# Peering Inside the Black Box: Comparative Analysis of Explainable AI Methods

Sathammai Sathappan
*Independent Research Project*
sathammai2005@gmail.com

***Abstract*—This study investigates interpretability differences between convolutional and transformer-based architectures. Explanation methods including Grad-CAM, Grad-CAM++, Integrated Gradients, and Attention Rollout, were applied to datasets of varying complexity: CIFAR-10 and RSNA Pneumonia X-ray dataset. Both qualitative and quantitative assessments were conducted to asses the inter-dependence of model architecture, explanation faithfulness and human interpretability, thereby providing actionable insights for applying explainable AI in sensitive domains.**

## I. Introduction

With increasing model complexity, deep neural networks have become increasingly opaque. This makes explanaibility methods essential to understand model decisions and enhance trustworthiness. However, the behavior of the explanation methods varies significantly across architectures and datasets, leading to inconsistent and misleading interpretations. This study systematically compares popular explanations techniques to determine to evaluate their stability and faithfulness across varying contexts.

## II. Methodology

### A. Datasets

- CIFAR-10: A colored, low-resolution image classification dataset containing ten object classes.
- RSNA Pneumonia X-ray: A grayscale, medical-imaging dataset labeled for pneumonia detection.

### B. Model Architectures

- CNNs: ResNet-18 for CIFAR-10, DenseNet-121 for RSNA.
- ViTs: ViT-B for CIFAR-10, ViT-Tiny for RSNA.

### C. Explainability Methods

- Explainability methods include Grad-CAM, Grad-CAM++, Integrated Gradients (IG), and Attention Rollout.
- Evaluation metrics include Insertion/Deletion scores and heatmap Sparsity/Entropy scores.
- For RSNA, localization fine-tuning was applied to guide the model's focus towards the lung region.

## III. Results

### A. CIFAR-10

Integrated Gradients produced consistent results for both ResNet-18 and ViT-B. Grad-CAM variants had similar performance, with Grad-CAM slightly outperforming Grad-CAM++. Attention Rollout produced counterintuitive metrics indicating that attention weights alone were not sufficient to produce faithful explanations. For both the architectures, XAI methods converged for correctly classified samples and diverged for misclassified ones.

### B. RSNA X-rays

- DenseNet-121: Localization fine-tuning was applied to improve spatial focus. Grad-CAM++ was the most faithful method here due to its finer pixel weighing mechanism. Integrated Gradients, despite its strong performance on CIFAR-10, underperformed here due to its linear path integration design. Overall, Grad-CAM variants provided the best coverage of clinically relevant pixels.
- ViT-Tiny: Before localization, IG heatmaps indicated the model's awareness of the lung region, but the heatmaps produced by Grad-CAM variants revealed that the model relied on spurious features. After localization, because the model was guided to focus on the lung region, Attention and IG faithfulness improved while Grad-CAM variants' faithfulness decreased.

## IV. Discussion

1) *Method-Level Behaviour and Faithfulness:* Grad-CAM excels in coarse, single-object datasets (like CIFAR-10), while Grad-CAM++ handles fine-grained domains (like RSNA) better. Integrated Gradients had a consistent performance across all architectures, affirming its model-agnostic nature. Attention Rollout, though visually informative, lacks the fidelity required to be a standalone explanation.
2) *Architecture-Level Patterns (CNNs vs. ViTs):* Localization influences architectures differently. CNNs have local receptive fields. In complex dataset, this causes them to initially attend to semantically irrelevant but high-activation regions. Localization refines this by inducing spatial awareness. ViTs, inherently global, benefit less from localization. Their improvement lies in learning more meaningful inter-patch relations.

3) *Dataset-Level Effects (Simplicity vs. Spatial Complexity):* In simple datasets like CIFAR-10, divergence of XAI methods often correlates with misclassification. In complex datasets like RSNA, such divergence can be a reflection of the nuanced and non-trivial reasoning of the model instead of confusion.
4) *Faithfulness vs. Plausibility:* Visually convincing explanation are not always faithful to the model's reasoning, and vice versa. Faithfulness (alignment with model's reasoning) and Plausibility (alignment with human intuition) needs to be integrated to reliably evaluate interpretability.
5) *Localization as an Attention Enhancer:* Before localization, Attention Rollout produced highly counterintuitive metrics. After localization, attention became spatially concentrated, demonstrating how guided fine-tuning can transform raw attention into quantitatively meaningful explanations.

## V. Conclusion

Interpretability is inherently context-dependent, influenced by the interplay between model architecture, dataset complexity and the chosen explanation method. Comprehensive assessment of explainability requires balancing quantitative metrics with qualitative understanding of model behaviour and dataset structure. This work therefore highlights the need to evaluate faithfulness and plausibility jointly for the advancement of reliable and human-aligned interpretability in AI.