When Saliency Over-Credits Anatomy: Token-Level Evidence Flow in Frozen Medical Vision Transformers

Keun Woo Kim[0009−0001−4884−2173], Jemyoung Lee[0000−0001−7248−3652], and Mingyu Kim[0000−0003−1236−2220]

Promedius Inc., Seoul, Republic of Korea
{kwkim, jmlee, mgkim}@promedius.ai

Abstract. In medical imaging, saliency maps are among the most common ways to communicate model reasoning, yet a plausible-looking map may not faithfully reflect the evidence a model actually uses. This limitation is amplified in frozen foundation-model pipelines, where large vision transformers separate representation learning from downstream decision-making and combine global `[CLS]` summaries with local patch-token evidence. We audited the *plausibility–faithfulness* gap in a frozen chest X-ray foundation model for osteoporosis screening. Exploiting its linear pooled-patch classifier pathway, we derived an exact per-patch attribution for the pooled-patch contribution to the decision margin and used it as a branch-level faithfulness reference. On external radiographs ($n = 295$), input-gradient saliency assigns, on average, 1.92× more mass to the spine than the exact pooled-patch attribution and exceeds it in all 295 cases, while the spine is clinically plausible for osteoporosis. This reveals a false-trust failure mode: saliency can make a model appear more clinically grounded than its exact decision attribution supports. We present this as a focused interpretability audit and argue for quantitative faithfulness checks before deploying heatmap-based explanations on medical foundation models.

Keywords: Foundation models · Faithfulness · Saliency maps · Interpretability · Chest X-ray.

1 Introduction

Interpretability is a central requirement for the clinical adoption of artificial intelligence (AI) in medicine, where model outputs must be assessed not only for predictive performance but also for trustworthiness and safe use [6, 3]. This challenge has become increasingly important with the adoption of vision transformers (ViTs), foundation models, and vision–language models in medical imaging. While recent large vision models, including DINO and radiology-specific variants such as Rad-DINO, have begun to expose internal structure through token visualization, principal-component analysis of patch embeddings, or attention-based inspection [8, 9], such analyses do not necessarily establish whether a displayed explanation reflects the evidence used for a specific downstream decision.

As frozen foundation models become increasingly common in medical imaging pipelines, there is a growing need to audit whether their visual explanations are not only clinically plausible, but also faithful to the model's decision pathway.

In medical imaging, class activation maps (CAMs), gradient-based saliency maps, and related heatmap-based approaches remain widely used to make predictions more interpretable, because they are simple to generate, visually intuitive, and easy to display alongside clinical images [6]. This reflects a core distinction between an explanation's *plausibility*, whether it looks clinically reasonable, and its *faithfulness*, whether it reflects the evidence the model actually used. A plausible heatmap can be unfaithful, encouraging inappropriate trust or automation bias when it overlaps meaningful anatomy without driving the decision. Consistent with this concern, Nicolson et al. [7] showed that explanations of AI predictions had heterogeneous effects on clinicians, improving performance for some while worsening it for others.

Motivated by these concerns, we ask a complementary question in frozen ViT-based foundation models: when an exact token-level decision attribution is available for part of the classifier, does the displayed saliency map assign importance to the same anatomical evidence? We examine this through a focused interpretability audit of a frozen chest X-ray foundation model for osteoporosis prediction. Osteoporosis provides a clinically intuitive test case, as skeletal anatomy is relevant to the disease [13], and bony thoracic structures may appear visually plausible in explanation maps on frontal chest radiographs. By comparing input-gradient saliency with exact pooled-patch decision attribution derived from the model's patch tokens, we investigate whether clinically plausible saliency reflects decision evidence in a frozen medical foundation model.

Our contribution is not the general observation that saliency may be unfaithful, but a quantitative audit of the magnitude, consistency, and clinically misleading direction of this gap in a frozen medical foundation-model pipeline. Specifically, we derive an exact per-patch attribution for the linear mean-pooled patch pathway and use it as a branch-level faithfulness reference, showing that visually *plausible* input-gradient saliency can be systematically *unfaithful*, overcrediting clinically plausible anatomy relative to the exact pooled-patch attribution.

2 Related Works

Gradient-based saliency [12] and CAM-family methods are among the most common explanation tools in medical imaging, where they are typically evaluated by visual plausibility or by localization agreement with clinician-defined regions such as lesion segmentations or bounding boxes [17]. A growing body of work shows that such maps can be unstable, sensitive to implementation choices, and inconsistently aligned with annotated regions [1, 10, 15, 14], motivating caution before reading a heatmap as evidence that a model relies on clinically meaningful regions.

These limitations reflect a broader distinction between *plausibility* and *faithfulness* [4]: plausibility concerns whether an explanation appears clinically reasonable to a human observer, whereas faithfulness concerns whether it reflects the model's actual decision process. Agreement with clinician-defined anatomy speaks to the former but does not establish the latter. We adopt this perspective, treating plausible localization as insufficient on its own and comparing displayed saliency against an exact decomposable decision attribution as a reference for faithfulness.

ViTs and foundation models pose additional challenges because their representations are structured around global `[CLS]` and local patch tokens [2], rather than the spatial feature maps of convolutional networks. Recent ViT-specific explanation methods, including token-removal approaches such as Token Insight [5] and distribution-aware attribution methods such as DAVE [16], aim to better account for token-level structure when producing attribution maps. However, these methods still rely on approximate, gradient-based, or perturbation-based attribution. In contrast, the linear mean-pooled patch pathway of a frozen medical ViT admits a closed-form decomposition of the decision margin into per-patch contributions, providing an exact branch-level reference against which displayed saliency maps can be compared.

3 Methodology

We analyzed a frozen chest X-ray foundation-model pipeline for osteoporosis screening, referred to here as OsteoFM. The image encoder uses a DINOv2 ViT-L/16 backbone [8] with 24 transformer blocks and 1024-dimensional tokens, followed by a linear downstream head applied to the concatenation of the final-layer `[CLS]` token and mean-pooled patch tokens. OsteoFM is proprietary and not publicly available. The evaluation cohort comprised 295 anonymized external frontal chest radiographs collected from tertiary hospitals in North and South America, with DEXA-derived labels: 80 normal, 161 osteopenia, and 54 osteoporosis. The binary task was osteoporosis versus non-osteoporosis. All images were unseen during model development and were processed at 1024×1024 resolution. The classifier achieved ROC-AUC ≈ 0.81, indicating sufficient task performance for explanation auditing.

3.1 Exact Pooled-Patch Attribution

The downstream classifier is linear on the concatenation of the final-layer `[CLS]` token and the mean-pooled patch tokens. Let $c \in \mathbb{R}^d$ denote the `[CLS]` token and $p_i \in \mathbb{R}^d$ denote patch token i, with N patches. The classifier input z is

$$z = c \oplus \bar{p}, \qquad \bar{p} = \frac{1}{N} \sum_{i=1}^{N} p_i,$$

where $\oplus$ denotes concatenation.

For the osteoporosis decision margin, let $w = W_{\text{osteo}} - W_{\text{non}}$ be the difference between the osteoporosis and non-osteoporosis linear-head weight vectors, partitioned as $w = w_{\text{cls}} \oplus w_{\text{pool}}$. We define the exact contribution a_i of patch i to the mean-pooled patch pathway as

$$a_i = \frac{1}{N} w_{\text{pool}}^{\top} p_i, \tag{1}$$

so that the margin s decomposes as

$$s = w_{\text{cls}}^{\top} c + \sum_{i=1}^{N} a_i + b,$$

where b is the corresponding bias difference.

We use the patch-level map $M_{\text{faith}} = \{a_i\}_{i=1}^{N}$ as the decision-attribution reference against which saliency maps are compared. This attribution is exact for the mean-pooled patch pathway and reconstructs the model margin, together with the `[CLS]` term and bias, to numerical precision. The `[CLS]` pathway remains a global term and is not spatially decomposed.

3.2 Saliency Comparison and Anatomical Over-Crediting

Input-gradient saliency [12] was computed as the absolute value of the gradient of the osteoporosis decision margin with respect to the input image, summed over the three input channels and average-pooled to the 64×64 patch grid, producing a non-negative saliency map $M_{\text{sal}} \in \mathbb{R}^N$, where $N = 4096$. For comparison, we also computed a Grad-CAM-style transformer attribution map, adapted from Grad-CAM [11], using activations and gradients from the final transformer block and projecting the resulting map to the same patch grid.

To quantify anatomical emphasis, we defined a fixed rectangular spine region of interest (ROI) on the chest radiograph grid. Let $\Omega_{\text{spine}} \subset \{1, \ldots, N\}$ denote the set of patch indices inside the spine ROI, and let M_i denote the value of map M at patch i. For any attribution or saliency map M, we define the spine fraction $\text{SF}(M)$ as

$$\text{SF}(M) = \frac{\sum_{i \in \Omega_{\text{spine}}} |M_i|}{\sum_{i=1}^{N} |M_i|}, \tag{2}$$

where the numerator measures the absolute map mass inside the spine ROI and the denominator measures the total absolute map mass. Since the spine ROI covers approximately 10% of the patch grid, a map with no special spinal emphasis yields $\text{SF} \approx 0.10$, giving an area-based reference for comparison.

Using the exact pooled-patch attribution map M_{faith} defined from Eq. 1, we define the anatomical over-crediting ratio AOR as

$$\text{AOR} = \frac{\text{SF}(M_{\text{sal}})}{\text{SF}(M_{\text{faith}})}, \tag{3}$$

so that values greater than one indicate that the saliency map assigns proportionally more mass to the spine ROI than the exact pooled-patch decision attribution. We report the AOR at the cohort level as the ratio of the mean saliency spine fraction to the mean exact spine fraction across the cohort. The paired difference between input-gradient and exact pooled-patch spine fractions was evaluated using a two-sided Wilcoxon signed-rank test. A natural objection is that M_{faith} assigns little mass to the spine simply because the frozen encoder does not represent spinal information, rather than because the decision pathway discounts it. To test this, we decoded ROI mean intensity from the `[CLS]` token and the mean-pooled patch representation using leave-one-out ridge regression.

4 Results and Discussion

As shown in Fig. 1, the saliency maps produced by OsteoFM consistently highlight the spine region. This appears clinically plausible, since the spine is anatomically relevant to osteoporosis assessment [13]. However, visual plausibility alone is insufficient: the maps assign greater relative mass to the spine than the model's exact pooled-patch decision contribution.

Under the localization-based criteria by which saliency maps are usually assessed [1, 10, 15, 14], a spine-focused osteoporosis map may appear clinically appropriate, yet plausibility does not imply faithfulness [4]. An explanation can emphasize meaningful anatomy more than the decision pathway supports.

As described in Section 3.1, the frozen encoder and linear head allow the pooled-patch pathway to be decomposed exactly into per-patch contributions to the decision margin, providing an exact pooled-patch reference for this branch. This is not a complete spatial decomposition of the full model, since the `[CLS]` contribution remains global, but it lets us evaluate whether spatial saliency maps assign anatomical mass consistently with the decomposable patch pathway.

Across the normal, osteopenic, and osteoporotic groups (Fig. 1), the displayed input-gradient saliency assigns high relative mass to the spine, with spine fractions of approximately 20–22%. In contrast, the exact pooled-patch attribution assigns lower spine fractions of 7–14% and is spatially diffuse. Final-block Grad-CAM also assigns lower spine emphasis than the displayed saliency.

We quantified this effect across the evaluation cohort ($n = 295$) using the spine-fraction metric defined in Eq. 2. As shown in Fig. 2, the exact pooled-patch attribution assigned a mean spine fraction of approximately 0.101, close to the anatomical area share of the ROI. In contrast, the displayed input-gradient saliency assigned a mean spine fraction of approximately 0.193, corresponding to an AOR of 1.92. This difference was significant under a two-sided paired Wilcoxon signed-rank test ($W = 0$, $p = 4.0 \times 10^{-50}$). Per image, the saliency spine fraction exceeded the exact pooled-patch spine fraction in all 295 cases, with every point lying above the equality line (Fig. 2b). Final-block Grad-CAM assigned a markedly lower mean spine fraction of approximately 0.06, below the area reference, indicating that the over-attribution was specific to input-gradient

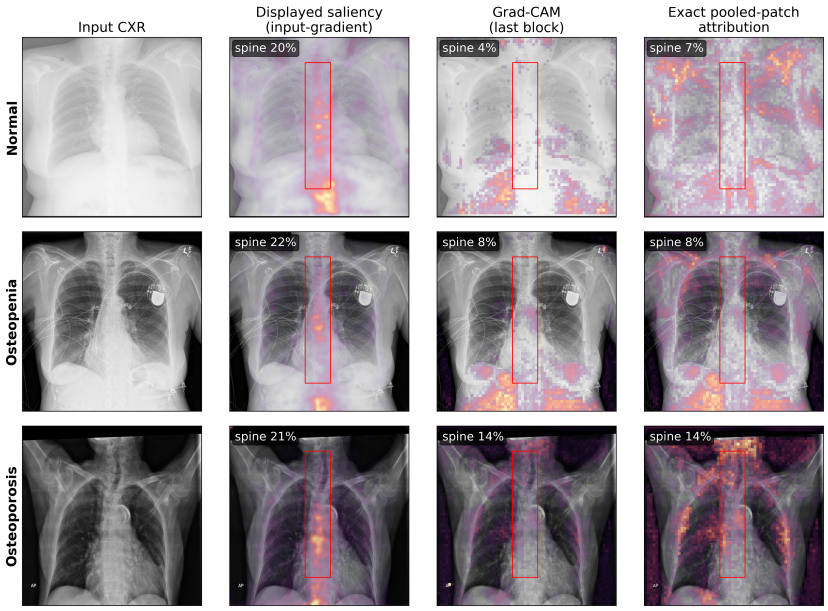
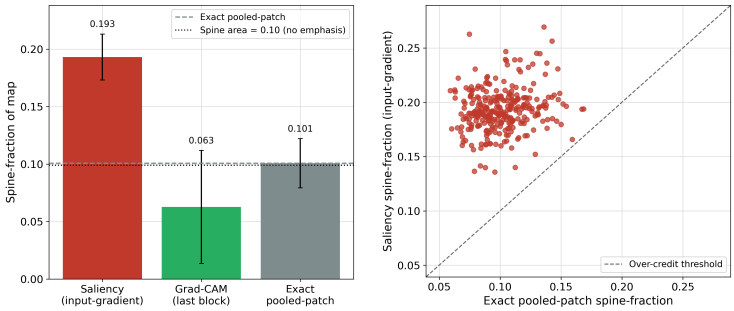


Fig. 1: Qualitative comparison across osteoporosis stages. Input-gradient saliency emphasizes the spine more than the exact pooled-patch attribution, whereas final-block Grad-CAM shows lower, less localized spine emphasis. Red boxes mark the spine ROI; percentages give the fraction of absolute map mass inside it.

saliency. However, the Grad-CAM maps were spatially diffuse at the resolution of the final block, limiting their localization as anatomical explanations.

The same pattern was preserved across osteoporosis severity strata (Table 1). Input-gradient saliency assigned high spine fractions in normal, osteopenic, and osteoporotic cases, whereas the exact pooled-patch attribution remained close to the spine-area reference across all strata. If the spine emphasis in the displayed saliency map reflected disease-specific evidence, the spine fraction would be expected to increase with osteoporosis severity. Instead, the saliency map highlights the spine to a similar degree across severity groups. This suggests that the observed spine emphasis is primarily associated with the attribution method rather than with disease-severity variation.

A possible alternative explanation is that the spine receives low exact attribution because the frozen representation does not encode spine-related information. To examine this, we decoded spine-region intensity from a single global vector per image using leave-one-out regression at the final block. Spine-region intensity was decodable from both the `[CLS]` token and the mean-pooled patch representation, with out-of-sample R^2 values of 0.70 and 0.62, respectively, where R^2 denotes the squared Pearson correlation between predicted and observed ROI mean intensity. These values were close to the local spine-patch



(a) (b)

Fig. 2: Quantitative spine over-crediting in the external evaluation cohort ($n = 295$). (a) Mean spine fraction for input-gradient saliency, final-block Grad-CAM, and exact pooled-patch attribution. The exact pooled-patch attribution lies near the spine-area share (≈ 0.10), whereas saliency lies well above it. (b) Per-image comparison: all 295 points lie above the equality line, indicating that saliency assigns more spine mass than the exact pooled-patch attribution in every case.

Table 1: Spine fraction (mean ± SD) by diagnostic group ($n = 295$). Input-gradient saliency over-credits the spine by approximately 1.9× relative to the exact pooled-patch attribution, while the exact pooled-patch map remains near the spine-area reference (≈ 0.10).

Diagnostic group	n	**Input-gradient**	**Exact pooled-patch**	**Ratio**
Normal	80	0.198 ± 0.021	0.106 ± 0.022	1.87×
Osteopenia	161	0.193 ± 0.019	0.099 ± 0.021	1.96×
Osteoporosis	54	0.185 ± 0.018	0.099 ± 0.021	1.87×
Spine area reference: ≈ 0.10				

reference ($R^2 = 0.67$) and substantially above the 95th-percentile permutation null ($R^2_{95} \leq 0.009$). These results indicate that the spine signal is available in the frozen representation and in the pooled-patch feature used by the classifier head. The low exact spine fraction should therefore not be interpreted as a failure to represent the spine. Rather, the discrepancy indicates that the displayed saliency assigns greater anatomical emphasis to the spine than is supported by the exact pooled-patch decision contribution. Although the full margin also includes a global `[CLS]` term that a_i in Eq. 1 omits, the mean absolute `[CLS]` and pooled-patch contributions were comparable (2.16 vs. 2.28), and over-crediting persisted when saliency was restricted to the pooled-patch term alone, matching the pathway used to compute a_i (AOR = 1.75, 294/295 images). Thus, the observed gap cannot be explained solely by omitted `[CLS]`-routed evidence.

In addition, the input-gradient saliency is computed through the classifier head, so the gradients of the trained head could, in principle, drive the observed over-crediting. To ensure that this behavior is not an artifact of the specific trained head, we recomputed the spine fraction under readouts that do not depend on it: a set of 20 random linear heads and a head-agnostic feature-norm readout ($\|\text{feature}\|$). Both over-credited the spine, with mean spine fractions of 0.158 ± 0.020 and 0.153, respectively, versus 0.099 for the area baseline and 0.193 for the trained osteoporosis head. Of the 0.094 excess over the area baseline, approximately 0.059 ($\sim$63%) is attributable to the frozen encoder and only approximately 0.035 ($\sim$37%) to the task-head increment, which sits just above the 95th percentile of the random-head distribution ($\text{SF}_{95} = 0.192$). These controls suggest that spine over-crediting is not solely explained by the trained osteoporosis head and may partly arise from encoder- or position-related properties of the frozen backbone. One possibility is that this reflects sensitivity to the spine's stable central position rather than task-specific decision contribution. However, further controls are required to isolate the precise mechanism.

These findings show that a visually plausible saliency map is not necessarily a decision-faithful explanation. We do not claim that the model ignores the spine, nor that all saliency or CAM methods fail. Rather, we show that the displayed input-gradient saliency assigns greater anatomical emphasis to the spine than is supported by the exact pooled-patch decision contribution. Although the exact pooled-patch attribution provides a more direct faithfulness reference for this branch, it should not be interpreted as a complete clinician-facing explanation due to its inability to spatially account for the global `[CLS]` contribution. In practice, saliency-based explanations could therefore be accompanied by branch-level faithfulness checks where exact decomposition is available, while future work should investigate spatial decomposition of the `[CLS]` pathway.

Limitations This study is a focused audit of one frozen chest X-ray foundation-model pipeline, and its conclusions should be interpreted within that scope. The reference attribution is exact only for the linear mean-pooled patch pathway; the `[CLS]` contribution remains global and is not spatially decomposed, so it should be treated as a branch-level faithfulness reference rather than a complete spatial explanation of the full decision. The two maps also measure different quantities: input-gradient saliency captures input sensitivity through the full model, whereas the exact pooled-patch attribution captures feature-level contribution to the pooled-patch margin. Their disagreement therefore indicates a faithfulness gap relative to this pathway, not that saliency methods are universally invalid.

In addition, the spine ROI is a fixed geometric region rather than an expert anatomical segmentation and may include adjacent tissue, while the decoding analysis uses ROI mean intensity only as a proxy for recoverable spine information, not disease-specific evidence. Finally, the evidence is limited to one cohort, one task, one backbone, and a limited set of explanation methods; extending the audit to additional models, regions, tasks, and attribution approaches remains future work.

5 Conclusion

In this study, we examined the plausibility and faithfulness of saliency maps for a frozen chest X-ray foundation model used for osteoporosis screening. The displayed input-gradient saliency maps consistently over-credited the spinal region: they were visually plausible, yet assigned greater anatomical emphasis to the spine than was supported by the model's exactly decomposable pooled-patch decision pathway. Grad-CAM, by contrast, did not over-credit the spine in this setting, but was spatially diffuse and therefore harder to interpret as an anatomical explanation. These findings suggest that clinically plausible heatmaps should not be treated as direct evidence of model reliance unless they are accompanied by quantitative faithfulness checks.

Disclosure of Interests. All authors are employees of Promedius Inc., which developed and owns OsteoFM, the proprietary model evaluated in this study; its implementation and training details cannot be disclosed for commercial and intellectual-property reasons. The authors declare no other competing interests.

References

1. Arun, N., Gaw, N., Singh, P., Chang, K., Aggarwal, M., Chen, B., Hoebel, K., Gupta, S., Patel, J., Gidwani, M., et al.: Assessing the Trustworthiness of Saliency Maps for Localizing Abnormalities in Medical Imaging. Radiology: Artificial Intelligence **3**(6), e200267 (2021). https://doi.org/10.1148/ryai.2021200267
2. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: Proceedings of International Conference on Learning Representations. IEEE (2021)
3. Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., Scardapane, S., Spinelli, I., Mahmud, M., Hussain, A.: Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence. Cognitive Computation **16**(1), 45–74 (2024). https://doi.org/10.1007/s12559-023-10179-8
4. Jacovi, A., Goldberg, Y.: Towards Faithfully Interpretable NLP Systems: How should we define and evaluate faithfulness? In: Proceedings of the 58th annual meeting of the Association for Computational Linguistics. pp. 4198–4205 (2020). https://doi.org/10.18653/v1/2020.acl-main.386
5. Kang, S., Vankerschaver, J., Ozbulak, U.: Identifying Critical Tokens for Accurate Predictions in Transformer-based Medical Imaging Models. In: International Workshop on Machine Learning in Medical Imaging. pp. 169–179. Springer (2024). https://doi.org/10.1007/978-3-031-73290-4_17
6. Lai, T.: Interpretable Medical Imagery Diagnosis with Self-Attentive Transformers: A Review of Explainable AI for Health Care. BioMedInformatics **4**(1), 113–126 (2024). https://doi.org/10.3390/biomedinformatics4010008
7. Nicolson, A., Bradburn, E., Gal, Y., Papageorghiou, A.T., Noble, J.A.: The human factor in explainable artificial intelligence: clinician variability in trust, reliance, and performance. npj Digital Medicine **8**(1), 658 (2025). https://doi.org/10.1038/s41746-025-02023-0

8. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: DINOv2: Learning Robust Visual Features without Supervision. arXiv preprint arXiv:2304.07193 (2023). https://doi.org/10.48550/arXiv.2304.07193
9. Pérez-García, F., Sharma, H., Bond-Taylor, S., Bouzid, K., Salvatelli, V., Ilse, M., Bannur, S., Castro, D.C., Schwaighofer, A., Lungren, M.P., et al.: Exploring scalable medical image encoders beyond text supervision. Nature Machine Intelligence **7**(1), 119–130 (2025). https://doi.org/10.1038/s42256-024-00965-w
10. Saporta, A., Gui, X., Agrawal, A., Pareek, A., Truong, S.Q., Nguyen, C.D., Ngo, V.D., Seekins, J., Blankenberg, F.G., Ng, A.Y., et al.: Benchmarking saliency methods for chest X-ray interpretation. Nature Machine Intelligence **4**(10), 867–878 (2022). https://doi.org/10.1038/s42256-022-00536-x
11. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 618–626 (2017)
12. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. arXiv preprint arXiv:1312.6034 (2013). https://doi.org/10.48550/arXiv.1312.6034
13. Sözen, T., Özışık, L., Başaran, N.Ç.: An overview and management of osteoporosis. European journal of rheumatology **4**(1), 46 (2016). https://doi.org/10.5152/eurjrheum.2016.048
14. Stubbin, A., Chyrikov, T., Zhao, J., Chajo, C.: The Limits of Perception: Analyzing Inconsistencies in Saliency Maps in XAI. arXiv preprint arXiv:2403.15684 (2024). https://doi.org/10.48550/arXiv.2403.15684
15. Venkatesh, K., Mutasa, S., Moore, F., Sulam, J., Yi, P.H.: Gradient-Based Saliency Maps Are Not Trustworthy Visual Explanations of Automated AI Musculoskeletal Diagnoses. Journal of Imaging Informatics in Medicine **37**(5), 2490–2499 (2024). https://doi.org/10.1007/s10278-024-01136-4
16. Wróbel, A., Gairola, S., Tabor, J., Schiele, B., Zieliński, B., Rymarczyk, D.: Dave: Distribution-aware attribution via vit gradient decomposition. arXiv preprint arXiv:2602.06613 (2026). https://doi.org/10.48550/arXiv.2602.06613
17. Xi, S., Wang, S., Safari, M., Hu, M., Tian, Z., Yang, X.: Uncertainty as a Foundation for Trustworthy Medical Imaging AI: A Comprehensive Review. Authorea Preprints (2025). https://doi.org/10.36227/techrxiv.176347970.02094782/v1