Proceedings of Machine Learning Research **318** (2026)
The 39th Canadian Conference on Artificial Intelligence

Canadian AI 2026
25–29 May 2026, Vancouver, BC

A Quantitative Evaluation Protocol for Assessing the Clinical Usefulness of 3D Saliency Explanations for MRI-based Alzheimer's Classification

Tamal Chakroborty*, Yang Liu,
Department of Physics and Computer Science, Wilfrid Laurier University,
Waterloo, ON N2L 3C5, Canada

Abstract

While Explainable AI (XAI) is widely considered essential for building clinical trust in MRI-based 3D deep learning models for Alzheimer's disease (AD) detection, the clinical validation of these explanations is insufficiently rigorous. Current evaluation protocols for assessing clinical usefulness rely mainly on subjective visual inspections or limited attributions 'top-k' regional overlap measures. These methods do not offer a standardized benchmark, making it difficult to objectively determine which explanation method most accurately aligns with the complex and distributed nature of neurodegenerative pathology. To address this gap, this paper proposes a quantitative evaluation protocol for assessing the clinical usefulness of 3D saliency maps through metric-based anatomical alignment. We implement a comprehensive scoring system based on AD neuropathology that assigns clinical importance weights to anatomical regions, allowing for mathematical verification of explanation integrity. This protocol provides a structured approach for evaluating explanation methods, advancing empirical alignment between XAI outputs and established pathological evidence.

Keywords: Explainable AI, Deep Learning, Alzheimer's Disease, Saliency Maps

1. Introduction

While 3D deep learning models have achieved high diagnostic accuracy in distinguishing AD from Cognitively Normal (CN) or other impaired subjects using sMRIs [1], their clinical adoption is hindered by their "black-box" nature. For the clinical application of such diagnostic models, especially when distinguishing changes from normal aging is challenging, it is essential that the reliability of the model is verified not only with classification accuracy but also by ensuring explanatory integrity. This emphasizes ensuring that a model's decisions are based on the genuine neuropathology of AD, rather than imaging artifacts or non-specific global markers.

To ensure explanatory integrity, post-hoc XAI methods, such as saliency maps, have become standard practices for visualizing model attention [2]. Apart from visual inspections, the validity of these explanations is mostly evaluated along three complementary dimensions: faithfulness or sanity check [3], robustness [4], and clinical alignment [5]. In the context of assessing explanations of AD diagnosis, while previous studies have extensively covered faithfulness and robustness [2, 6], a critical gap remains in validating the clinical alignment or usefulness of these explanations through a rigorous assessment. Current evaluation methods primarily rely on attribution hit rates for the top 5 to 6 important brain regions [2, 6]. However, simply identifying a small set of top-ranked regions from saliency attribution does not provide a meaningful basis for comparing saliency methods in terms of their clinical relevance. This distinction is essential, as the spatially distributed patterns of neurodegeneration are a key clinical characteristic of AD diagnosis.

To address this gap, we propose a quantitative evaluation protocol to validate and compare 3D saliency maps based on their holistic clinical relevance. We replace traditional

* chak0440@mylaurier.ca

hit-rate tables with a comprehensive scoring system based on AD pathology. By assigning importance weights to specific anatomical regions, we quantify the alignment between the attributions of explanation methods and established neuropathological biomarkers. We measure this using two metrics that encompass two dimensions: the Cosine Similarity metric measures distributional alignment and Partial-Order Consistency (POC) assess ordinal priority. The rationale behind this combination is that a clinically useful explanation should not only align with the correct anatomy but also respect the localized hierarchy of clinical importance in the attribution distribution.

We applied the proposed protocol to five gradient-based XAI methods: Grad-CAM, Grad-CAM++, HiResCAM, Backpropagation, and Guided Backpropagation, using a DenseNet architecture trained for the binary classification of AD vs CN subjects. Our experimental results revealed considerable clinical alignment issues that standard protocols fail to isolate distinctively. Notably, the outcome revealed significant instability in Grad-CAM++, making it unsuitable for clinical 3D imaging, despite its popularity in the field of computer vision. By integrating clinical importance scoring, this study provides a reproducible methodology for selecting clinically trustworthy explanation methods.

2. Related Work

2.1. Post-hoc Explainability

Post-hoc explainability methods interpret a trained classifier without modifying its structure by attributing predictions to input features. Saliency methods are widely used in this setting, particularly for 3D models, as they highlight spatial regions influencing decisions [5, 7]. Among them, gradient-based approaches have shown strong effectiveness for AD classification [2, 6]. Accordingly, we evaluate five representative methods: Grad-CAM[8], Grad-CAM++[9], HiResCAM[10], Backpropagation[11], and Guided Backpropagation[12].

2.2. Evaluating Clinical Usefulness

Clinical usefulness evaluates whether the regions highlighted by an explanation method genuinely correspond to established clinical pathology. In the high-risk field of medical imaging, it is crucial to evaluate the explanations within the context of clinical localization [5]. Additionally, recent studies indicate that without rigorous clinical validation, a saliency map might be erroneously misinterpreted as biological evidence during visual inspection, even when it lacks true causal relevance or stability [13]. Therefore, verifying XAI outputs with established neuropathological biomarkers is essential to ensure clinical usefulness in AD diagnosis.

While saliency methods have been widely used for diagnosing 3D MRI scans, few studies have quantitatively assessed their clinical effectiveness. Most existing research has relied on a single approach, typically involving the overlap of saliency attribution with segmented brain MRI regions to calculate hit rates or total saliency values in selected areas and presented in tables for comparisons [2, 6]. To fill this critical gap, we replace fragmented regional checks with a comprehensive quantitative evaluation protocol that integrates an importance-weighted scoring system based on AD neuropathology.

3. Proposed Evaluation Protocol

To evaluate clinical usefulness of saliency maps, we establish a holistic metric protocol that bridges the gap between voxel-level saliency maps and region-level pathological knowledge. The proposed protocol consists of a three-stage process aimed at assessing the clinical alignment of saliency-based explanations for 3D neuroimaging models illustrated in Figure 1. First, we aggregate anatomical regions and weigh them using clinical importance

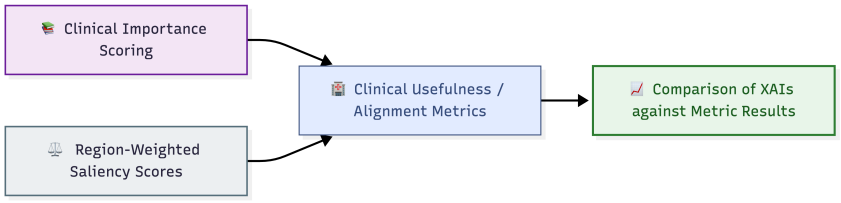


Figure 1. Overview of the proposed evaluation protocol.

derived from neuropathological evidence. Second, we generate a saliency distribution across anatomical regions by quantifying XAI attributions generated for the classifier. Finally, we compute quantitative evaluation metrics to verify clinical alignment and localization correctness of different explanation methods.

3.1. **Clinical Importance Scoring**

As finding a perfect ground truth in neuroimaging to compare the saliency findings is subjective from patient to patient, we categorized important regions and scored them for AD based on prior neuropathological findings as listed in Table 1. We use discrete numerical values instead of categorical labels because our evaluation protocol requires numerical priors for computing clinical alignment. We selected a simple ordinal scale (0, 1, 2, 3) over larger intervals (e.g., 10, 20, 30) for two reasons. First, these scores are normalized to a probability distribution to derive the metrics. Second, our evaluation metrics focus on relative ordinal ranks. Hence, larger values confer no additional advantage.

Table 1. Assigned Clinical Importance Scores for Anatomical Regions

Region Group	**Clinical Evidence**	**Importance Score (I)**
Hippocampus, ParaHippocampal, Amygdala, Temporal (Mid, Inf, Sup), Temporal Pole (Sup, Mid)	[14, 15]	3
Precuneus, Cingulum (Post, Mid), Parietal (Inf, Sup), Angular, SupraMarginal, Frontal (Sup Medial, Mid, Sup), Olfactory	[14, 16]	2
Fusiform, Insula, Cingulum Ant, Frontal (Inf Tri, Inf Oper, Sup/Mid/Med/Inf Orb), Rectus, Putamen, Thalamus, Caudate, Pallidum, Precentral, Postcentral, Supp Motor Area, Paracentral Lobule, Rolandic Oper, Heschl	[16–18]	1
Occipital (Mid, Inf, Sup), Lingual, Calcarine, Cuneus, Cerebellum (Crus1, Crus2, 3–10), Vermis (1–10)	[14, 18]	0

3.2. **Region-weighted Saliency Score**

To quantify the biological relevance of the generated saliency maps, we introduce a region-weighted saliency score based on the AAL atlas, a standardized anatomical brain map that assigns a unique integer label to each region [19]. For example, the left hippocampus is assigned the label `4101`. Given that the classifier operates directly on raw voxel grids, we ensure coordinate system consistency by normalizing both the subject MRI and the MNI152 atlas to an identity affine prior to registration. This step guarantees that atlas-defined regions for each test sample are spatially aligned with the corresponding saliency outputs, thereby eliminating potential mismatches. Building on this alignment, the following definitions establish the foundation for the downstream metrics. Let N denote the total number of regions in the AAL atlas; j index of a specific region; R_j represent the set of

voxels belonging to region j; $V_{R_j}^T$ is the total number of voxels in R_j; S_n denote the saliency value at voxel n; and I_j be the clinical importance score assigned to region j.

(1) **Normalized Importance Score:** Clinical importance scores (shown in Table 1) are normalized to form a probability distribution: $I'_j = \frac{I_j}{\sum_{k=1}^{N} I_k}$.

(2) **Raw Average Saliency:** The average saliency within region j is computed as $S_{raw}(R_j) = \frac{\sum_{n \in R_j} |S_n|}{V_{R_j}^T}$, which mitigates bias due to differences in regional volume.

(3) **Normalized Average Saliency:** To facilitate comparison across methods, regional saliency is further normalized: $S_{\text{norm}}(R_j) = \frac{S_{\text{raw}}(R_j)}{\sum_{k=1}^{N} S_{\text{raw}}(R_k)}$, yielding the proportion of total model attention attributed to region j.

The quantity S_{raw} accounts for anatomical volume effects, while I'_j and $S_{\text{norm}}(R_j)$ address scale differences. Together, these formulations provide a consistent basis for evaluating clinical relevance across XAI methods with varying output magnitudes.

3.3. Clinical Usefulness / Alignment Metrics

To evaluate clinical usefulness, we employ two complementary metrics, each ranging from 0 (complete disagreement) to 1 (perfect agreement).

Cosine Similarity: This metric quantifies the angular alignment between the normalized saliency distribution and the normalized clinical importance distribution. It captures the overall directional agreement across regions while remaining invariant to global scaling:

$$\text{Cosine Similarity} = \frac{\sum_{j=1}^{N} I'_j \, S_{\text{norm}}(R_j)}{\sqrt{\sum_{j=1}^{N} (I'_j)^2} \cdot \sqrt{\sum_{j=1}^{N} \left(S_{\text{norm}}(R_j)\right)^2}} \tag{3.1}$$

Partial-Order Consistency (POC): Adapted from constraint-based evaluation frameworks, POC measures whether the relative ordering of regions induced by saliency scores is consistent with clinically defined importance rankings. The set of pairwise ordering constraints is defined as $P = \{(R_i, R_j) \mid I_i \neq I_j\}$. The POC score is then computed as:

$$\text{POC} = \frac{\sum_{(R_i,R_j) \in P} \mathbb{1}\left[\left(I_i > I_j\right) \iff \left(S_{\text{norm}}(R_i) > S_{\text{norm}}(R_j)\right)\right]}{|P|} \tag{3.2}$$

where $\mathbb{1}[\cdot]$ denotes the indicator function, taking the value 1 when the saliency-based ranking preserves the clinical ordering between regions, and 0 otherwise.

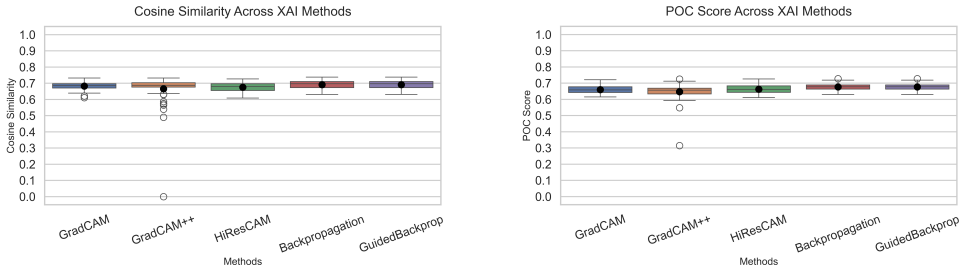
4. Experimental Setup

A total of 62 pre-processed 3D T1-weighted AD-positive cases' brain masks from the ADNI 1 cohort [20] were used. These samples were originally part of a larger dataset comprising 1300 scans (600 AD, 700 CN), which was pre-processed and split into training, validation, and test sets in prior work [2]. The selected subset consists exclusively of AD-positive cases from the test set and was strictly held out from all training and model development phases. The overall cohort from which this subset was drawn included participants aged 55-93 years (mean = 76.58 ± 6.35), with 51.92% male and 48.08% female participants.

To ensure metric variance originates from the XAI methods rather than model instability, we use a pre-trained DenseNet architecture trained for the binary classification [2]. As the post-hoc explanation parallels the classification task, researchers can readily adopt any classifier across different datasets or alternative neuroimaging diagnostic tasks that meet their specific needs without compromising the validity of the alignment metrics.

5. Results

To measure the clinical capture efficiency and spatial consistency of the explanations, we evaluated two clinical alignment metrics defined in Section 3.3. The results illustrated in Figure 2 collectively reveal a distinct performance hierarchy among the five interpretation methods, categorized by their ability to maintain spatial and ordinal clinical relevance. In terms of Cosine Similarity, Backpropagation and Guided Backpropagation achieve the highest mean demonstrating superior stability compared to other methods. Notably, Grad-CAM++ shows a high number of outliers, suggesting that it is unstable to generate a clinically aligned explanation. The result on POC scores reveals all interpretation methods



(a) Cosine Similarity (Vector Alignment) *(b)* POC Score (Localization Correctness)

Figure 2. Regional saliency evaluation metrics results.

exhibit the localized constraint of important clinical orders, with scores between 65% and 70%. Backpropagation and Guided Backpropagation tend to show that they hold the constraint at the higher end of this range. The statistical analysis of the alignment metrics in Appendix A reports the same findings, confirming the consistency of these results across the sample cohort.

6. Conclusions and Future Work

This study bridges the gap between XAI explanations and clinical usefulness by proposing a clinically grounded quantitative evaluation protocol for AD diagnosis. Our results reveal clear performance differences among methods that are not evident from visual inspection alone. In particular, Grad-CAM++ exhibits significant instability, raising concerns for clinical use, while Backpropagation and Guided Backpropagation yield identical outcomes, suggesting limited added value from ReLU-based gating in this setting. These findings emphasize the need for rigorous, multi-metric evaluation when assessing diagnostic explanations.

This work is limited to five gradient-based methods and a single DenseNet model for binary classification. Future research will extend the protocol to a wider range of XAI methods, architectures, and disease stages, as well as validate findings on larger datasets to further assess robustness and clinical reliability.

Acknowledgment

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) under Grant No. RGPIN-2023-05471.

References

[1] A. Ebrahimighahnavieh et al. "Deep Learning to Detect Alzheimer's Disease from Neuroimaging: A Systematic Literature Review". In: *Comput. Methods Programs Biomed.* 187 (2020), p. 105242.

[2] T. Chakroborty et al. "Beyond Accuracy: Explainable Deep Learning for Alzheimer's Disease Detection Using Structural MRI Data". In: *Information* 16 (2025), p. 1058.

[3] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim. "Sanity Checks for Saliency Maps". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2018, pp. 9525–9536.

[4] S. Hooker, D. Erhan, P.-J. Kindermans, and B. Kim. "A Benchmark for Interpretability Methods in Deep Neural Networks". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2019.

[5] N. Arun, N. Gaw, P. Singh, K. Chang, J. Kalpathy-Cramer, et al. "Assessing the Trustworthiness of Saliency Maps for Localizing Abnormalities in Medical Imaging". In: *Radiology: Artificial Intelligence* 3.6 (2021), e200267.

[6] J. Rieke et al. "Visualizing convolutional networks for MRI-based diagnosis of Alzheimer's". In: *MICCAI Workshop*. 2018, pp. 24–31.

[7] S.-H. Wang et al. "Saliency-based 3D CNN for categorising common focal liver lesions". In: *Insights Imaging* 12 (2021), p. 173.

[8] R. R. Selvaraju et al. "Grad-CAM: Visual explanations from deep networks". In: *IJCV* 128.2 (2020), pp. 336–359.

[9] A. Chattopadhay et al. "Grad-CAM++: Generalized gradient-based visual explanations". In: *WACV*. 2018, pp. 839–847.

[10] R. L. Draelos and L. Carin. "Use HiResCAM instead of Grad-CAM for faithful explanations". In: *arXiv preprint arXiv:2011.08891* (2020).

[11] K. Simonyan et al. "Deep inside Convolutional Networks". In: *arXiv preprint arXiv:1312.6034* (2013).

[12] J. T. Springenberg et al. "Striving for simplicity: The all convolutional net". In: *arXiv preprint arXiv:1412.6806* (2014).

[13] W. Jin et al. "One map does not fit all: Evaluating saliency on multi-modal images". In: *arXiv preprint arXiv:2107.05047* (2021).

[14] H. Braak and E. Braak. "Neuropathological Staging of Alzheimer-Related Changes". In: *Acta Neuropathologica* 82 (1991), pp. 239–259.

[15] A. Convit, M. J. De Leon, C. Tarshish, et al. "Specific Hippocampal Volume Reductions in Individuals at Risk for Alzheimer's Disease". In: *Neurobiology of Aging* 18 (1997), pp. 131–138.

[16] E. K. Miller and J. D. Cohen. "An Integrative Theory of Prefrontal Cortex Function". In: *Annual Review of Neuroscience* 24 (2001), pp. 167–202.

[17] C. T. Morgan. "The Cerebral Cortex of Man: A Clinical Study of Localization of Function". In: *Science* 112 (1950), p. 567.

[18] C. J. Stoodley, E. M. Valera, and J. D. Schmahmann. "Functional Topography of the Cerebellum for Motor and Cognitive Tasks: An fMRI Study". In: *NeuroImage* 59 (2012), pp. 1560–1570.

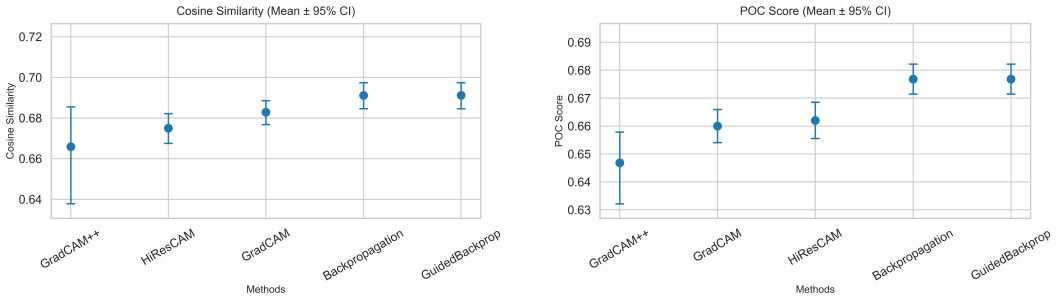
[19] N. Tzourio-Mazoyer et al. "Automated Anatomical Labeling of Activations in SPM Using a Macroscopic Anatomical Parcellation of the MNI MRI Single-Subject Brain". In: *NeuroImage* 15 (2002), pp. 273–289.

[20] *Alzheimer's Disease Neuroimaging Initiative (ADNI)*. Dataset, accessed 2025-08-07. 2025.

Appendix A. Statistical Analysis Results of Alignment Metrics

To verify the consistency of clinical alignment across the cohort, we report the mean and 95% confidence intervals (CIs) for each metric. Given the non-Gaussian distributions of saliency-based metrics and the moderate clinical sample size, we used a nonparametric bootstrap procedure with 10,000 resamples to estimate CIs without assuming normality.

The mean performance and corresponding 95% CIs for each XAI method across two evaluation metrics are presented in Figure 3. The CI ranges were found closer to each of the saliency maps for the two metrics.



(a) Cosine Similarity *(b)* POC Score

Figure 3. Mean Performance Scores and 95% Confidence Intervals for XAI Clinical Alignment Metrics.

For Cosine similarity, Backpropagation and Guided Backpropagation achieved the highest mean similarity score of 0.691, with a narrow 95% CI of [0.685, 0.697], indicating strong consistency. Similar to Cosine Similarity, both Backpropagation and Guided Backpropagation demonstrated the highest alignment with predefined clinical orders in POC metric. Notably, Grad-CAM++ exhibited the lowest mean score and the widest CI across two metrics.