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

Sathammai Sathappan
Independent Research Project
sathammai2005@gmail.com

***Abstract*—This study investigates the interpretability differences in different architectures, including Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs), using multiple explainability methods, mainly Grad-CAM, Grad-CAM++ and Integrated Gradients (IG). Evaluation was conducted using ResNet-18 and ViT-B on CIFAR-10, and DenseNet-121 and ViT-Tiny on RSNA X-ray dataset. Apart from the key quantitative metrics, such as Insertion/Deletion scores, heatmap Sparsity/Entropy scores, qualitative human centered assessments are also reported. Results show that while these methods produced meaningful and intuitive explanations for simple datasets like CIFAR-10, they did not produce coherent explanations for complex and sensitive datasets like medical X-rays. This suggests that the effectiveness of the explainability method is tied to both dataset and the underlying model, and that models can make seemingly correct predictions for spurious or incorrect reasons. Overall, this report highlights the trade-offs between model architecture, explanation faithfulness, and human interpretability, providing the necessary insights for applying XAI in sensitive domains.**

I. Introduction

The domain of artificial intelligence has grown to account for increasingly complex real world problems, and therefore, model complexity has increased too. Simple models like linear regression, while they offer highly intuitive explanations (white-box models), it fails to capture the intricate details of many real world problems. Convolutional Neural Networks and Vision Transformers are widely adopted for this reason, because of their ability to model complex data, but their decision-making process often lack transparency (black-box models). This creates a major challenge for human trust and accountability because of the high-stakes nature of the domains in which these methods are deployed. Explainable AI (XAI) addresses this by providing insights into which parts of the input influences the model's predictions. This enables both validation and critical assessment. However, the reliability of these methods are not uniform. It depends strongly on various factors, including the dataset and model architecture. Simple datasets like CIFAR-10 often yield heatmaps and attributes that are visually intuitive, whereas complex datasets, especially those that require intricate explanations, such as medical X-rays, can produce explanations that are often diffuse, misleading or counter-intuitive. Understanding what causes these differences is crucial for drawing boundaries between limitations of the explainability methods and the errors arising from the model itself. This study explores these dynamics by applying multiple XAI methods - including Grad-CAM, Grad-CAM++ and Integrated Gradients, across two architectures- CNN and ViT and across two datasets - CIFAR-10 and RSNA X-rays. This implementation is designed to explore the trade-offs between model complexity, explanation faithfulness and human centered interpretability.

II. Methodology

A. Datasets

This study utilizes two datasets to evaluate interpretability across domains of varying complexity.

- CIFAR-10: It is a widely used image classification benchmark dataset, consisting of 60,000 low resolution (32x32) color images divided into ten classes. It has a total of 50,000 training images and 10,000 testing images. These images were normalized and resized to match the input dimensions required by the model. This dataset enables controlled experimentation to assess the intuitiveness of the XAI methods, thereby providing a baseline for evaluating their limitations.
- RSNA X-ray dataset: It is a real-world medical imaging dataset provided by the Radiological Society of North America (RSNA), comprising high resolution grayscale chest X-ray images, labeled for pneumonia detection. Images were normalized and resized to match the input dimensions required by the model. This dataset allows the assessment of the intuitiveness of the XAI methods in critical and complex domain, highlighting the extent to which these methods rely on the underlying dataset.

B. Model Architectures

Four model architectures were evaluated across the two datasets:

- Convolutional Neural Networks (CNNs): CIFAR-10 was evaluated using ResNet-18 and RSNA was evaluated using DenseNet-121. Both networks were

initialized with pretrained weights and fine tuned on the respective datasets using standard optimization methods.

- Vision Transformers (ViTs): ViT-B was used for evaluating CIFAR-10 and ViT-Tiny for evaluating RSNA. These networks were also initialized using pertained weights and fine tuned on the respective datasets using standard optimization methods.

These models were selected to compare conventional CNNs with transformer based architectures, enabling analysis of interpretability across models with differing architectures and varying complexities.

C. Explainability Methods

Four popular XAI methods were applied to generate input attributions:

- Grad-CAM: Compares gradients of the target class with gradients of the feature maps in the last convolutional layer. This was used to produce class-specific heatmaps.
- Grad-CAM++: An extension of Grad-CAM that provides more precise heatmaps and can handle multiple occurrences of the same class within an image.
- Integrated Gradients (IG): It attributes feature importance of each pixel by computing the integral of the gradients along the path from a baseline input to the actual input. It is a model agnostic method.
- Attention Rollout: It visualizes how information flows through the network by computing the cumulative attention across all transformer layers. This method was only applied to Vision Transformers (ViTs), as it allows us to trace their attention patterns and understand which regions consistently contributed to the model's predictions.

D. Quantitative Evaluation Metrics

Interpretability was assessed using several quantitative metrics:

- Insertion / Deletion Scores: It measures the influence of the highlighted pixels on the predictions by progressively inserting or deleting them. Higher Insertion AUC and lower Deletion AUC indicates better faithfulness of the explanations.
- Heatmap Sparsity / Entropy: These metrics provide insights into the focus versus noise of the explanations by measuring how concentrated or diffuse the attributions are across the input. Higher Sparsity and lower Entropy indicate compact explanations.

All metrics were computed for each XAI method and across the four model-dataset pairs, allowing systematic comparisons. Furthermore, Localization fine-tuning was used on the RSNA dataset to guide the model's attention towards the lung regions. This was done to provide a reference point for evaluating whether the XAI methods highlighted regions that aligned with clinically relevant pixels or not.

E. Analysis Plan

- Compare interpretability across model architectures (CNN vs ViT) for each dataset.
- Compare performance of XAI methods for each model and dataset.
- Investigate the distinction between method-level failures (XAI method produces unintuitive heatmaps) and model-driven failures (heatmaps reflect model errors).
- Insights from both quantitative metrics and qualitative analysis are used to discuss trade-offs between model complexity, explanation faithfulness and human interpretability.

III. Results

A. CIFAR-10

- ResNet-18

Quantitative evaluation was performed using four metrics- Insertion AUC, Deletion AUC, Sparsity and Entropy. This helps assess the faithfulness and intuitiveness of the explanations produced by the three XAI methods - Grad-CAM, Grad-CAM++, and Integrated Gradients. The results are summarized in Table I.

XAI Method	Insertion AUC	Deletion AUC	Entropy	Sparsity
Grad-CAM	0.2192	0.5752	0.5279	10.3225
Grad-CAM++	0.2084	0.5511	0.4919	10.4297
Integrated Gradients	14.9451	6.2685	0.9973	11.1803

TABLE I: Evaluation metrics computed per class (5 images each) and averaged across classes

1) Results:

- Grad-CAM++ showed a slightly higher faithfulness compared to Grad-CAM, reflected by its lower Deletion AUC (0.5511 vs, 0.5752) and slightly lower Insertion AUC (0.2084 vs. 0.2192). It also produced more concentrated heatmaps, indicated by its lower Entropy, and had a broader focus region, reflected by its lower Sparsity.
- Integrated Gradients, with its higher Insertion AUC and lower Deletion AUC, provided highly intuitive explanations compared to the two Grad-CAM variants. The higher Sparsity (0.99) and moderate to high Entropy (11.18) indicate that the heatmaps had concentrated attributions while relevant background context was preserved.

- Overall, for this dataset, Grad-CAM++ offered the best balance between interpretability and localization sharpness, whereas Integrated Gradients provided precise yet informative attributions, highlighting the varying strength of different XAI method.

2) Observation 1-Method-Level Reliability vs. Model Performance:

- For a model with slightly lower accuracy (80% vs. 89%), the Grad-CAM variants produced highly intuitive Insertion and Deletion metrics. As accuracy increased, these metrics started behaving counterintuitively, which might initially suggest a model-level issue. However, this is a method level limitation, because while Grad-CAM variants failed to produce faithful explanations for a higher accuracy model, Integrated Gradients still generated highly intuitive saliency maps (see Table II). This is evident from its increasing Insertion curve, decreasing Deletion curve, higher Insertion AUC and lower Deletion AUC. Therefore, this demonstrates that a particular XAI method can remain reliable even when others falter.

3) Conceptual Reflection on Evaluation Metrics:

- When analyzing Grad-CAM and Grad-CAM++ performance on CIFAR-10, a conflict emerged between 'faithfulness' and 'plausibility' of explanations. For a higher accuracy model (89%), Grad-CAM's Insertion and Deletion metrics appeared counterintuitive, even though the heatmaps highlighted the correct object regions aligned with human reasoning. This might again suggest a model-level failure, but it continues to remain a method-level limitation because while it produced plausible heatmaps, it did not produce faithful explanations.
- Conversely, in the RSNA dataset, heatmaps often highlighted spurious regions. At first glance, this may appear to be a method-level failure, but it reflects a model-level issue, where the model itself relied on spurious features. As a result, XAI methods produced explanations that were faithful to the model but implausible to humans.
- This reflection highlights the importance of distinguishing between faithfulness (how accurately does an explanation reflect the model's reasoning?) and plausibility (how sensible is the explanation to humans?). Therefore, this emphasizes the need for multiple evaluation perspectives, because human-centered evaluation complements quantitative metrics by highlighting instances where the explanations are visually meaningful even though the metrics appear counterintuitive.

XAI Method	Insertion AUC	Deletion AUC	Entropy	Sparsity
Grad-CAM	17.3988	6.6362	0.9537	7.3516
Grad-CAM++	16.5347	7.6386	0.9973	7.3643

TABLE III: Evaluation Metrics for Each Class (5 Images per Class) and Class-Averaged Results for the 80% Accuracy Model

4) Observation 2-Divergence of XAI Methods in Misclassifications:

- For samples that were correctly classified, both Grad-CAM and Grad-CAM++ consistently highlighted the same object regions, indicating agreement in the model's internal focus. However, for misclassified samples, the two methods highlighted completely different regions, indicating significance divergence. Grad-CAM++ often highlighted the true object regions, whereas Grad-CAM often focused on irrelevant background regions like the sky or the ground (see Table IV for visual comparisons).
- This suggests that the model's misclassification could be a result of varying and inconsistent feature attributions within its deepest layers. When both the strongest and the consistent contributions were from the same region (the object region), the model tends to predict correctly. Conversely, divergence of feature attributions in terms of strength and consistency reflect uncertainty and potential confusion is the model's decision making process.

5) Quantitative Explanation:

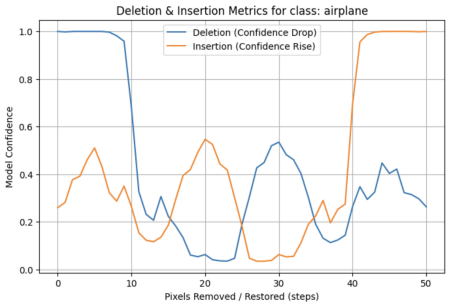
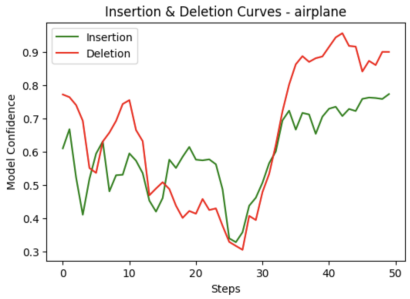
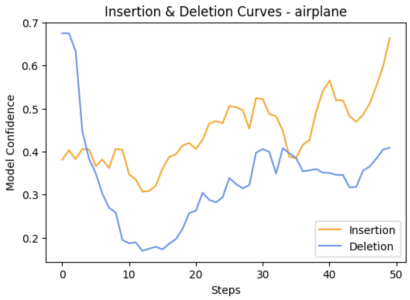
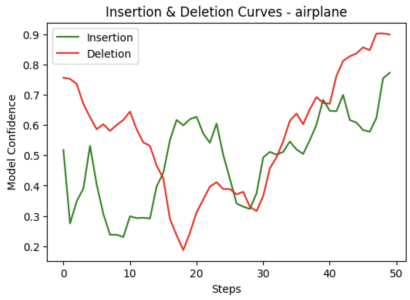
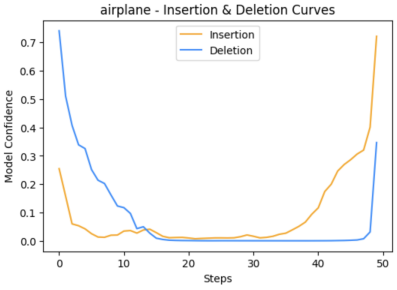
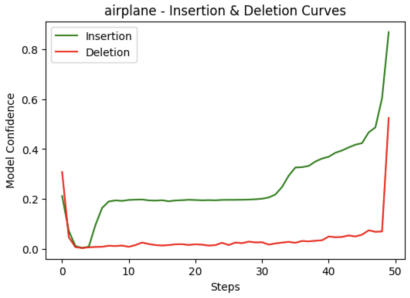
- MAD (Mean Absolute Difference): Measure of the average absolute pixel wise difference between two heat maps. Higher value indicate greater divergence.
- SSIM (Structural Similarity Index): Evaluates similarity in structure and patterns between two heatmaps. Lower values indicate less alignment.

Metric	Average	Min	Max
MAD	0.1221	-	0.4866
SSIM	0.6827	0.1318	-

TABLE V: MAD and SSIM Metrics Computed for Misclassified Images

- The divergence between Grad-CAM and Grad-CAM++ was quantitatively assessed using SSIM and MAD scores. For misclassified samples, the average SSIM was 0.6827 (min: 0.1318) and the average MAD was 0.1221 (max: 0.4866), indicating notable divergence between the two methods. In contrast, for samples that were correctly classified, there was a higher similarity, indicated by higher SSIM and lower MAD. This confirms that the convergence of XAI methods corresponds to correct

TABLE II: Visualization of Insertion/Deletion Graphs for Two Models (80% and 89% Accuracy)

	Model (80% Accuracy)	**Model (89% Accuracy)**
Grad-CAM		
Grad-CAM++		
Integrated Gradients		

model predictions. These metrics reinforce the qualitative analysis that inconsistent feature attribution results in misclassifications (see Table V).

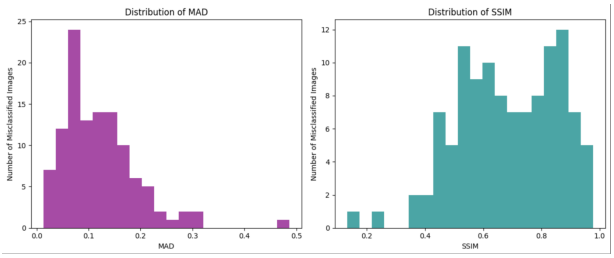


Fig. 1: Distribution of MAD and SSIM across the misclassified images (see Table V for reference values)

- ViT-B

Quantitative evaluation was performed using four metrics - Insertion AUC, Deletion AUC, Sparsity and Entropy. This helped assess the faithfulness and intuitiveness of the explanations produced by the three XAI methods - Grad-CAM, Grad-CAM++ and Integrated Gradients. Further, these metrics were used to quantify the performance of Attention Rollout too. The results are summarized in Table VI.

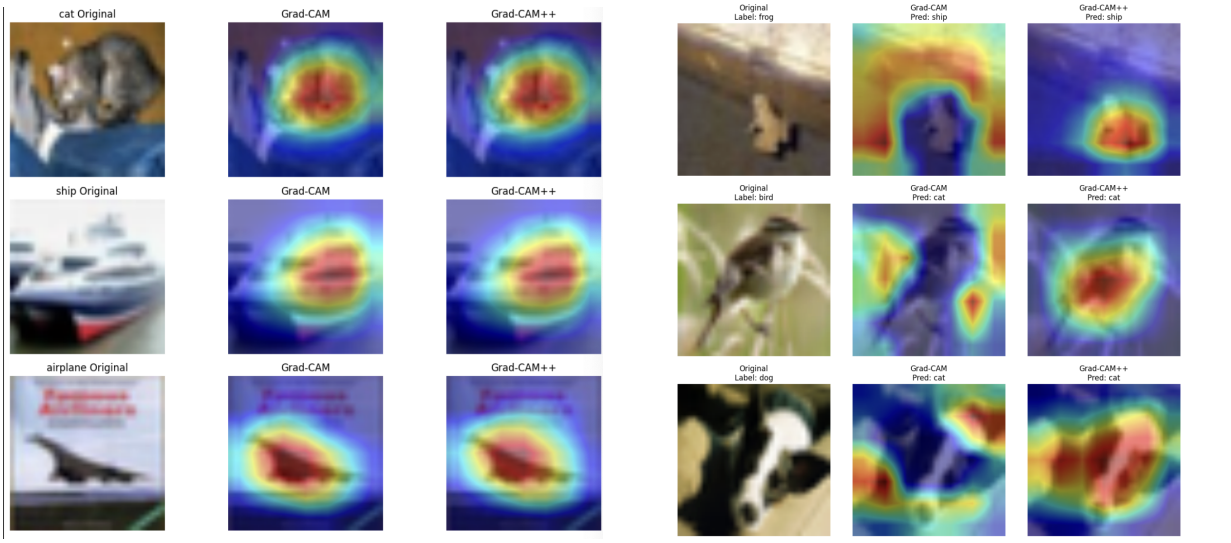
XAI Method	Insertion AUC	Deletion AUC	Entropy	Sparsity
Grad-CAM	0.4594	0.4094	0.5731	7.332
Grad-CAM++	0.3455	0.4542	0.4393	9.423
Integrated Gradients	0.2816	0.1124	0.8559	10.3471
Attention Rollout	0.0968	0.8666	0.0371	5.2085

TABLE VI: Evaluation metrics computed per class (5 images each) and averaged across classes for ViT-B.

1) Results:

- Grad-CAM exhibits slightly higher Insertion AUC (0.4594 vs. 0.3455) and lower Deletion AUC (0.4094 vs. 0.4542) compared to Grad-CAM++, indicating that Grad-CAM produces a more faithful explanation in this case. Although Grad-CAM++ is generally regarded a more faithful method due to

TABLE IV: Visualization of Accurately Classified and Misclassified Images



its precise explanations, its higher Entropy (9.423 vs. 7.332) and lower Sparsity (0.4393 vs. 0.5731) suggest that it highlighted a larger set of contributing pixels, including many less relevant ones. This finer pixel weighting mechanism can dilute its effectiveness on low resolution datasets like CIFAR-10, where a broader object-level focus performs better. The additional contextual information captured by Grad-CAM++ becomes a limitation on low-resolution data.

- Integrated Gradients (IG) produced highly intuitive explanations, indicated by its lower Deletion AUC (0.1124) and comparatively higher Insertion AUC (0.2816). Its high Sparsity (0.8559) combined with relatively high Entropy (10.3471) indicates that it produced concentrated attributions while maintaining some background context. Overall, IG's saliency maps provide precise yet informative attributions, making it the most human aligned method for ViT-B on this dataset.
- Attention Rollout produced highly counterintuitive metrics, with a very low Insertion AUC (0.0968) and high Deletion AUC (0.8666). The Insertion curve remained nearly flat along the bottom axis, while the Deletion curve stayed flat near the top – a trend observed across all ten classes. This behavior reflects how this methods failed to highlight pixels that actually influenced the model's decision, representing a clear method-level failure. Its extremely low Sparsity (0.0371) and low Entropy (5.2085) further indicate that very few pixels were assigned significant importance, and the method largely ignored relevant object regions.

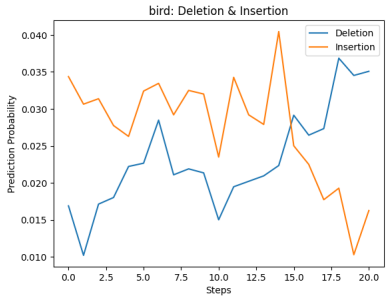
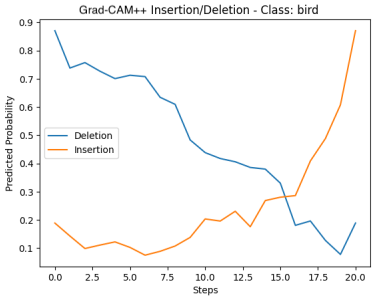
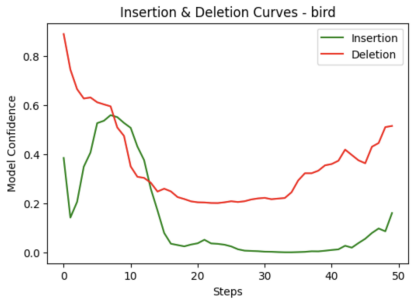
2) Observation 3 – Variability in Insertion and Deletion Curves Across Architectures:

- The Insertion and Deletion curves produced by the ViT model structurally differed from those produced by the CNN model, particularly for Grad-CAM – a trend observed consistently across all ten classes. This variability was less noticeable for Integrated Gradients and largely absent for Grad-CAM++ because their weighting mechanisms tend to smoothen the attribution responses. The increased spikiness in ViT curves could be a result of the model's non-local attention mechanism. This can cause abrupt shift in pixel importance when different patches gain or lose contextual relevance. Conversely, CNNs have localized spatial inductive biases, which shows smoother transitions in Insertion and Deletion curves. This reflects spatial continuity in feature activations.

3) Observation 4 – Architectural Differences in Explanation Faithfulness:

For all four evaluation metrics - Insertion AUC, Deletion AUC, Entropy and Sparsity, ViT-B consistently produced lower values than the ResNet-18 CNN model. This pattern indicates that ViTs generate more distributed, context dependent attributions, where importance is spread across multiple non-contiguous regions than just local concentration. As a result of this, perturbation-based metrics may underestimate the faithfulness of ViT explanations because it assumes compact attributions.

TABLE VII: Comparison of XAI explanation maps for ViT and CNN models. Each row shows results from a different method (Grad-CAM, Grad-CAM++, Integrated Gradients), illustrating how the attribution patterns differ across architectures.

	ViT	**CNN**
Grad-CAM		
Integrated Gradients		
Grad-CAM++		

This highlights a key architectural difference: CNNs leverage local spatial relevance, whereas ViTs leverage relational, global context, resulting in inherently subtler explanation maps that don't fit well under standard XAI evaluation schemes.

B. RSNA Dataset

Quantitative evaluation was performed using four metrics - Insertion AUC, Deletion AUC, Sparsity and Entropy. This helped assess the faithfulness and intuitiveness of the explanations produced by the three XAI methods - Grad-CAM, Grad-CAM++, and Integrated Gradients. The results are summarized in Table VIII.

XAI Method	Insertion AUC	Deletion AUC	Entropy	Sparsity
Grad-CAM	3.7544	2.1934	0.7814	10.2290
Grad-CAM++	0.9814	2.1321	0.8814	10.4414
Integrated Gradients	0.5874	0.5486	0.0029	10.3757

TABLE VIII: Quantitative evaluation metrics for DenseNet-121 on 30 RSNA images, averaged across the samples.

1) Results:

- Grad-CAM++ exhibits higher Entropy (10.4414 vs. 10.2290) and higher Sparsity (0.8814 vs. 0.7814) compared to Grad-CAM, indicating a more diffuse heatmap, which might indicate instability at first. However, further evaluation shows that its higher Insertion AUC (3.9814 vs. 3.7544) and lower Deletion AUC (2.1321 vs. 2.1934) reflect a more faithful explanation rather than instability. Its finer pixel weighing mechanism distributes attribution

across all relevant regions, highlighting the model's decision making process in a more comprehensive manner.

- Integrated Gradients (IG), in contrast, produced highly focused attributions, indicated by its extremely low Sparsity (0.0029). Its relatively high entropy (10.3757) suggest that, even though only a few pixels contributed, they exhibit varying attribution strength, which provides precise yet informative explanations.
- Overall, Grad-CAM variants covered a broader range of contributing pixels, whereas IG concentrated on fewer regions with variable strengths. The differences aligns with the respective design principles of the methods. They further demonstrate how Insertion and Deletion metrics capture subtle distinctions in attribution behavior. The results indicate that the method performance is context-dependent, with diffusion and sparsity influencing metrics in ways that reflect the model's true reasoning.

2) Observation 5 – Dataset Structure Drives XAI Agreement and Divergence:

- In the CIFAR-10 dataset, when different XAI methods highlighted distinct regions, these instances consistently coincided with misclassifications. Conversely, in the RSNA dataset, while the XAI methods substantially diverged for nearly every image, the model still produced correct predictions. This is seen in the coherent and intuitive quantitative metrics across the methods.
- This difference is largely attributed to the dataset characteristics: CIFAR-10 images are small, low-resolution and typically single object oriented, with minimal and irrelevant background. Consequently, XAI methods naturally converged on similar salient regions. RSNA images, in contrast, are high resolution, and contain multiple sub-regions capable of providing independent discriminative cues and exhibit greater spatial complexity. Therefore, even when the XAI methods emphasize different regions, their explanations remain valid, highlighting the dataset-dependent nature of explanation divergence.
- Table IX provides a visual comparison of the heatmaps produced by Grad-CAM, Grad-CAM++ and Integrated Gradients on one sample – an illustration of how dataset structure influences method agreement and divergence.

3) Observation 6 – Localization Supervision Enhances Activation-Based Interpretability:

- Before localization supervision, Insertion and Deletion curves for Grad-CAM, Grad-CAM++ and Integrated Gradients were largely similar, which indicates weakly faithful attributions. After localization using bounding boxes, Grad-CAM and Grad-CAM++ showed remarkable improvement: Insertion curves rose sharply, Deletion curves dropped steeply, indicating a more precise and reliable explanation (see Table X). Integrated Gradients improved less noticeably, as its pixel-level operation does not fully exploit the spatial alignment provided by localization. This demonstrated that the effectiveness of activation-based XAI methods depends on spatial learning and dataset complexity. This explains why Grad-CAM variants can outperform IG on high resolution medical-images like RSNA, whereas IG outperforms on small and low-resolution datasets like CIFAR-10.

4) Observation 7 – Coarse vs. Fine Pixel Attributions Depend on Dataset Complexity:

- A clear pattern emerged when comparing Grad-CAM and Grad-CAM++ across datasets of different complexity. In CIFAR-10, Grad-CAM constantly outperforms Grad-CAM++. This suggests that for small, low-resolution images dominated by a single object, major object pixels are sufficient for explanations, and finer-grained attributions add little value. This is confirmed by the observation that an 80% accurate CIFAR-10 model produced more intuitive Grad-CAM metrics than a higher accuracy model (89%) model, implying that only major object pixels are necessary for faithful explanations.
- In contrast, RSNA images are high resolution and spatially diffuse, where multiple subregions contribute to classification. Here, Grad-CAM++ performs better due to finer pixel sensitivity, producing lower Deletion AUC and higher Insertion AUC. Thus, the metrics of a higher accuracy model (93%) outperforms the lower accuracy one (81%), highlighting the importance of subtle yet informative regions.
- This observation ties together the prior findings: coarse attributions are sufficient to produce meaningful explanations in simple datasets, whereas faithful explanations in complex datasets like medical images requires attention to both major and finer-grained spatial features.

- ViT-Tiny

Quantitative evaluation was performed using four metrics- Insertion AUC, Deletion AUC, Sparsity and Entropy. This helps assess the faithfulness and intuitiveness of the explanations produced by the three XAI methods - Grad-CAM, Grad-CAM++ and Integrated Gradients. Further, these metrics were used to quantify the performance of Attention Rollout too. The results are summarized in Table XII.

TABLE IX: Comparison of heatmaps for an RSNA X-ray image across XAI methods (Grad-CAM, Grad-CAM++, Integrated Gradients), illustrating method-specific emphasis and divergence.

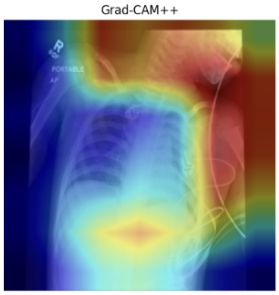
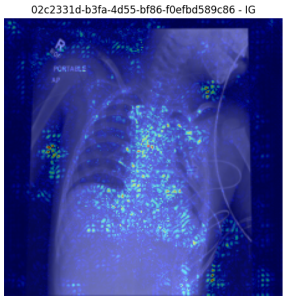
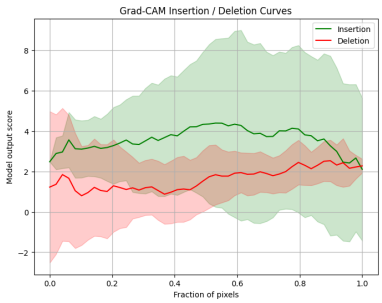
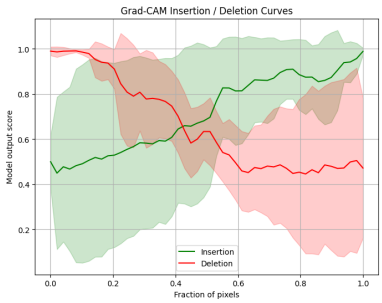
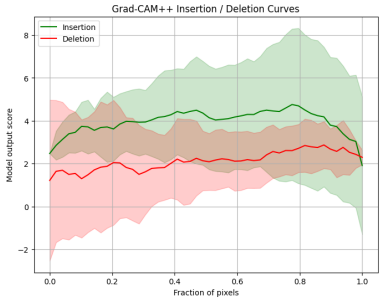
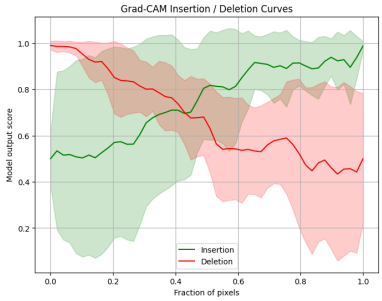
Grad-CAM	**Grad-CAM++**	**Integrated Gradients**
		

TABLE X: Comparison of Explanations Before and After Localization Training

	Before Localization Training	**After Localization Training**
Grad-CAM		
Grad-CAM++		

XAI Method	Insertion AUC	Deletion AUC	Entropy	Sparsity
Grad-CAM	0.6573	0.6417	0.3440	9.8505
Grad-CAM++	0.6576	0.6588	9.0063	0.6825
Integrated Gradients	0.6700	0.6809	0.6108	10.4069
Attention Rollout	9.0000	9.0000	0.9500	10.7804

TABLE XI: Evaluation metrics computed for 30 images and averaged across the dataset for ViT-Tiny.

1) Results:

- Grad-CAM, Grad-CAM++ and Integrated Gradients produced comparable quantitative metrics. Grad-CAM produced the most faithful explanations, indicated by its higher Insertion AUC (0.6573) and lower Deletion AUC (0.6417). The Insertion and Deletion curves also followed an intuitive trend. Grad-CAM++ exhibited similar behavior, with a slightly higher Insertion (0.6576) and Deletion (0.6588) AUC values, yielding largely intuitive curves.
- Integrated Gradients also produced similar Insertion (0.6700) and Deletion (0.6809) AUC values. However, its curves were less intuitive, indicating that the highlighted pixels sometimes confused the model instead of reinforcing the model's predictions.
- Attention Rollout, due its global attention mechanism, produced high Entropy (10.7804) and Sparsity (0.9500) values. Its identical Insertion and Deletion AUC values (9.0000) suggest a lack of concentrated focus on the regions that influence the

model's decision. This aligns with the observations of ViT-B on CIFAR-10, where Attention Rollout produced less faithful explanation too due to the attention being distributed rather than being concentrated.

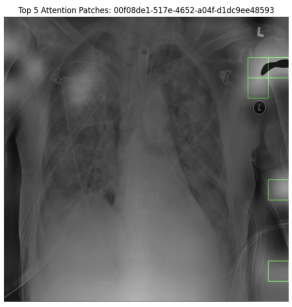
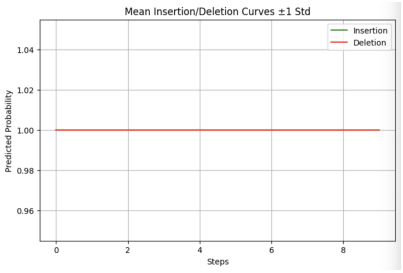
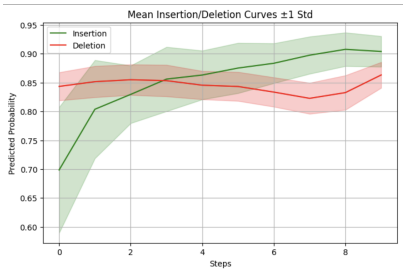
5) Observation 8 – Localization Enhances Attention Focus and Faithfulness:

- Before and after localization, the top regions highlighted by Attention Rollout continued to be the same, but, before localization, the Insertion and Deletion curves were largely flat and coinciding (see Table XII). This indicated that the attention map, though visually plausible, wasn't qualitatively faithful. The model's prediction was not strongly dependent on these regions. However, after localization, the attention values became concentrated for the same regions. As a result, the Insertion and Deletion curves also became highly intuitive, clearly reflecting the influence of these regions in the model's predictions. This demonstrated that localization amplified the contribution of already correct regions and that raw attention alone may be insufficient as a faithful metric. Localization, thus makes attention based methods more meaningful quantitatively.

6) Observation 9 – Model Behavior and Faithfulness Before and After Localization:

- Before localization, Integrated Gradients (IG) mostly focused on the lung regions, suggesting that the model recognized their relevance. However, the Insertion and Deletion curves were counterintuitive, indicating that the highlighted pixels did not consistently reflect what the model was actually relying on for predictions. Grad-CAM and Grad-CAM++ often highlighted peripheral or corner regions. While these areas might seem arbitrary, the model did use them in making predictions, which is reflected in the more intuitive curves for Grad-CAM, despite the overall model accuracy being only around 70% (see Table XIV). This suggests that latching onto these specific regions doesn't always help the model make correct predictions.
- Attention Rollout adds another perspective. The top-K regions, before and after localization, remained exactly the same. Before localization, the Insertion and Deletion curves were flat and coinciding, suggesting that the model did not strongly focus on the regions. This was expected as the model was free to attend wherever it wanted, therefore, these top-K patches were not yet influential. This led to its uninformative curves. After localization, the model was guided to focus on the lung regions. But because it struggled to extract meaningful pattern from these pixels, it increased the attention to the top-5 patches, which, in many cases, were the corner regions. This strong latching drove the predictions even after localization, even though these patches weren't clinically relevant regions.
- After localization, the IG curves became more intuitive, because the highlighted regions now aligned with where the model was guided to focus. But even then, the Insertion-Deletion curves did not become fully intuitive. The Insertion curve increased slightly at first but then gradually declined, while the Deletion curve rose slowly. This suggest that while the model did start paying attention to key lung pixels, it struggled to extract meaningful relationships as more pixels were added. This idea was also reinforced by the plateau observed in localization training and validation loss and AUROC.
- Interestingly, the Grad-CAM and Grad-CAM++ heatmaps continued to highlight the same region before and after localization, yet their curves, which were intuitive before localization, became counter-intuitive after localization. This suggests that the model's reliance on these regions changed in subtle ways. Post-localization, the models relied even more on these corner regions, reflected in the higher AUCs (see Table XIV), even though they were not clinically relevant. Conversely, the Insertion and Deletion curves revealed another behavior. Insertion curve decreased and Deletion curve increased, suggesting that these regions, even though they were heavily focused on, it pushed the model away from the correct predictions. Therefore, removing these spurious attributions helped the model to be more confident in the right predictions.
- Essentially, while localization forced the model to focus on the lungs, its difficultly in interpreting these regions and forming meaningful relationships between the pixels caused the model to compensate by relying more on previously attended corner patches. But since these patches weren't clinically relevant, it distorted the faithfulness of Grad-CAM based explanations.
- Overall, this observation suggests that while localization can only guide the model to focus on a specific region, it cannot force the model to intuitively understand this region, especially when the model already understands its relevance before localization but still doesn't rely on it to make its predictions. This also highlights how localization can only help a model when the model isn't spatially aware of the specific region (like for CNNs). It does not have a huge meaningful impact on the explanation methods when the model intentionally decides to not focus

TABLE XII: Top-K regions highlighted by Attention Rollout, with corresponding Insertion and Deletion curves shown before and after localization training.

Top-K regions	Insertion/Deletion Curves Before Localization	Insertion/Deletion Curves After Localization
		

on these attributions (like ViTs).

XAI Method	Insertion AUC	Deletion AUC	Entropy	Sparsity
Grad-CAM	0.9055	0.7803	0.9937	14.4395
Grad-CAM++	0.8958	0.7744	0.7306	9.0065
Integrated Gradients	0.9232	0.7918	0.9933	10.4103
Attention Rollout	7.7185	7.5917	0.9500	10.7611

TABLE XIII: Evaluation Metrics for the XAI Methods After Localization Training

7) Observation 10 – Architectural Differences in Localization Response: CNN vs. ViT:

- A clear difference arises when comparing CNNs and Vision Transformers (ViTs) on the RSNA dataset. For DenseNet-121 (CNN), Grad-CAM heatmaps before localization often highlighted corner regions, producing counterintuitive Insertion and Deletion curves. However, after localization, the heatmaps started highlighting the lung regions, and the curves became more intuitive, indicating that the model now leverages clinically relevant pixels to make predictions. This behavior is consistent with CNN's mechanism - without guidance, they latch onto any local patterns that are highly activated, which are often irrelevant areas. Localization steers them towards meaningful regions.
- In contrast, ViT-Tiny models, because of their global self-attention mechanism, naturally highlight relevant regions even before localization. Nevertheless, as seen in the previous observation, these regions are not fully exploited for predictions until the model learns effective inter-patch relationships. Therefore, ViTs can attend globally from the start, but their predictive power depends on integrating patch-level information meaningfully.
- Overall, this observation emphasizes that CNNs require spatial guidance to steer the attention towards the correct regions, whereas ViTs inherently attend globally, therefore their effectiveness relies on learning relationships between the patches. This architecture difference highlights how model design affects interpretability and localization training.

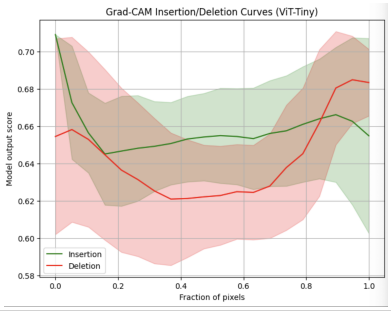
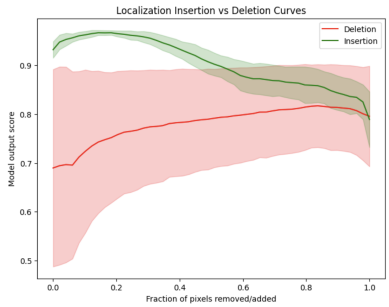
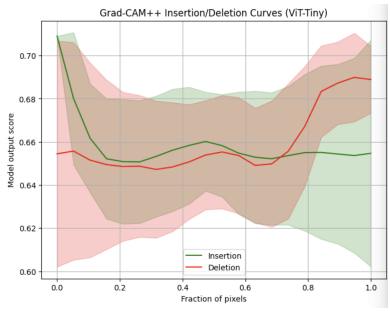
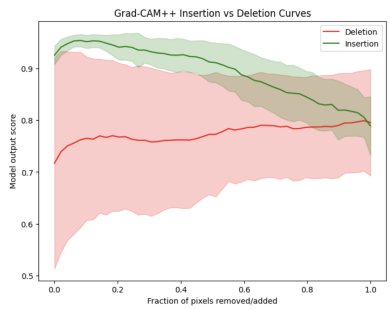
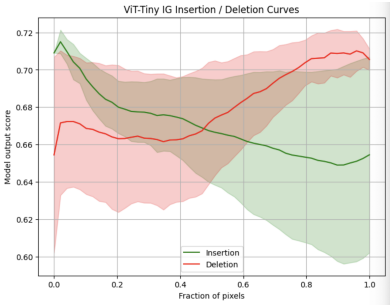
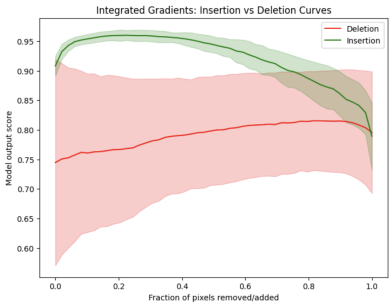
IV. Discussion

This study aimed to investigate how the behavior of the explainable AI (XAI) methods is influenced by model architecture and dataset complexity. By comparing Grad-CAM, Grad-CAM++, Integrated Gradients and Attention Rollout across convolutional and transformer based models on both simple and complex datasets, it became evident that interpretability is highly context-based. Each method responds differently to different model architecture, mechanisms and the characteristics of the dataset. This demonstrated that the quality of the explanations depends not only on the XAI method chosen, but also on the underlying model.

1) Method-Level Behavior and Faithfulness:

- Grad-CAM++ typically produced smoother, more spatially distributed heatmaps, which proved useful for high-resolution medical images such as RSNA, where subtle features carry important predictive weight. On smaller, low-resolution datasets like CIFAR-10, however, this broader sensitivity sometimes diluted the relevance of highlighted regions, whereas Grad-CAM, with its coarser, object-focused approach, often better mirrored the model's actual decision-making.
- Integrated Gradients had a consistent behavior across datasets and architectures. Its heatmaps captured both local and global influence without excessive variations. Therefore, its explanations remained quite stable. Even when activation-based methods demonstrated counterintuitive behavior, IG maintained coherent and faithful attributions. This makes IG the methodologically robust amongst the four methods used.

TABLE XIV: Visualization of Insertion/Deletion Curves Before and After Localization

	Before Localization	After Localization
Grad-CAM		
Grad-CAM++		
Integrated Gradients		

- Attention Rollout provided another perspective. While its attention maps often remained intuitive, the Insertion and Deletion metrics, with its near flat curve, remained largely uninformative. This indicates that although the attention maps suggested the correct focus area, because of ViTs inherent nature of globally distributed attributions, they did not reliably affect the model's predictions. This highlights a key point: attention-based visualizations alone cannot be used to provide faithful explanations.

2) Model-Level Patterns: CNNs vs. Vision Transformers:
   - Architectural differences between CNNs and ViTs influence the effect of localization. CNNs, because of their locally receptive field, initially latched onto spurious peripheral regions to make predictions. After applying localization, the model shifted its attention towards the relevant regions. This improved the faithfulness of the corresponding metrics, thus demonstrating that spatial guidance can align the CNNs internal attribution with the meaningful regions.
   - ViTs, in contrast, exhibited global attention from the start. However, this visual alignment is not leveraged fully until the model learns inter-patch relationships. Thus, CNNs require external guidance to focus on relevant regions, whereas ViTs inherently have global awareness, but depend on relational reasoning for utilizing this awareness for predictions.
3) Dataset-Level Effects: Simplicity vs. Spatial Complexity:
   - Dataset exerted a significant influence on the behavior of the XAI methods. In CIFAR-10, which is single subject oriented and had minimal background complexity, the XAI methods converged on salient regions. Any divergence often

correlated with incorrect predictions. Conversely, RSNA images has multiple sub-regions that can independently influence the model's predictions. Therefore, even when the predictions were correct, different XAI methods highlighted different regions. This doesn't indicate failure, rather, it reflects the dataset's intrinsic complexity, where multiple subregions can provide valid reasoning cues.

- Hence, alignment of the XAI methods as an indicator for reliability is mostly valid only on simpler datasets. In complex, high-resolution domains, diversity in explanations doesn't always suggest inconsistency, rather, it may reflect the nuanced model reasoning.

4) Faithfulness vs. Plausibility:
   - A recurring theme was the distinction between faithfulness and plausibility. Explanations that appeared visually convincing, did not necessarily highlight the regions the model highly relied on. Conversely, counterintuitive visualizations demonstrated high fidelity to the model's decision making process. For example, the higher-accuracy model on CIFAR-10 often produced heatmaps that looked plausible, but the poor metrics indicated that it wasn't faithful to the model's internal reasoning. Conversely, the Grad-CAM variants on RSNA, often produced heatmaps that highlighted spurious regions, but the excellent metrics suggest that even though they look implausible, they're still faithful. This emphasizes that faithfulness (how well an explanation reflects the model's reasoning) and plausibility (how convincing it appears to humans) are distinct terms. This also necessitates the integration of both quantitative evaluation and human judgement in assessing interpretability.
5) Broader Implications:
   - These findings demonstrates that interpretability is highly context-dependent, and is largely influenced by architecture, supervision and dataset characteristics. Overall, explainability should be considered as a process rather than a fixed property, Quantitative metrics alone are insufficient to evaluate the faithfulness of the explanations. The effect of model design, dataset complexity and evaluation methodology must be considered to understand the model behavior comprehensively.

V. Conclusion

This study examined how model architecture, dataset complexity, and XAI method choice – their subtle interplay – affects explanation reliability and interpretability. Key findings show that CNNs benefit more from localization whereas ViTs benefit more when they understand inter-patch relationships. In terms of XAI method choices - Grad-CAM performs well on simple datasets, compared to Grad-CAM++, which produces high fidelity explanations in complex, high resolution dataset. Integrated Gradients has a consistent behavior across datasets and architectures, making it a robust method. Importantly, faithfulness and plausibility are distinct. Visually convincing explanations do not always reflect the model's true reasoning. Overall, interpretability is context-dependent. Therefore, the integration of quantitative metrics, knowledge of model behavior, and dataset characteristics is essential for a comprehensive assessment.

VI. Future Work

The findings suggest several directions for further research. First, the differences between convolution and transformer-based models indicate a need for architecture aware XAI methods. Current XAI methods, including Grad-CAM variants and Integrated Gradients were developed with CNNs in mind. XAI frameworks designed specifically for ViTs, that leverages global attention and patch-level interaction mechanisms can improve both interpretability and faithfulness.

Second, model reasoning and interpretability are closely linked. Counterintuitive explanations were often a result of the model's reliance on spurious features rather than flaws in the XAI methods itself. This suggests that improving interpretability may require improving the model behavior itself. Future work could explore how architectural choices, regularization strategies, or localization-guided supervision can make models inherently more transparent. This reduces the dependence on post-hoc explanations.

Finally, incorporating causality-driven evaluation: assessing whether the highlighted features overlaps with the relevant region, can help bridge the gap between faithfulness and human trust. This approach would steer the field towards proactive model design, where interpretability is embedded from the beginning, rather that solely relying on post hoc explanations.

References

[1] A. Krizhevsky, G. Hinton, and V. Nair, "CIFAR-10 and CIFAR-100 datasets," 2009. [Online]. Available: https://www.cs.toronto.edu/~kriz/cifar.html

[2] RSNA Pneumonia Detection Challenge Dataset, "Radiological Society of North America," 2018. [Online]. Available: https://www.kaggle.com/c/rsna-pneumonia-detection-challenge

[3] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, et al., "PyTorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, vol. 32, 2019. [Online]. Available: https://pytorch.org

[4] Torchvision, "Models and datasets for computer vision," 2020. [Online]. Available: https://pytorch.org/vision/stable/models.html

[5] R. Wightman, “TIMM: PyTorch image models,” 2021. [Online]. Available: https://github.com/rwightman/pytorch-image-models
[6] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual explanations from deep networks via gradient-based localization,” in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2017. [Online]. Available: https://github.com/jacobgil/pytorch-grad-cam
[7] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, “Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks,” in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018.
[8] M. Sundararajan, A. Taly, and Q. Yan, “Axiomatic attribution for deep networks,” in *Proc. 34th Int. Conf. Machine Learning (ICML)*, 2017. [Online]. Available: https://captum.ai
[9] S. Abnar and W. Zuidema, “Quantifying attention flow in transformers,” in *Proc. 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.
[10] NumPy Developers, “NumPy: Fundamental package for scientific computing with Python,” 2020. [Online]. Available: https://numpy.org
[11] G. Bradski, “The OpenCV library,” *Dr. Dobb’s Journal of Software Tools*, 2000. [Online]. Available: https://opencv.org
[12] J. D. Hunter, “Matplotlib: A 2D graphics environment,” *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007. [Online]. Available: https://matplotlib.org
[13] A. Clark and Contributors, “Pillow (PIL Fork) Python Imaging Library,” 1990. [Online]. Available: https://python-pillow.org