BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2025
http://excel.fit.vutbr.cz

Interpretable and Explainable Machine Learning

Bc. Šimon Šmída*

Abstract

Machine learning (ML) models, particularly deep neural networks (DNNs), have achieved state-of-the-art performance across complex tasks. Yet, their opaque "black-box" nature limits trustworthiness in sensitive domains (e.g., healthcare, autonomous driving). This work systematically analyzes methods of interpretability and explainability (XAI) of DNNs in computer vision, covering model-agnostic, model-specific, gradient-based, and mechanistic approaches. Experiments on image classification datasets of increasing complexity (MNIST → ImageNet) reveal the strengths and limitations of selected methods under natural noise, invariance shifts, and adversarial attacks. The insights provide valuable guidance for selecting XAI techniques, improving model transparency, and deepening understanding of complex models.

*xsmida03@vutbr.cz, *Faculty of Information Technology, Brno University of Technology*

1. Introduction

The impressive performance of DNNs has driven their widespread adoption across diverse fields, including healthcare, finance, and autonomous driving. However, their complexity poses significant challenges for interpretability, accountability, and trustworthiness [1]. This work addresses limited transparency by systematically evaluating XAI methods, focusing on their ability to clarify model predictions reliably.

Effective interpretability methods must provide stable, comprehensible, and accurate explanations that facilitate trust in model decisions. Both qualitative (visual clarity of explanations) and quantitative (fairness, robustness, complexity metrics) criteria are essential for evaluating the quality of explanations.

Existing solutions to enhancing model interpretability include intrinsically interpretable models (e.g, decision trees), which offer a degree of transparency but might struggle with complex, high-dimensional data [2], and post-hoc methods (e.g, SHAP, Integrated Gradients), which generate explanations *after* model training but vary in stability and robustness [1].

This study compares these post-hoc interpretability methods through structured experimentation, identifying their practical strengths, limitations, and conditions affecting their reliability.

2. Selected XAI Methods

Interpretability and explainability, though related, address distinct aspects of understanding ML models. Interpretability typically refers to understanding the model's internal workings, while explainability focuses explicitly on justifying individual decisions [3].

SHAP, a model-agnostic technique grounded in cooperative game theory, offers theoretically sound attributions but suffers from high computational cost [4]. Model-specific methods often exploit internal gradients: vanilla gradients (saliency) are simple but noisy; Integrated Gradients improve attribution reliability via path integration [5]

These methods were chosen for their diversity, adoption, and complementary strengths, enabling robust and meaningful comparative evaluation.

3. Experimental Methodology

The case study is set in the domain of computer vision, focusing on image classification. Experiments were designed to progressively assess XAI methods across datasets of varying complexity: MNIST, CIFAR-10, and ImageNet. The evaluation targeted three primary dimensions: the relationship between model quality and explanation clarity, robustness under realistic perturbations (natural noise), adversarial attacks (on both models and explanations), and effectiveness of ensemble and mechanistic interpretability approaches.

3.1 Model Quality vs. Explanation Quality

Custom-trained CNN models with controlled generalization levels (underfit, properly trained, and overfit) were evaluated on MNIST. Explanation methods were qualitatively assessed via heatmap visualizations and quantitatively analyzed using metrics from the Quantus library [6]. The experiments confirmed a clear correlation: explanations produced by well-trained models were consistent and interpretable, whereas explanations from underfit or overfit models often appeared fragmented or misleading.

3.2 Robustness Evaluation of Explanations

Robustness was assessed using realistic perturbations: Gaussian noise, invariance transformations, and adversarial attacks. Integrated Gradients demonstrated superior robustness to Gaussian noise and invariance shifts, where vanilla gradients were sensitive. Ensemble approaches, aggregating multiple explanation methods, provided significant robustness improvements, suggesting their utility in adversarial contexts.

3.3 Mechanistic Interpretability and Ensembles

Mechanistic interpretability was explored as a complementary perspective, analyzing the internal representation structures (features, circuits) of CNNs. This structural analysis supplemented the pixel-wise explanations from post-hoc methods, offering deeper insight into the decision-making process [7]. Ensemble explanations further enhanced stability and mitigated method-specific biases and vulnerabilities.

4. Discussion of Results

The experiments revealed essential insights into the practical use of interpretability methods. Integrated Gradients emerged as a reliable, broadly applicable baseline method, balancing robustness, precision, and interpretability. SHAP excelled in precision but exhibited instability under perturbations, particularly problematic in high-stakes scenarios. Ensemble approaches significantly boosted robustness, providing practical strategies for reducing explanation volatility.

5. Conclusions

This work provides a systematic, experimental assessment of interpretability methods in computer vision. It clarifies dependencies between model quality and explanation reliability, quantifies robustness issues, and demonstrates practical approaches like ensembles and mechanistic interpretability for enhancing explanatory quality, helping to open up the "black-box" nature of the models.

Future work should extend interpretability research to underexplored domains such as transformer architectures. To assess explanation quality in terms of human trust and usability, user studies are essential. 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.

Acknowledgements

I would like to thank my supervisor prof. Ing. Lukáš Sekanina. Ph.D. for his mentoring and help during the creation of this work.

References

[1] Christoph Molnar. *Interpretable Machine Learning*. 2 edition, 2022.

[2] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, 2019.

[3] Zachary C. Lipton. The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, June 2018.

[4] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions, 2017.

[5] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, 2017.

[6] Anna Hedström, Leander Weber, Daniel Krakowczyk, Dilyara Bareeva, Franz Motzkus, Wojciech Samek, Sebastian Lapuschkin, and Marina Marina M.-C. Höhne. Quantus: An explainable ai toolkit for responsible evaluation of neural network explanations and beyond. *Journal of Machine Learning Research*, 24(34):1–11, 2023.

[7] Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 2020. https://distill.pub/2020/circuits/zoom-in.