Comparative Plausibility Evaluation of Heatmaps for Vision Transformers in Digital Mammography

Ramsha Aasim[1], Ghada Zamzmi[2], Jana G. Delfino[2], Joseph Jaja[1] and Miguel A. Lago[2]

[1]University of Maryland's Institute of Health Computing, North Bethesda, MD, USA
[2]Office of Science and Engineering Laboratories, Center for Devices and Radiological Health, U.S. Food and Drug Administration, Silver Spring, MD, USA

Abstract

To effectively integrate Artificial Intelligence (AI) into healthcare workflows, robust AI explainability (XAI) techniques are needed to build clinician trust and understanding of AI-driven predictions. With numerous XAI methods available, objectively determining the most suitable method for a given task — beyond subjective human assessment —remains a challenge. This study tackles this challenge by introducing an evaluation framework that enables explainability assessments to foster trust and accelerate AI adoption in clinical settings. We demonstrate the plausibility criterion - defined as the degree to which AI explanations align with established ground truth for a given task —as a means to quantitatively assess XAI. To illustrate our evaluation criteria in medical imaging, we compared five post hoc heatmap methods applied to a Vision Transformer (ViT) trained for lesion detection in digital mammography. For each technique, we generate and compare heatmaps, and then examine how varying the mammogram's radiation dose affect their plausibility. Our analysis shows how this framework quantitatively measures each heatmap method's ability to highlight diagnostically relevant regions, providing an objective way to evaluate AI explanation quality. We conclude by emphasizing that, while plausibility is a key metric, it should be considered alongside other relevant criteria. This work is part of our broader efforts to develop a comprehensive framework to quantitatively assess the quality of AI explanations in the regulatory evaluation of medical imaging devices.

Keywords

AI Explainability (XAI), Regulatory Evaluation, Vision Transformers (ViTs), Heatmaps

1. Introduction

Vision Transformers (ViTs) [1] have rapidly gained prominence in computer vision as they offer a compelling alternative to Convolutional Neural Networks (CNNs) for tasks such as image classification, segmentation, and object detection. ViTs primarily utilize the self-attention mechanism [2], in order to capture global dependencies, offering minimal inductive bias. In medical imaging, ViTs have proved [3] to perform well for applications such as segmentation [4], detection [5], classification [6], and reconstruction [7].

Despite their widespread popularity, transformer models have a complex architecture, making it difficult to understand and explain their predictions. Explainable AI (XAI) aims to tackle this issue by offering techniques that provide insights into the model's reasoning [8, 9, 10]. However, the field still lacks a comprehensive and standardized framework for evaluating the effectiveness of these explanation techniques [11, 12].

This study proposes an evaluation framework specifically designed to assess the effectiveness of heatmap-based explanations for ViTs. In particular, we focus on heatmap-based explanations generated by attribution methods, which assign a relevance score to each input feature (image pixels, in our case) based on its contribution to the final prediction; a higher positive value indicates a greater impact on the model's prediction. While each explainability method is designed to generate 'localized' and 'relevant' explanations, the resulting heatmaps often vary significantly. This inconsistency, along with the lack of objective evaluation techniques for these generated explanations, makes it difficult

EXPLIMED 2025 - Second Workshop on Explainable Artificial Intelligence for the medical domain - 25-30 October 2025, Bologna, Italy

raasim@umd.edu (R. Aasim); miguel.lago@fda.hhs.gov (M. A. Lago)

to confidently rely on and trust these explanations in practice. To address this challenge, our study explores "plausibility"—defined as the alignment of a generated heatmap with an established ground truth—as a criterion to assess the quality and localization of heatmaps. To demonstrate this approach, we evaluate and compare the plausibility of heatmaps generated by five AI explainability methods applied to a ViT trained for lesion detection in digital mammography data.

2. Related Work

One of the most widely used explainability methods is heatmaps, which are designed to visualize the regions of input data that contribute most significantly to a model's prediction. Heatmaps have become an essential tool for explaining CNNs and highlighting discriminative regions in the input. CAM based techniques are among the most prevalent explainability techniques for heatmap generation for CNNs [8, 13, 14].These explainability have since been extented to Vision Transformers (ViTs) [15]. The main component of a ViT is the multi-head self-attention mechanism to model relationships between patch embeddings. To that end, several attention-based methods have also been introduced, including raw attention visualizations, the attention-rollout technique [9], which aggregates information from different attention heads across transformer blocks, and relevance-based explainability techniques [10, 16] that incorporate the relevance of attention heads in their explanations. Perturbation-based methods, such as ViT-CX [17], have also been utilized for generating explanations.

While the field of XAI is rapidly advancing, the research around the quantitative evaluation of these explanations remains less explored. This gap is particularly evident in medical imaging, where explanations often lack thorough evaluation [18]. For instance, Lysdahlgaard [19] compares various AI models on wrist and hand X-ray data based on the explanations generated by GradCAM. However, in the absence of standardized evaluation protocols, there is no way to tell if the GradCAM is actually generating the "right" explanations and if such explanations can be trusted. Some research studies suggest various techniques to evaluate explanations, based on sensitivity, complexity and faithfulness of explanations to the model. [20, 12, 21]. However, the field lacks a consensus on how to quantitatively assess AI explainability methods, often resorting to illustrative but potentially subjective examples [22]. Our work aims to contribute to establishing a standardized framework for evaluating heatmap explanations. We propose plausibility as a key evaluation criterion and explore different metrics for its quantification. Although the concept of evaluating heatmaps based on their ability to highlight relevant regions has been touched upon in general XAI research [23, 24], our work specifically investigates its application within the domain of medical imaging.

3. Methods

In this section, we outline the methodology used to quantitatively evaluate the plausibility of heatmap-based explainability methods applied to ViTs in the context of medical imaging.

3.1. Heatmap Explanations for ViT

We applied three CAM-based methods and two attention-based methods for explainability. For the class-discriminative explainability methods, the explanations are generated based on the model's predicted class.

3.1.1. CAM based Methods

Class Activation Map (CAM) explainability techniques were initially designed to explain CNN predictions. The core idea is that the final convolutional layer's feature maps retain information about image regions crucial for classifying the target category. A heatmap is then generated by taking a weighted combination of these feature maps and projecting it back onto the input image to highlight important regions. Different CAM methods vary in how they calculate the weights for these feature

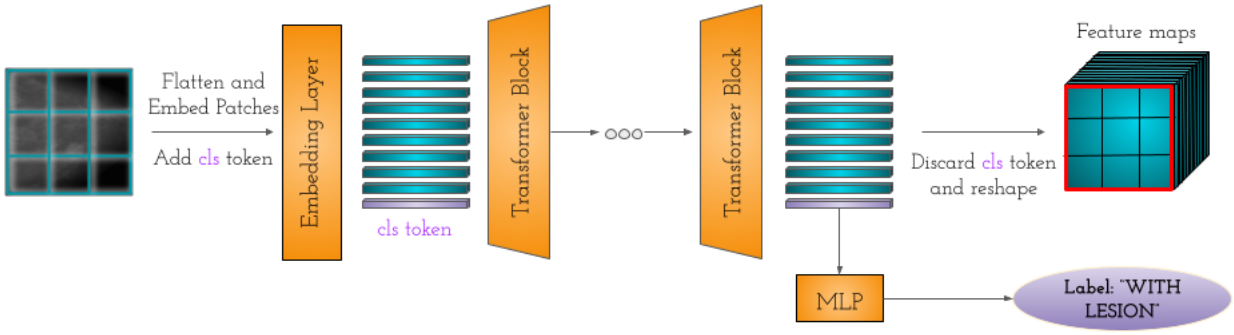


Figure 1: Vision Transformer Pipeline: Input images are divided into non-overlapping, fixed sized patches. These patches are flattened, embedded, and positional encodings are added. A class token (cls) is prepended to the sequence for classification. The embedded representations are then processed as tokens through multiple transformer blocks. The final embedding of the class token is used by a Multi-Layer Perceptron (MLP) head, consisting of fully connected layers, for the final prediction. The remaining token embeddings can be reshaped into feature maps, which can be used for CAM-based explainability techniques. (Note that since the classification head is only attached to the class token, there is no gradient flow through the feature embeddings of the final block.

maps. For example, GradCAM [8], a foundational and widely used approach, calculates the gradient of each feature map with respect to the target class and then the weights for each feature map is calculated to be the average of the gradients of that feature map. GradCAM Element-wise [25], is a variation of GradCAM, but in contrast, weighs individual feature map values with their corresponding gradient values before combining them. EigenGradCAM, an extension of EigenCAM [13], is another method that leverages the eigenvalues of the element-wise gradient weighted feature maps to generate heatmaps.

The utility of CAM methods has also been extended to explain ViT models as illustrated in Figure 1. In contrast to CNNs, ViTs process images as a sequence of fixed size, non-overlapping patches, which are embedded and subsequently transformed through multiple self-attention-based transformer blocks. Each patch is represented by a vector embedding of dimension D, a dimensionality that is maintained throughout the transformer blocks. To adapt CAM techniques for ViTs, we extract patch representations from an intermediate transformer layer, typically the second to last due to the lack of gradient flow in the final layer. These patch vectors are then rearranged to reflect their original spatial arrangement in the image, with the embedding dimension D forming the third dimension. This reshaped representation is then treated as a set of D feature maps, enabling the application of standard CAM methods by determining a suitable weighting scheme for their combination (Figure 1). For our experiments, we apply the GradCAM, EigenGradCAM and GradCAM Element-wise techniques on Vision Transformers.

3.1.2. Attention based Methods

For the attention-based explanations, we use the attention rollout method and the relevance-based explainability method.

Attention Rollout [9] is a label-agnostic technique that consolidates the flow of information within the Vision Transformer's attention mechanism. In each transformer block, multiple attention heads assign attention scores based on the importance of individual image patches for propagating relevant information to the next layer, creating a set of attention matrices that focus on different aspects of the input. The rollout method then aggregates these attention matrices—using operations such as averaging, minimum, or maximum across the attention heads—and incorporates the identity matrix (I) to account for residual connections. The resulting attention matrix (**A**) for each transformer block is recursively multiplied across all layers to produce a final matrix, which is used for generating the heatmap.

The Transformer Multi-Modal Explainability method [16], extends the Layerwise Relevance Propagation-Based Method, Beyond Attention [10]. It utilizes the relevance matrices for each attention head instead of raw attention scores. These matrices are weighted by their corresponding gradients (discarding negative values), and the resulting positive relevance is propagated across transformer layers, similar to Attention Rollout.

3.2. Heatmap Plausibility Evaluation

The plausibility of an AI model's explanation refers to how well it aligns with established ground truth, leveraging relevant background knowledge or general consensus. This is also known as "coherence" [22] or "localization" [23]. In medical imaging, a plausible explanation, such as a heatmap, should highlight clinically significant regions that correspond to a clinician's area of focus for a given task. For instance, in lesion detection, a plausible heatmap would emphasize the lesion's location.

To quantify plausibility, an ideal heatmap should assign high attribution values to pixels within the ground truth region and low values to pixels outside this area. We use three metrics to evaluate plausibility in our analysis.

3.2.1. Intersection over Union

The Intersection over Union (IoU) [26] metric measures the spatial overlap between the ground truth lesion annotation and the predicted high-intensity region derived from the heatmap. This predicted region is determined by identifying the pixel with the highest attribution value and creating a bounding box of the same dimensions as the ground truth, with the high intensity point as the center. The IoU ranges from 0 (no overlap) to 1 (perfect overlap), with higher values indicating better plausibility.

3.2.2. Relevance Mass Accuracy

Relevance Mass Accuracy quantifies the proportion of total attribution mass concentrated within the ground truth region. Given a heatmap with at least one positive attribution value, Arras et al. [24] defines it as the ratio of the sum of the positive attribution values R_k of the pixels that lie inside the ground truth (GT) region to the sum of attributions R_n of the total pixels N in the heatmap. It is mathematically given by:

$$MassAccuracy = \frac{\sum_{k \in GT} R_k}{\sum_{n=1}^{N} R_n} \tag{1}$$

The Relevance Mass Accuracy will have a value of 1 if all positive attributions of the heatmap lie inside the identified ground truth region.

3.2.3. Relevance Rank Accuracy

The relevance rank accuracy measures the number of pixels in the ground truth region that have high positive attribution values. To calculate the high attribution values, we maintain the top 10% of the attribution values in the heatmap and set the rest to 0. Let R_k be the set of pixels that contain the top k=10% of the attributions and GT is the set of all the pixels that lie inside the ground truth, the rank accuracy is given by:

$$RankAccuracy = \frac{|R_k \cap GT|}{|GT|} \tag{2}$$

3.3. Experimental Setup

3.3.1. Dataset

We used M-SYNTH dataset [27] for training the lesion detection model. M-SYNTH, developed at the U.S. Food and Drug Administration (FDA), is a synthetic dataset of 45,000 digital mammography images constructed using stochastic knowledge-based models to simulate realistic variations based on the breast fibroglandular density, lesion size and density, and radiation dose. M-SYNTH mammograms include both lesion-containing and lesion-free images. One of the key advantages of M-SYNTH is its precise annotation of lesion locations, provided in the form of pixel-level masks that accurately delineate the lesion regions. In this study, we use bounding boxes derived from these masks as the ground truth. This reliable ground truth enables a meaningful evaluation of how effectively explainability methods can highlight the relevant areas within the image.

The data used was a subset of the M-SYNTH dataset consisting of 900 images with fatty breast density, 100% recommended radiation dose, and containing lesions with diameters of 5.0, 7.0, and 9.0 mm (equally distributed) and a lesion density of 1.06. This dataset maintained a balanced 50/50 distribution of images with and without lesions. We divided this dataset into 70% training, 15% tuning, and 15% testing.

To evaluate the heatmaps for plausibility, we selected a random sample of 150 lesion-present mammogram images from the subset of M-SYNTH data with equal representations of lesions with sizes 5.0, 7.0, and 9.0.

3.3.2. Model Training and Evaluation

We employed a vanilla ViT [1] as a binary classifier to detect the presence or absence of lesions in mammograms. The lesion detection is labeled with "WITHLESION" or "NOLESION". We fine-tuned the "vit-base-patch16-224" model, pre-trained on ImageNet [28] and available via Hugging Face, using an input image of 224×224 and a patch size of 16×16. Prior to training and inference, the input images were preprocessed by normalizing pixel values with a mean and standard deviation of 0.5 and resized to 224×224 pixels. We trained the model for 7 epochs with a batch size of 12 and a learning rate of 6×10^{-6}. The model achieved a test accuracy of 99.3% and an AUC score of 0.99, which is important for ensuring that the visual explanations generated by heatmaps are meaningful.

3.3.3. Target Layer Selection for CAM methods

CNNs process images through convolutional blocks, inherently preserving spatial information. Thus, leveraging the last convolutional layer, which carries significant discriminatory information and is subsequently used for classification via fully connected layers (Multi-Layer Perceptron head, MLP), is a logical choice for CAM based techniques. In contrast, the self-attention layers in ViTs inject global information into each patch as it propagates through the network. As a result, the patch representations evolve to not only contain information about their own spatial location but also about other patches in the image. Consequently, towards the end of the network, these patch representations tend to become more abstract [9, 29], potentially rendering CAM methods that rely on the final layers less interpretable. Therefore, selecting the appropriate target layer for the generation of ViT heatmaps requires careful consideration. We explore various ViT transformer block outputs as the target layers, across the three CAM methods. We present our results in Figure 2, observing that earlier layers tend to emphasize the lesion, while later layers result in heatmaps that are more dispersed and less focused.

In order to determine the optimal target layer for our experiments, we evaluate the average plausibility score achieved by the three CAM methods across our balanced sample set. The results for this experiment are shown in Figure 3. The target layer that yielded the highest average score across all CAM methods i.e., output layer for the sixth transformer block, was selected for subsequent analysis for the CAM based methods.

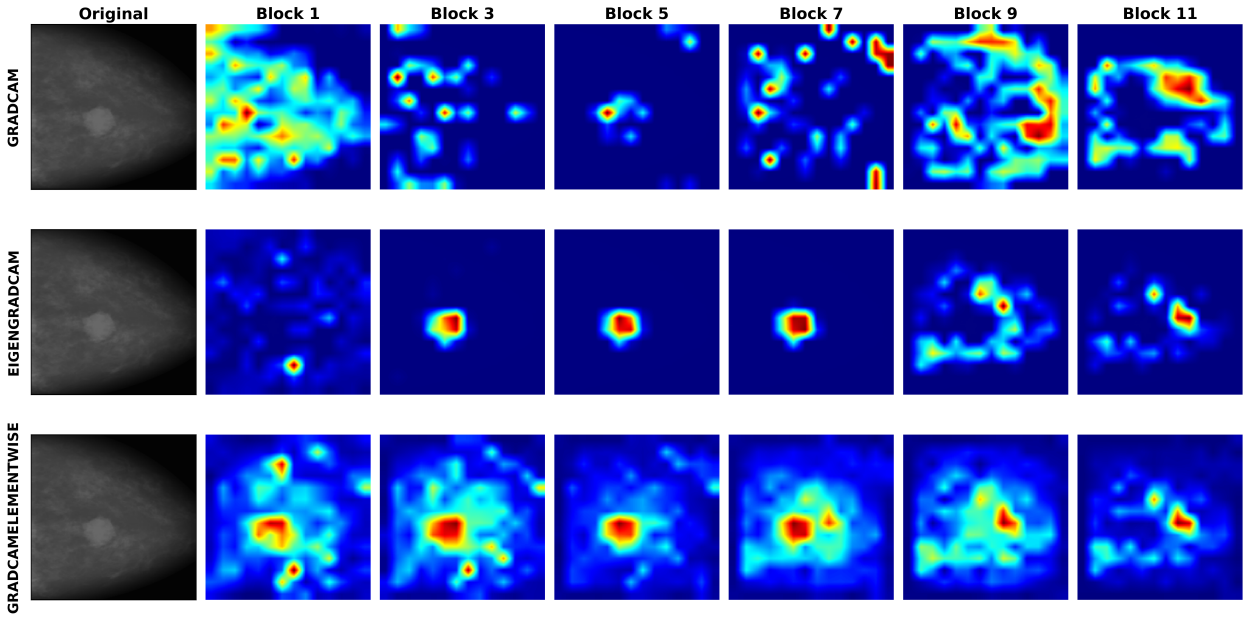


Figure 2: Comparison of heatmaps generated by GradCAM, EigenGradCAM, and GradCAM Element-wise across different transformer block output layers for a single input image.

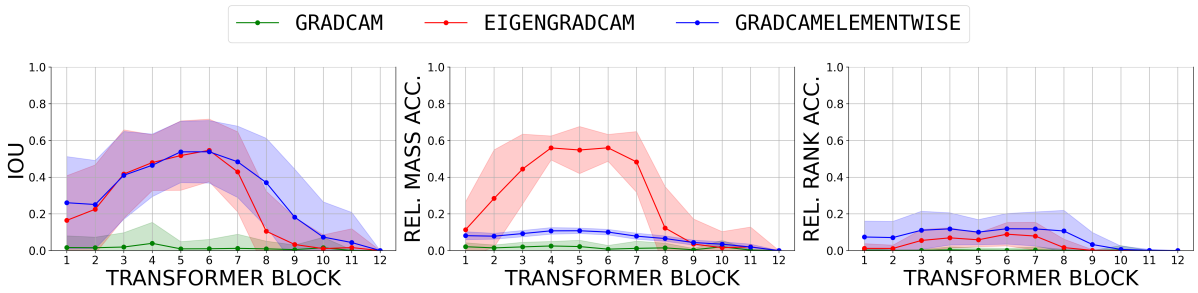


Figure 3: Change in Average plausibility scores for GradCAM, EigenGradCAM, and GradCAM Element-wise, on our sample set by using different transformer blocks output layers as target layers, with peak plausibility at layer 6

4. Results

Figure 4 presents a comparative analysis of the average plausibility scores for the explainability methods, GradCAM, EigenGradCAM, GradCAM Element-wise, Attention Rollout, and the Transformer Multi-Modal Explainability method, evaluated on our sample set. Notably, GradCAM yields very low plausibility scores for all the metrics, close to zero, indicating that its generated heatmaps are not plausible, and are broadly dispersed rather than localized. The other methods exhibit a range of plausibility scores, with EigenGradCAM and GradCAM Element-wise achieving the highest values. An interesting observation is that while EigenGradCAM and GradCAM Element-wise exhibit similar values for IoU and relevance rank accuracy—suggesting comparable performance in identifying the ground truth region—EigenGradCAM achieves a higher relevance mass accuracy. This indicates that although both methods highlight the correct region, GradCAM Element-wise also attributes high relevance scores to surrounding breast tissue, leading to less precise localization. A similar pattern is observed when comparing the attention rollout method with the Transformer MM-Explainability method. While attention rollout yields higher IoU and relevance rank accuracy, the MM-Explainability method achieves better relevance mass accuracy, suggesting more focused and localized heatmaps. Such analysis can help identify the strengths and weaknesses of various explainability methods.

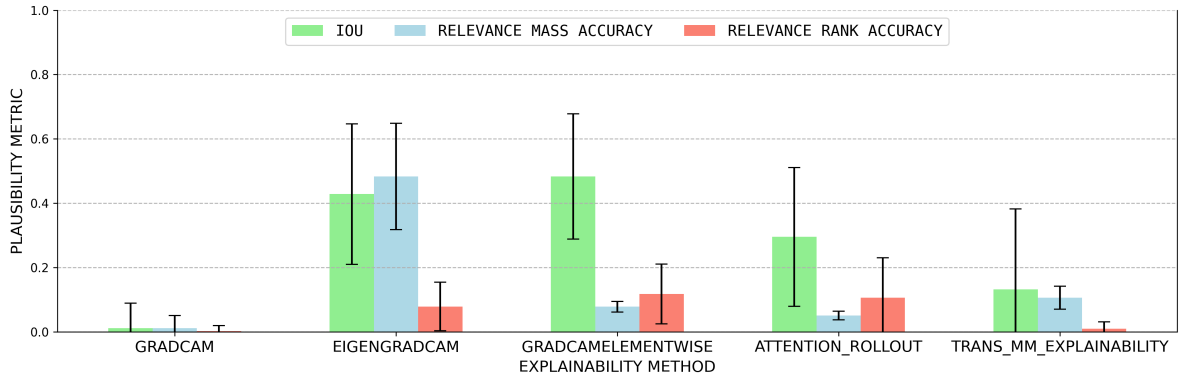


Figure 4: Average plausibility scores for heatmaps generated by different explainability methods. Higher is better.

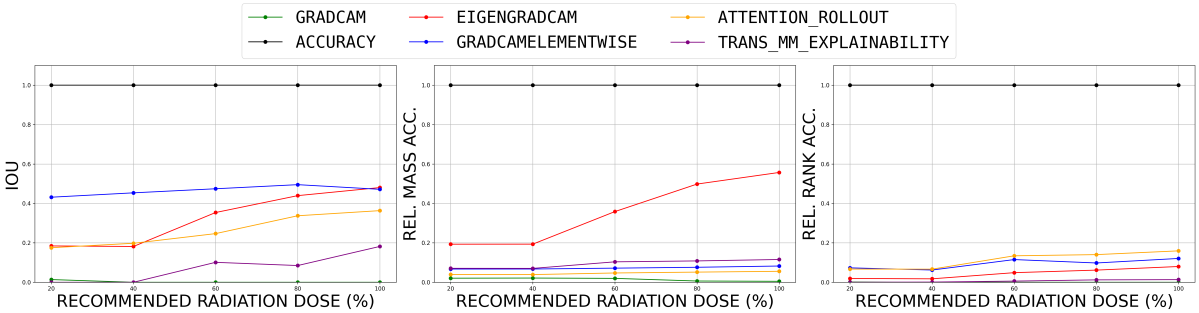


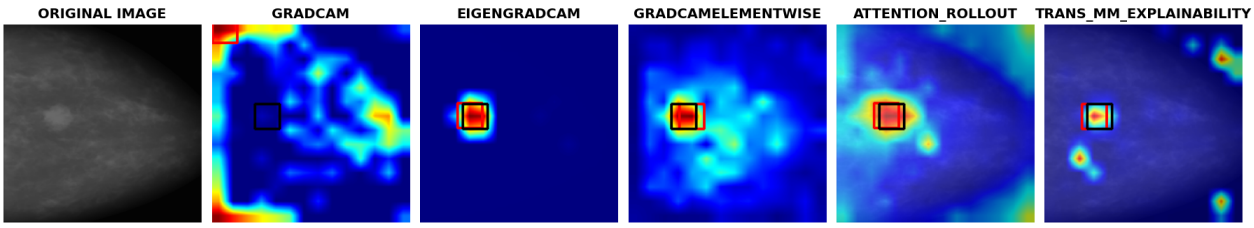
Figure 5: Plausibility Plots: Plausibility trends across varying radiation doses for (a) IoU score (b) Relevance Mass Accuracy (c) Relevance Rank Accuracy.

Figure 5 examines how varying levels of radiation dose affect the plausibility of heatmap explanations. While the model maintains consistent accuracy across all dose levels (black line) despite being trained only on the 100% relative radiation dose, the plausibility of its explanations—measured using IoU, Relevance Rank Accuracy, and Relevance Mass Accuracy—is noticeably lower at lower doses and improves as the dose approaches the recommended 100% level. This suggests that reduced image quality at lower doses may limit the model's ability to generate focused and clinically meaningful explanations. Notably, GradCAM yields consistently low plausibility scores regardless of the radiation dose, indicating limited robustness of this method across varying imaging conditions. These findings highlight the importance of considering image acquisition factors when evaluating the clinical reliability of explainability methods.

Figure 6 presents example heatmaps and their corresponding plausibility scores, illustrating the relationship between visual explanations and quantitative metrics. Consistent with our findings, the heatmaps generated by EigenGradCAM and GradCAM Element-wise appear highly localized and plausible. We also observe that while both EigenGradCAM and GradCAM Element-wise effectively highlight the lesion region, EigenGradCAM focuses more precisely on the ground truth area, assigning near-zero attribution to pixels outside of it. In contrast, GradCAM Element-wise highlights additional surrounding regions. This demonstrates that our quantative analysis captures not only the localization capabilities but also the complexity of the generated heatmap explanations.

5. Discussion and Conclusion

This study introduces plausibility as a key evaluation criterion for explainability methods in medical imaging. We quantitatively assessed the plausibility of explanations generated by a ViT model trained for lesion detection in mammograms, comparing five post hoc heatmap techniques. Our results highlight the significance of plausibility, measured using three metrics: IoU with ground truth, Relevance Mass Accuracy, and Relevance Rank Accuracy. These metrics enable objective comparison of heatmaps and



IoU	0.00	0.65	0.51	0.61	0.65
MA	0.00	0.53	0.09	0.06	0.15
RA	0.00	0.15	0.12	0.24	0.02

Figure 6: Example heatmaps generated by different explainability methods, along with their corresponding plausibility metrics.

provide insight into how well AI-generated explanations align with clinically meaningful regions, which is an important consideration for building trustworthy and interpretable medical AI systems.

A key limitation of plausibility evaluation is its reliance on a well-defined ground truth. In our study, the use of a synthetic dataset allowed us to access precise ground truth lesion annotations, only available for lesion-present trials. However, in some real-world clinical tasks, ground truth is often subjective and based on expert interpretation. This makes plausibility harder to define and measure. Further, plausibility is less applicable to global tasks such as breast density assessment, where no specific localized ground truth exists. Another challenge arises when models are trained on poor-quality data and fail to learn meaningful patterns. In such cases, an explanation that does not align with the ground truth might still be valuable, as it could reveal the model's limitations. Therefore, while plausibility is a useful evaluation criterion, it should be interpreted alongside other criteria to provide a more comprehensive assessment of explanation quality. Examples of other criteria that might need to be considered when assessing explanation quality include consistency, fidelity, and usefulness. In future work, we aim to incorporate a broader range of evaluation criteria for ViT explainability to provide a more complete and nuanced evaluation of the generated heatmaps.

This work underscores the importance of quantitative evaluation methods in selecting appropriate XAI methods for medical applications and contributes to the broader goal of developing a comprehensive framework for assessing the quality of explanations provided in medical imaging tasks.

Declaration on Generative AI

For the preparation of this work, the authors have not employed any Generative AI tools.

References

[1] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL: https://arxiv.org/abs/2010.11929. arXiv:2010.11929.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[3] R. Azad, A. Kazerouni, M. Heidari, E. K. Aghdam, A. Molaei, Y. Jia, A. Jose, R. Roy, D. Merhof, Advances in medical image analysis with vision transformers: A comprehensive review, Medical Image Analysis 91 (2024) 103000. URL: https://www.sciencedirect.com/science/article/pii/S1361841523002608. doi:https://doi.org/10.1016/j.media.2023.103000.

[4] M. Heidari, A. Kazerouni, M. Soltany, R. Azad, E. K. Aghdam, J. Cohen-Adad, D. Merhof, Hiformer: Hierarchical multi-scale representations using transformers for medical image segmentation, in: Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2023, pp. 6202–6212.

[5] Y. Wu, Q. Kong, L. Zhang, A. Castiglione, M. Nappi, S. Wan, Cdt-cad: Context-aware deformable transformers for end-to-end chest abnormality detection on x-ray images, IEEE/ACM Transactions on Computational Biology and Bioinformatics (2023).

[6] X. Huo, G. Sun, S. Tian, Y. Wang, L. Yu, J. Long, W. Zhang, A. Li, Hifuse: Hierarchical multi-scale feature fusion network for medical image classification, Biomedical Signal Processing and Control 87 (2024) 105534.

[7] A. Luthra, H. Sulakhe, T. Mittal, A. Iyer, S. Yadav, Eformer: Edge enhancement based transformer for medical image denoising, arXiv preprint arXiv:2109.08044 (2021).

[8] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 618–626.

[9] S. Abnar, W. H. Zuidema, Quantifying attention flow in transformers, CoRR abs/2005.00928 (2020). URL: https://arxiv.org/abs/2005.00928. arXiv:2005.00928.

[10] H. Chefer, S. Gur, L. Wolf, Transformer interpretability beyond attention visualization, 2021. URL: https://arxiv.org/abs/2012.09838. arXiv:2012.09838.

[11] D. Gunning, E. Vorm, J. Y. Wang, M. Turek, Darpa's explainable ai (xai) program: A retrospective, Applied AI Letters 2 (2021) e61. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/ail2.61. doi:https://doi.org/10.1002/ail2.61. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/ail2.61.

[12] C.-K. Yeh, C.-Y. Hsieh, A. S. Suggala, D. I. Inouye, P. Ravikumar, On the (in)fidelity and sensitivity for explanations, 2019. URL: https://arxiv.org/abs/1901.09392. arXiv:1901.09392.

[13] M. B. Muhammad, M. Yeasin, Eigen-cam: Class activation map using principal components, in: 2020 International Joint Conference on Neural Networks (IJCNN), IEEE, 2020, p. 1–7. URL: http://dx.doi.org/10.1109/IJCNN48605.2020.9206626. doi:10.1109/ijcnn48605.2020.9206626.

[14] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, X. Hu, Score-cam: Score-weighted visual explanations for convolutional neural networks, 2020. URL: https://arxiv.org/abs/1910.01279. arXiv:1910.01279.

[15] S. Stassin, V. Corduant, S. A. Mahmoudi, X. Siebert, Explainability and evaluation of vision transformers: An in-depth experimental study, Electronics 13 (2024). URL: https://www.mdpi.com/2079-9292/13/1/175. doi:10.3390/electronics13010175.

[16] H. Chefer, S. Gur, L. Wolf, Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers, 2021. URL: https://arxiv.org/abs/2103.15679. arXiv:2103.15679.

[17] W. Xie, X.-H. Li, C. C. Cao, N. L. Zhang, Vit-cx: Causal explanation of vision transformers, 2023. URL: https://arxiv.org/abs/2211.03064. arXiv:2211.03064.

[18] P. Komorowski, H. Baniecki, P. Biecek, Towards evaluating explanations of vision transformers for medical imaging, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE, 2023, p. 3726–3732. URL: http://dx.doi.org/10.1109/CVPRW59228.2023.00383. doi:10.1109/cvprw59228.2023.00383.

[19] S. Lysdahlgaard, Utilizing heat maps as explainable artificial intelligence for detecting abnormalities on wrist and elbow radiographs, Radiography 29 (2023) 1132–1138.

[20] U. Bhatt, A. Weller, J. M. F. Moura, Evaluating and aggregating feature-based model explanations, 2020. URL: https://arxiv.org/abs/2005.00631. arXiv:2005.00631.

[21] D. Alvarez-Melis, T. S. Jaakkola, On the robustness of interpretability methods, 2018. URL: https://arxiv.org/abs/1806.08049. arXiv:1806.08049.

[22] M. Nauta, J. Trienes, S. Pathak, E. Nguyen, M. Peters, Y. Schmitt, J. Schlötterer, M. van Keulen, C. Seifert, From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai, ACM Computing Surveys 55 (2023) 1–42. URL: http://dx.doi.org/10.1145/3583558. doi:10.1145/3583558.

[23] A. Hedström, L. Weber, D. Bareeva, D. Krakowczyk, F. Motzkus, W. Samek, S. Lapuschkin, M. M. C. Höhne, Quantus: An explainable ai toolkit for responsible evaluation of neural network explanations and beyond, 2023. URL: https://arxiv.org/abs/2202.06861. arXiv:2202.06861.

[24] L. Arras, A. Osman, W. Samek, Clevr-xai: A benchmark dataset for the ground truth evaluation of neural network explanations, Information Fusion 81 (2022) 14–40. URL: http://dx.doi.org/10.1016/j.inffus.2021.11.008. doi:10.1016/j.inffus.2021.11.008.

[25] J. Gildenblat, contributors, Pytorch library for cam methods, https://github.com/jacobgil/pytorch-grad-cam, 2021.

[26] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, S. Savarese, Generalized intersection over union: A metric and a loss for bounding box regression, 2019. URL: https://arxiv.org/abs/1902.09630. arXiv:1902.09630.

[27] E. Sizikova, N. Saharkhiz, D. Sharma, M. Lago, B. Sahiner, J. G. Delfino, A. Badano, Knowledge-based in silico models and dataset for the comparative evaluation of mammography ai for a range of breast characteristics, lesion conspicuities and doses, Advances in Neural Information Processing Systems (2023).

[28] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, L. Fei-Fei, Imagenet large scale visual recognition challenge, 2015. URL: https://arxiv.org/abs/1409.0575. arXiv:1409.0575.

[29] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, A. Dosovitskiy, Do vision transformers see like convolutional neural networks?, CoRR abs/2108.08810 (2021). URL: https://arxiv.org/abs/2108.08810. arXiv:2108.08810.