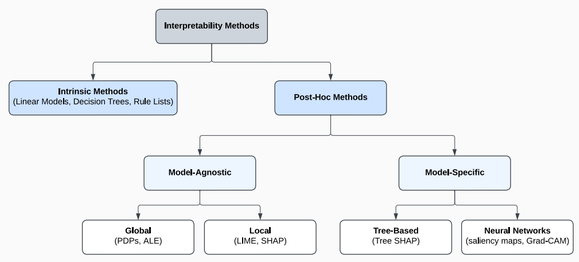


Figure 2: Overview of ML interpretability methods, categorized into intrinsic methods and post-hoc methods, which are further divided into model-agnostic and model-specific approaches.

Experimental Methodology

- The case study is set in the domain of **computer vision**, focusing on **image classification**.
- Experiments were designed to progressively assess selected XAI methods across **datasets of varying complexity**: MNIST, CIFAR-10, and ImageNet.

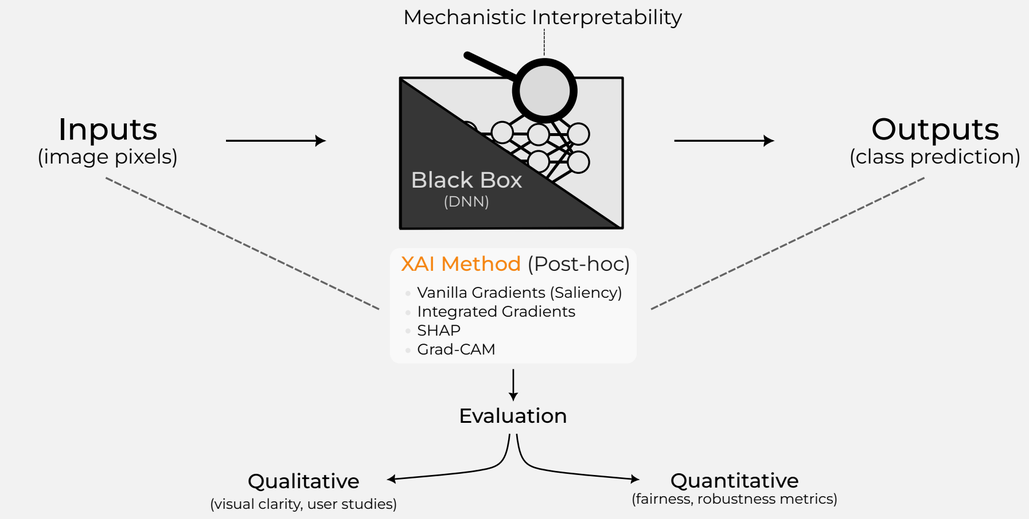


Figure 3: Interpretability workflow for DNNs: mechanistic analysis via inner inspection and post-hoc XAI methods (saliency, integrated gradients, SHAP, Grad-CAM) with ability to evaluate via qualitative (visual clarity, user studies) and quantitative (fairness, robustness) metrics..

Experimental Analysis

- The **analysis** targeted following dimensions:
  - (1) Relationship between **model quality and explanation clarity**
  - (2) **Robustness** under realistic perturbations (natural noise, invariant transformations)
  - (3) **Adversarial attacks** (on both models and XAI explanations)
  - (4) Effectiveness of **ensemble** and **mechanistic interpretability** approaches

Model Quality vs. Explanation Quality

Setting:

- Custom-trained CNN models with controlled generalization levels (underfit, properly-trained, overfit) were evaluated on MNIST dataset

Analysis:

- Clear correlation: explanations produced by properly-trained model were consistent and useful, while explanations from underfit or overfit models often appeared fragmented or misleading

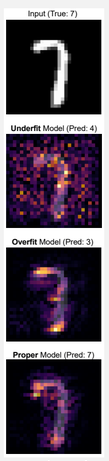


Figure 4: Comparing **impact of model quality** on explanation.

Selected XAI Methods **Comparison** and **Evaluation**

Setting:

- Analysis of selected XAI methods on datasets of varying complexity (MNIST, CIFAR-10, ImageNet)

Analysis:

- **Qualitative** and **quantitative** evaluation of the methods

Figure 6: Qualitative assessment of selected XAI methods (ImageNet).

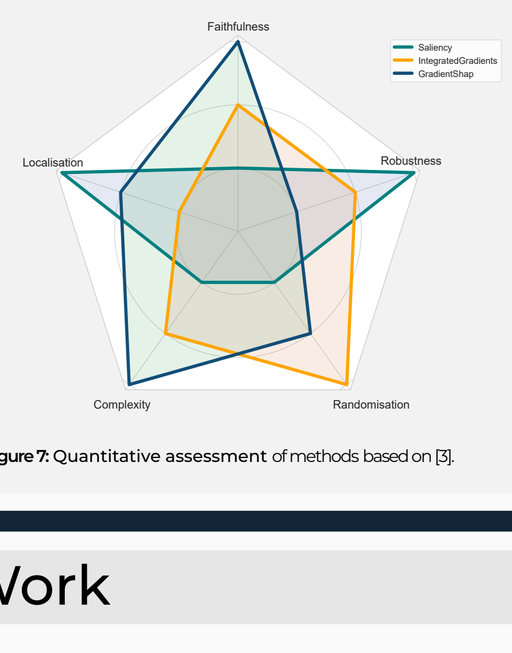
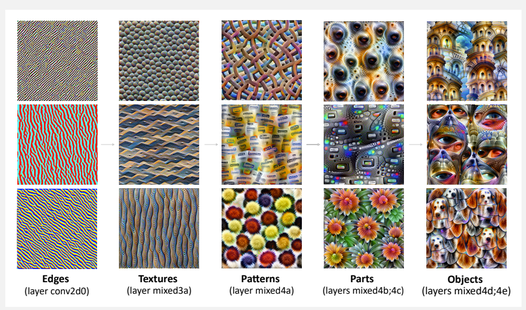


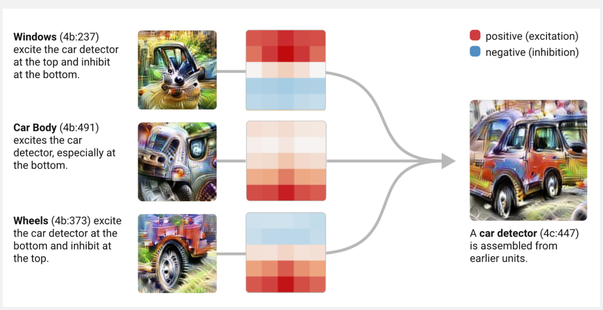
Figure 7: Quantitative assessment of methods based on [3].

Mechanistic Approach

- Mechanistic Interpretability for **analyzing InceptionV1** image classification network (ImageNet).



(a) **Feature complexity** increases across layers.



(b) **Car detector** emerges by composing lower-level features into a circuit.

Figure 5: (a) Progressive feature development across network layers. Layers detect edges and textures, while deeper layers capture object parts and higher-level concepts. **(b)** Circuit-based composition of a car detector. The visualization illustrates how a high-level car detector emerges from lower-level feature detectors through learned connections. The Windows, Car Body, and Wheels units contribute excitatory (red) and inhibitory (blue) signals, forming a neural circuit that integrates these components into a coherent car representation. Adapted from [1], CC BY 4.0

Robustness Evaluation of Explanations

Setting:

- Robustness was assessed using realistic perturbations

Analysis:

- Gaussian Noise
- Invariance Transformations Tests
- Adversarial Attacks

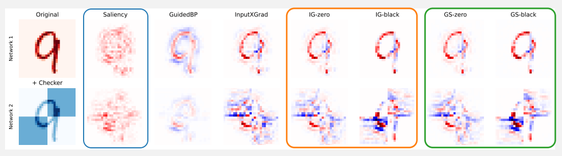
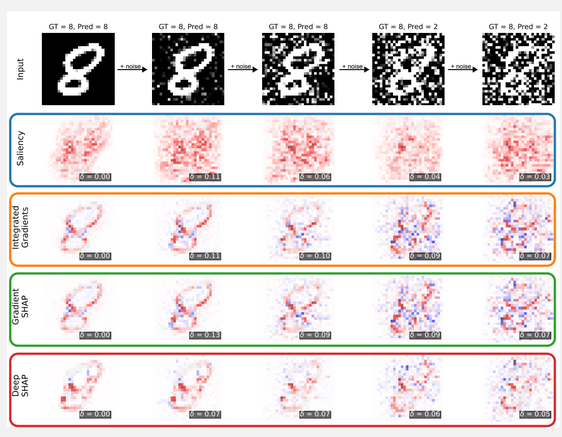
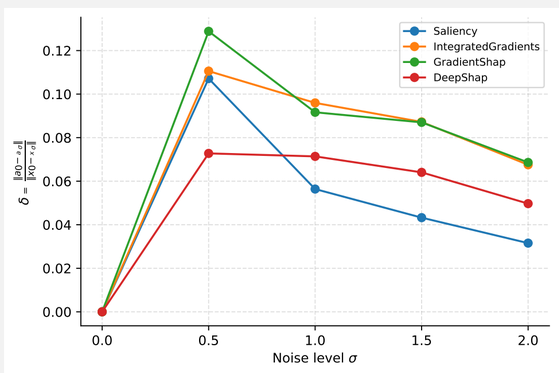


Figure 8: Effect of invariant input transformations on explanation quality. Shows robustness of GuidedBP and sensitivity of Integrated Gradients (IG) and Gradient SHAP (GS) to **baseline choice.**



(a) Explanations under Gaussian noise



(b) Relative attribution change

Figure 9: (a) **Impact of natural noise** on explanation quality across XAI methods. **(b)** Lower relative attribution change (δ) indicates **greater stability.**

Summary and Future Work

The experiments revealed valuable insights into the practical use of interpretability and explainability methods.

The study provides a systematic **experimental assessment of XAI methods** in the field of computer vision (image classification). It clarifies dependencies between model quality and explanation clarity, quantifies robustness issues, and demonstrates practical approaches of ensembles or mechanistic interpretability for enhancing explanatory quality in order to help **open up the black-box** nature of the ML models.

Future work should extend interpretability research to other, underexplored domains such as transformer architectures. Further directions include mapping internal representations to semantic concepts via mechanistic approaches, deepening the study of ensemble methods, and systematically evaluating explanation robustness across diverse models and tasks.

References

[1] Olah, et al., "Zoom In: An Introduction to Circuits", Distill, 2020.
[2] Viswan, V., Shaffi, N., Mahmud, M. et al. Explainable Artificial Intelligence in Alzheimer's Disease Classification: A Systematic Review. Cogn Comput 16, 1–44 (2024).
[3] Hedström et al. "Quantus: An Explainable AI Toolkit for Responsible Evaluation of Neural Network Explanations and Beyond," Journal of Machine Learning Research, vol. 24, no. 34, pp. 1–11, 2023.