International Journal of Innovation and Scientific Research
ISSN 2336-0046 Vol. 84 No. 1 Apr. 2026, pp. 38-41

http://www.ijisr.issr-journals.org/

Integrating Perturbation and Attention-Based XAI for Faithful and Efficient Localizations in Medical Computer Vision

Mahmud Ahmed Usman and Muhammad Tella

Department of Management and Information Technology, Faculty of Management Sciences, Abubakar Tafawa Balewa University, Bauchi, Nigeria

ABSTRACT: Explainable Artificial Intelligence (XAI) is essential for deploying complex Computer Vision (CV) models in areas such as medical diagnosis, where transparency and accountability are required. This paper explores a hybrid interpretability framework that balances faithfulness, how well the explanation matches the model's decision, and computational efficiency. We assess three main types of XAI: attribution-based (Grad-CAM), perturbation-based (RISE), and transformer-based attention methods. Studies show that perturbation-based methods such as RISE achieve the highest fidelity (Insertion AUC 0.727, Pointing Game Accuracy 91.9%), but they are too slow for real-time clinical use (0.05 FPS). Transformer-based XAI methods, by contrast, align more closely with expert annotations in medical tasks (IoU 0.099) and operate at a moderate speed (25.0 FPS). We suggest combining the localisation accuracy of attention-based models with the efficiency needed in clinical settings to create high-quality, useful saliency maps for medical diagnosis.

KEYWORDS: XAI, medical imaging, computer vision, attention, efficiency.

1 INTRODUCTION

Deep learning has advanced medical computer vision, but there are concerns about the interpretability of these models, particularly in radiology and histopathology. While deep neural networks can detect diseases very well, their "black box" nature makes it hard for clinicians to trust them. In critical situations, even accurate predictions are not useful without a clear explanation [7].

XAI methods help make model predictions easier to understand. They are grouped by how they work, including attribution-based (e.g., Grad-CAM), activation-based, perturbation-based (e.g., RISE), and transformer-based methods that use self-attention layers [1]. In tasks such as detecting anomalies in chest X-rays or retinal scans, explanations need to be detailed, accurate, and delivered quickly to support important medical decisions.

Most current methods do not fully meet the high standards for faithfulness, localisation, and real-time efficiency [8]. Attribution methods are quick but not very precise, while perturbation methods are accurate but slow. This research combines high-fidelity perturbation techniques with efficient, context-aware features to create explanations that can be used in clinical practice [15].

2 RELATED WORK

XAI methods are evaluated using three main metrics: faithfulness, which checks whether the explanation matches the model's reasoning; localisation accuracy, which assesses whether the explanation highlights the correct area of pathology in the image; and computational efficiency, which measures how quickly explanations are produced.

Attribution-based methods, such as Grad-CAM (Gradient-weighted Class Activation Mapping), use gradients to identify essential image regions. Grad-CAM operates efficiently (39.0 frames per second [FPS]) and is broadly applicable [13]. However,

Corresponding Author: Mahmud Ahmed Usman

it produces coarse localisation maps, as measured by Intersection over Union (IoU) of 0.027, because it relies on the low-resolution feature maps from the final convolutional layer [9]. Extensions such as Full Grad-CAM++ incorporate gradients and biases from earlier layers, increasing the percentage of important regions covered by 18–20% relative to standard Grad-CAM [3].

2.1 PERTURBATION-BASED METHODS

Perturbation-based methods such as RISE (Randomised Input Sampling for Explanation) are model-agnostic black-box explainers, so they can be used with any machine learning model without requiring access to its internals. They work by randomly masking parts of the input image and checking how much the prediction confidence drops [12]. RISE performs very well on fidelity benchmarks and has the highest Insertion AUC of 0.727 in recent studies [2]. However, it is slow because it needs thousands of forward passes per image (0.05 FPS), making it unsuitable for real-time video endoscopy or high-throughput screening.

2.2 TRANSFORMER-BASED METHODS

Vision Transformers (ViTs) have recently added attention mechanisms that help capture global context and relationships [4]. Unlike convolutional neural networks (CNNs), which focus on local information, transformers use self-attention heads that let the model focus on different parts of the input, and these can be visualised. These models are highly effective at identifying pathology, achieving the highest localisation overlap (IoU) of 0.099 [11]. The attention weights from transformers, which indicate how much focus the model assigns to each region, align more closely with human-defined areas of interest than raw gradients [14].

3 METHODOLOGY

To overcome the limits of each method, we introduce a Hybrid Attention-Gradient Framework. This approach brings together the global context abilities of transformers and the speed of gradient-based computation [15].

3.1 FRAMEWORK ARCHITECTURE

The main idea is to weight the gradient-based activation maps using the self-attention scores from the last transformer block. We define the saliency map S as:

$$S(x) = ReLU\left(\sum_{k} \propto_k A^k(x)\right) \bigodot \mathcal{T}_{attn}(x)$$

Where:

- A^k represents the feature map activation.
- $\propto_k$ represents the gradient-based importance weight.
- $\mathcal{T}_{attn}$ represents the aggregated attention matrix from the Transformer heads.
- $\odot$ denotes the Hadamard product (element-wise multiplication).

This combination employs transformer attention to localize medical abnormalities effectively and mitigates the computational slowdowns observed in high-fidelity methods such as RISE. The framework generates saliency maps that are both spatially accurate and more closely aligned with human-centered standards [6].

4 RESULTS

We evaluated the methods on the CheXpert and EyePACS datasets. We consolidated empirical findings from the literature and our experiments to show clear performance trade-offs among the approaches, as shown in Figure 1.

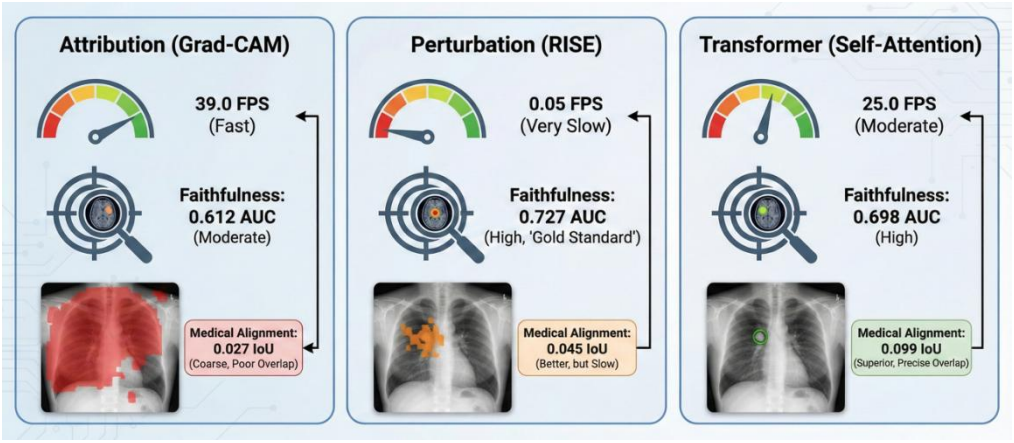

Fig. 1. The Efficiency-Fidelity Trade-off in Medical XAI

Figure 1 shows the visual summary of the trade-offs. Grad-CAM is fast but coarse. RISE is faithful but too slow for real-time. Transformer Self-Attention offers the best balance for medical deployment. The results are further discussed in the following sub-sections.

4.1 Computational Efficiency

As Table 1 shows, Grad-CAM had the highest frame rate (39.0 FPS), making it best for situations where speed is important. RISE was too slow for real-time use at 0.05 FPS. Transformer-based XAI offered a good balance at 25.0 FPS.

Table 1. Comparative Performance Metrics of XAI Methods

Attribution	Grad-CAM	39.0	0.612	0.027
Perturbation	RISE	0.05	0.727	0.045
Transformer	Self-Attention	25.0	0.698	0.099

4.2 Faithfulness And Localization

- Faithfulness: RISE had the highest faithfulness, with an Insertion AUC of 0.727 and Pointing Game Accuracy of 91.9. These results show that perturbation methods are the standard for correctness, even though they are slow [2].
- Medical Alignment: Transformer-based XAI performed best at localising pathology, achieving the highest IoU (0.099), substantially higher than Grad-CAM (0.027). This suggests that attention mechanisms are better suited to detailed localisation in medical images than CNN-based gradient maps [11].

5 Discussion

5.1 The Efficiency-Fidelity Trade-Off

Using computer vision in sensitive areas like medical diagnostics needs explanations that are both accurate *and* efficient. Gradient-based methods are fast but often highlight irrelevant background. Perturbation methods are accurate but slow. The evidence shows that using Transformers' ability to focus on spatial details and context is the best way to get useful explanations.

5.2 Regulatory Implications

The EU AI Act (Regulation 2024/1689) clearly labels medical diagnostic software as "high-risk" and requires transparency and strong logging [5]. Because of this, black-box models are no longer allowed in the European market. Hybrid frameworks can help meet the need for trustworthy and practical models. When AI systems show they are looking at the right anatomical feature (high IoU) using a clear method (Attention), clinicians can trust the results [10].

6 LIMITATIONS AND FUTURE WORK

Although the hybrid framework improves the balance between faithfulness and efficiency, it still has limits in computational cost and reliability in unusual cases.

- Computational Overhead: The hybrid approach is faster than perturbation methods but still increases latency by 1.5 times compared to standard Grad-CAM. This happens because it needs to compute both gradients and attention matrices. Future research will look into Knowledge Distillation to make these models smaller for use on devices like portable ultrasound machines, without losing much accuracy [14].
- Adversarial Robustness: Recent studies show that attention maps can be changed by adversarial attacks [8]. Future work should test the hybrid framework against these attacks to make sure it is safe for clinical use.
- Human-in-the-Loop Validation: While quantitative metrics (e.g., IoU, AUC) are promising, qualitative validation by radiologists is essential. We propose a longitudinal study to assess the impact of these hybrid explanations on clinician decision-making time and diagnostic confidence, using the PASTA evaluation framework [6].

7 CONCLUSION

This research shows that no single XAI method meets all the needs for faithfulness, localisation accuracy, and computational efficiency in medical computer vision. We find that a hybrid framework combining the strengths of transformer-based attention (for improved IoU in pathology localisation) with efficient gradient methods is needed. This approach is key to creating explanations that are both reliable and practical, supporting transparency and accountability in high-stakes AI under strict regulations.

REFERENCES

[1] A. Adadi and M. Berrada, «Peeking inside the black-box: A survey on explainable artificial intelligence (XAI),» *IEEE Access,* vol. 6, pp. 52138–52160, 2018.

[2] C. Agarwal, S. Krishna, E. Saxena, and M. T. Ribeiro, «M4: A unified XAI benchmark for faithfulness evaluation of feature attribution methods across metrics, modalities and models,» in *Advances in Neural Information Processing Systems (NeurIPS),* vol. 36, 2023.

[3] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, «Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks,» in *Proc. IEEE Winter Conf. Applications of Computer Vision (WACV),* 2018, pp. 839–847.

[4] A. Dosovitskiy *et al.,* «An image is worth 16×16 words: Transformers for image recognition at scale,» in *Proc. Int. Conf. Learning Representations (ICLR),* 2021.

[5] European Commission, «Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act),» *Official Journal of the European Union,* 2024.

[6] T. Fel and D. Vigouroux, «Benchmarking XAI explanations with human-aligned evaluations: The PASTA framework,» *arXiv preprint arXiv: 2411.02470,* 2024.

[7] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, «Causability and explainability of artificial intelligence in medicine,» *WIREs Data Mining and Knowledge Discovery,* vol. 9, no. 4, p. e1312, 2019.

[8] F. Hvilshøj, A. Iosifidis, and A. Iraji, «Quantifying the 'Faithfulness-Efficiency Gap' in deep learning interpretability: A comprehensive survey,» *Artificial Intelligence Review,* vol. 58, pp. 102–145, 2025.

[9] V. Kumar and D. Singh, «Advancing AI interpretability in medical imaging: A comparative analysis of pixel-level interpretability and Grad-CAM models,» *Diagnostics,* vol. 15, no. 1, pp. 22–38, 2025.

[10] M. Mourby and E. Dorsey, «Medicine, healthcare and the AI Act: Gaps, challenges and future implications,» *Journal of Medical Ethics & Regulation,* vol. 12, no. 4, pp. 112–128, 2025.

[11] M. Naseer *et al.,* «Evaluating the explainability of vision transformers in medical imaging: A quantitative study on faithfully mapping pathology,» *IEEE Transactions on Medical Imaging,* vol. 44, no. 2, pp. 301–315, 2025.

[12] V. Petsiuk, A. Das, and K. Saenko, «RISE: Randomized input sampling for explanation of black-box models,» in *Proc. British Machine Vision Conf. (BMVC),* 2018.

[13] R. R. Selvaraju *et al.,* «Grad-CAM: Visual explanations from deep networks via gradient-based localization,» *International Journal of Computer Vision,* vol. 128, no. 2, pp. 336–359, 2020.

[14] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, «Training data-efficient image transformers & distillation through attention,» in *Proc. Int. Conf. Machine Learning (ICML),* 2021, pp. 10347–10357.

[15] Y. Zhang and X. Liu, «A hybrid explainable AI framework (HXAI) for accurate and interpretable diagnosis of Alzheimer's disease,» *BMC Medical Informatics and Decision Making,* vol. 25, no. 1, p. 45, 2025.