


A Systematic Evaluation of Saliency Methods for 3D MRI-Based Alzheimer’s Disease Classification

Shubham Joshi¹ 

SHUBHAM_J@AMSC.IITR.AC.IN

Millie Pant^{1,2}

PANT.MILLI@AS.IITR.AC.IN

Kusum Deep³

KUSUM.DEEP@MA.IITR.AC.IN

¹ *Applied Mathematics and Scientific Computing, Indian Institute of Technology Roorkee, Roorkee, India*

² *Mehta Family School of Data Science and Artificial Intelligence, Indian Institute of Technology Roorkee, Roorkee, India*

³ *Department of Mathematics, Indian Institute of Technology Roorkee, Roorkee, India*

Editors: Under Review for MIDL 2026

Abstract

Structural MRI is essential to the assessment of Alzheimer’s disease (AD), yet the limited interpretability of deep neural networks continues to hinder their broader clinical adoption. Class Activation Map (CAM) approaches are used to visualize the decision making of deep neural networks, but their behaviour in three-dimensional neuroimaging contexts is still not well characterized. This study provides a systematic assessment of six saliency methods viz. Grad-CAM, Grad-CAM++, EigenCAM, LayerCAM, ScoreCAM, and ReciproCAM with a backbone 3D ResNet-18 classifier trained on 1540 T1-weighted MRI scans from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database. We also present an anatomically based evaluation framework that quantifies regional activation distributions through predefined volumetric regions of interest (ROIs) encompassing five structures associated with Alzheimer’s disease (AD). CAM energy fractions are calculated by region for Cognitively Normal (CN), Mild Cognitive Impairment (MCI), and Alzheimer’s Disease (AD) cohorts. In addition to these spatial analyses, we conduct a comprehensive quantitative evaluation employing deletion, insertion, and ROAD faithfulness metrics on the complete test set. Our results show significant variability between CAM approaches in faithfulness and anatomical localisation, suggesting that qualitative heatmaps by themselves might not give a complete picture of network reasoning. The study offers a reproducible benchmark of six CAM methods for evaluating explainability approaches in 3D neuroimaging, and highlights considerations for selecting reliable saliency tools for MRI-based AD diagnosis. Our code is available at https://github.com/sjiitr/Benchmarksaliency_alzheimer.git.

Keywords: Alzheimer, Neuroimaging, Class Activation Maps, Interpretable AI

1. Introduction

Neurodegenerative diseases such as Alzheimer’s disease (AD) are routinely investigated using structural magnetic resonance imaging (MRI), where characteristic patterns of cortical and subcortical atrophy provide clinically meaningful biomarkers of disease progression. With the growing availability of large, publicly accessible neuroimaging repositories such as the Alzheimer’s Disease Neuroimaging Initiative (ADNI), deep learning models have become increasingly prominent for automated MRI-based diagnosis. Three-dimensional convolutional neural networks, in particular, have demonstrated strong performance in distinguishing cognitively normal (CN) individuals from patients with mild cognitive impairment

(MCI) and AD. However, the opacity of these high-capacity models continues to hinder their acceptance in clinical settings, where transparency and interpretability are essential for decision support (Baselli et al., 2020).

Post-hoc saliency approaches, most notably Class Activation Map (CAM) based techniques (Jung and Oh, 2021) are widely employed saliency methods to visualize the regions of an image that contribute most strongly to a model’s prediction. These methods aim to reveal whether a neural network attends to clinically relevant neuroanatomy, such as the hippocampus or medial temporal lobe, or whether it inadvertently focuses on confounding or irrelevant structures. Despite their popularity, CAM methods are known to vary considerably in spatial localization, stability, and diagnostic fidelity. In the context of AD MRI, where disease-related changes can be subtle, diffuse, and highly region-specific, a systematic understanding of how different CAM techniques behave remains limited. Unlike classification accuracy, which is routinely benchmarked, the evaluation of interpretability methods has received relatively little quantitative scrutiny.

This work addresses this gap by presenting a comprehensive comparative study of six widely used saliency techniques viz. Grad-CAM, Grad-CAM++, EigenCAM, LayerCAM, ScoreCAM, and ReciproCAM applied to a 3D ResNet-18 (Hara et al., 2018) classifier trained on T1-weighted MRI scans from ADNI. By measuring region-specific activation distributions across important brain structures linked to AD, such as the hippocampus, temporal lobe, frontal lobe, parietal lobe, and ventricles, we conduct a structured anatomical assessment beyond qualitative overlays. Using faithfulness metrics based on deletion and insertion as well as the ROAD metric calculated over the whole held-out test set, we further assess each method’s dependability by providing an objective gauge of how closely each explanation captures the model’s actual decision-making process.

Overall, to the best of our knowledge this study provides one of the first systematic, anatomy-grounded evaluations of saliency methods in 3D MRI based AD classification. By integrating voxel-level region-of-interest (ROI) energy analysis, large-scale quantitative faithfulness assessment, and consistent benchmarking across multiple CAM variants, we highlight the strengths and limitations of these explainability techniques when applied to AD classification. Our findings offer practical guidance for selecting appropriate interpretation tools in clinical neuroimaging workflows and establish a foundation for future research on trustworthy MRI-based disease prediction.

2. Related Works

Deep learning has become the dominant paradigm for automated AD diagnosis from neuroimaging (Kaur et al., 2024). Early CNN-based classifiers demonstrated that volumetric neurodegeneration patterns could be reliably captured from 3D T1-weighted MRI, enabling discrimination between CN, MCI and AD subjects. More recently, lightweight residual architectures like 3D ResNet-18, have gained attention for their balance of representational capacity and computational efficiency and are now widely adopted in AD neuroimaging pipelines (Oh et al., 2023). Introducing transfer learning into 3D ResNet-18 has been shown to substantially improve AD detection over traditional 2D slice-based pipelines (Ebrahimi et al., 2020). More advanced architectures incorporating spatial-channel attention and contrastive domain adaptation further enhance robustness across heterogeneous MRI sites (Sam

et al., 2026). Other studies combine 3D CNNs with radiomic features from disease-relevant regions such as the hippocampus and amygdala to boost CN/MCI/AD discrimination (Zarei et al., 2024). Our work builds on this line of research by adopting a strong 3D ResNet backbone for three-way CN/MCI/AD classification as the basis for a systematic interpretability study. Explainable AI (XAI) has been increasingly explored in neuroimaging to mitigate the “black-box” nature of deep models and to examine whether decisions are driven by disease-relevant anatomy. Gradient-based CAM methods such as Grad-CAM and Grad-CAM++ have been widely used to visualize network attention in brain MRI, while gradient-free methods, including EigenCAM, ScoreCAM and ReciproCAM, have been proposed to improve stability and reduce sensitivity to local gradient noise. Several recent works highlight the need for more rigorous benchmarking of saliency methods in medical imaging. For instance, (Saporta et al., 2022) evaluated seven attribution techniques for chest X-ray interpretation using human-annotated pathology masks, demonstrating that all methods despite widespread clinical use significantly underperform expert localization benchmarks. Similarly, Afzal et al. (Afzal et al., 2024) conducted a large-scale study of supervised and self-supervised pre-training strategies for multi-view chest X-ray classification, emphasizing how model initialization and data regimes affect representation quality and downstream interpretability. One of the earliest quantitative CAM comparisons for brain hemorrhage detection in neuroimaging was presented by (Rafati et al., 2025) who demonstrated that localization fidelity varies significantly across CAM algorithms and network stages even in classification-only training settings. Together, these results show a growing consensus that although CAM techniques are widely used in clinical AI research, their anatomical faithfulness and dependability are still not well understood, especially for 3D MRI and progressive neurodegenerative diseases like AD. In order to fill this critical gap, our work provides a thorough, multi-method comparison of saliency techniques for structural MRI classification along with region-wise anatomical quantification and faithfulness evaluation.

3. Methods

Figure 1 presents an overview of our methodological pipeline. The study is organized into two sequential stages. In the first stage, a 3D ResNet-18 backbone is then trained from scratch using a stratified train-validation-test split that preserves the diagnostic distribution across the three clinical groups (CN, MCI, AD).

In the second stage, we employ six complementary saliency mapping techniques viz. Grad-CAM, Grad-CAM++, EigenCAM, LayerCAM, ScoreCAM, and ReciproCAM, to generate volumetric attribution maps for each subject in the held-out test set. All saliency maps are upsampled and saved in NIfTI format to enable standardized post-hoc analysis. These maps are then evaluated using region-of-interest (ROI) energy quantification as well as deletion, insertion faithfulness metrics. This integrated framework allows for a controlled and reproducible comparison of CAM methods under identical preprocessing, training, and evaluation conditions.

Implementation Details. All models are implemented in PyTorch and trained on an NVIDIA RTX 6000 Ada GPU with 48 GB of memory. Training is performed for 100 epochs using the Adam optimizer with a learning rate of 1×10^{-4} , a batch size of 4, and a cross-entropy loss objective. Data loading and preprocessing employ MONAI, NiBabel,

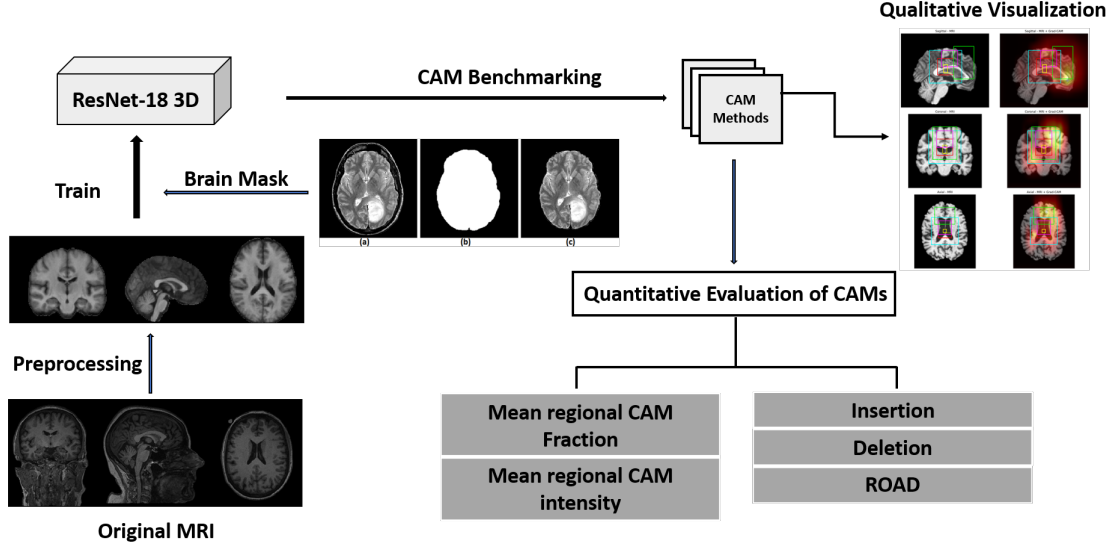


Figure 1: Framework for evaluating saliency methods

and custom PyTorch dataloaders optimized for 3D neuroimaging. All saliency methods operate on the final convolutional block of the ResNet-18 architecture and their outputs are upsampled to the original MRI resolution. A fixed random seed, preprocessing workflow, and test split are used across all experiments to ensure reproducibility.

3.1. Dataset

The Alzheimer’s Disease Neuroimaging Initiative (ADNI) ([adni](http://adni.loni.usc.edu)) is one of the most widely used public neuroimaging repositories for investigating AD, cognitive decline, and associated structural biomarkers. In this study, we utilize a curated subset comprising 1540 structural T1-weighted MRI volumes collected across the ADNI-1 and ADNI-2 phases. Each scan is provided in NIfTI format and annotated with one of three diagnostic categories having 433 CN, 748 MCI and 359 AD subjects.

All volumes were preprocessed using the *FSL* toolkit, including skull stripping, bias-field correction, AC–PC alignment, and rigid registration to the MNI152 template. The images were subsequently resampled to a standardized resolution of $182 \times 218 \times 182$ voxels. To emphasize disease-relevant neuroanatomical regions, we further applied a brain-region mask constructed from five predefined anatomical bounding regions: hippocampus, temporal lobe, frontal lobe, parietal lobe, and ventricles.

For all experiments, the dataset was partitioned into 1078 training MRI volumes, 154 validation MRI volumes, and 308 test MRI volumes, preserving the original diagnostic distribution. This fixed split was used consistently across classification and explainability analyses to ensure reproducibility and fair comparison of different CAM techniques.

3.2. Class Activation Mapping Methods

To investigate how the 3D classifier leverages different neuroanatomical structures when predicting CN, MCI, or AD, we evaluate five gradient-based and perturbation-based Class Activation Map (CAM) techniques, together with a causal attribution method. All CAM computations are adapted to 3D volumetric feature maps and subsequently upsampled to the native MRI resolution for quantitative and qualitative analyses.

Grad-CAM Grad-CAM (Selvaraju et al., 2017) produces a coarse localization volume by weighting the final-layer activations using gradients of the target class. Aggregating gradients across the depth, height, and width dimensions yields a class-specific heatmap highlighting regions that most strongly influence the prediction.

Grad-CAM++ Grad-CAM++ (Chattopadhyay et al., 2018) is the refinement of Grad-CAM to fine-tune the important weights linked to each activation channel. As a result, attribution volumes become sharper and more discriminative, facilitating the better detection of subtle disease-related patterns like early hippocampal atrophy.

Eigen-CAM Eigen-CAM (Muhammad and Yeasin, 2020) is a gradient-free technique that builds the saliency volume using the dominant eigenvector after performing principal component analysis on the activation tensor. This produces stable, class-agnostic explanations reflecting the underlying representational structure of the network.

Layer-CAM Layer-CAM (Jiang et al., 2021) uses element-wise interactions between activations and gradients at intermediate layers to form high-resolution attribution maps. By avoiding channel-wise spatial averaging, Layer-CAM retains fine-grained structural detail across the 3D feature space.

Score-CAM Score-CAM (Wang et al., 2020) is a perturbation based techniques. Individual activation maps are upsampled and injected into the input, and the resulting change in class confidence is used as an importance weight. The gradient-free strategy of Score-CAM yields smooth, noise-resistant, and class-specific saliency volumes.

Recipro-CAM Recipro-CAM (Byun and Lee, 2024) adopts a causal perturbation viewpoint by assessing how prediction scores change when activation slices are independently modified. This reciprocal interaction provides a causally grounded attribution that is less dependent on gradient quality and often more stable across volumetric scans.

All CAM volumes from these six methods are normalized, resampled to the subject’s MRI space, stored in NIfTI format, and subsequently used for ROI-level energy quantification and faithfulness evaluation.

4. Results

This section presents the empirical findings of our study, covering (i) the diagnostic performance of the 3D ResNet-18 classifier, (ii) comparative evaluation of six saliency methods using qualitative visualization, (iii) anatomical region-wise CAM energy and (iv) Faithfulness evaluation of saliency methods analysis across clinical groups. All experiments were performed on the fixed ADNI test split of 308 subjects to ensure fair comparison across explainability methods.

4.1. Classification Performance

The proposed 3D ResNet-18 classifier achieved strong diagnostic performance on the held-out test set. The overall accuracy was 84.42% across all three clinical categories. Class-wise accuracies were 82.76% for CN, 83.22% for MCI, and 88.89% for AD. These results indicate that the model distinguishes AD particularly well, while the CN–MCI boundary remains more challenging due to overlapping anatomical changes associated with aging and early cognitive decline. The high overall accuracy confirms that the learned representations are sufficiently reliable to support downstream explainability analysis.

4.2. Qualitative Visualization of Saliency Maps

We performed a qualitative comparison of six saliency approaches viz. Grad-CAM, Grad-CAM++, EigenCAM, LayerCAM, ScoreCAM, and ReciproCAM-using representative subjects from AD class. Figure 2 illustrates the spatial attribution patterns produced by these methods when overlaid on the structural T1-weighted MRI volume.

Across subjects, the gradient-weighted techniques (Grad-CAM and Grad-CAM++) produced compact and anatomically coherent hotspots, frequently centered along medial-temporal structures. LayerCAM, which incorporates spatially varying gradients, yielded finer-grained and more spatially precise attribution maps. In contrast, ScoreCAM and ReciproCAM generated higher-contrast responses with sharper boundaries, whereas EigenCAM distributed its saliency over broader cortical and subcortical territories owing to its PCA-based, gradient-free construction.

Although all methods highlighted regions plausibly associated with Alzheimer’s pathology, their spatial specificity varied considerably, underscoring the necessity of rigorous quantitative analysis extending beyond visual inspection.

4.3. Region-Wise CAM Energy Analysis

To examine whether the saliency maps highlight anatomically meaningful regions, we measured how each CAM method distributes its activation across five brain structures commonly affected in Alzheimer’s disease: the hippocampus, temporal lobe, frontal lobe, parietal lobe, and ventricular system. Since our dataset does not contain subject-specific anatomical labels, we approximated these regions by visually inspecting the preprocessed MR volumes and defining coarse three-dimensional bounding boxes using proportional scaling of the image dimensions. Although these boxes provide only an approximate representation of the underlying anatomy, they offer a consistent way to capture broad regional trends across subjects. Each CAM volume was normalized to the range $[0, 1]$ and evaluated using three complementary measures: *regional CAM energy* (S_R), *normalized regional contribution* (F_R), and *mean regional intensity* (M_R).

Let $C : \Omega \rightarrow [0, 1]$ denote the normalized CAM value at voxel $x \in \Omega$, where $\Omega \subset \mathbb{R}^3$ is the 3D brain volume. Let $R \subset \Omega$ denote one of the predefined anatomical regions, and let $B = \bigcup_{k=1}^K R_k$ denote the union of all $K = 5$ anatomical regions constituting the overall brain mask. The metrics are defined as:

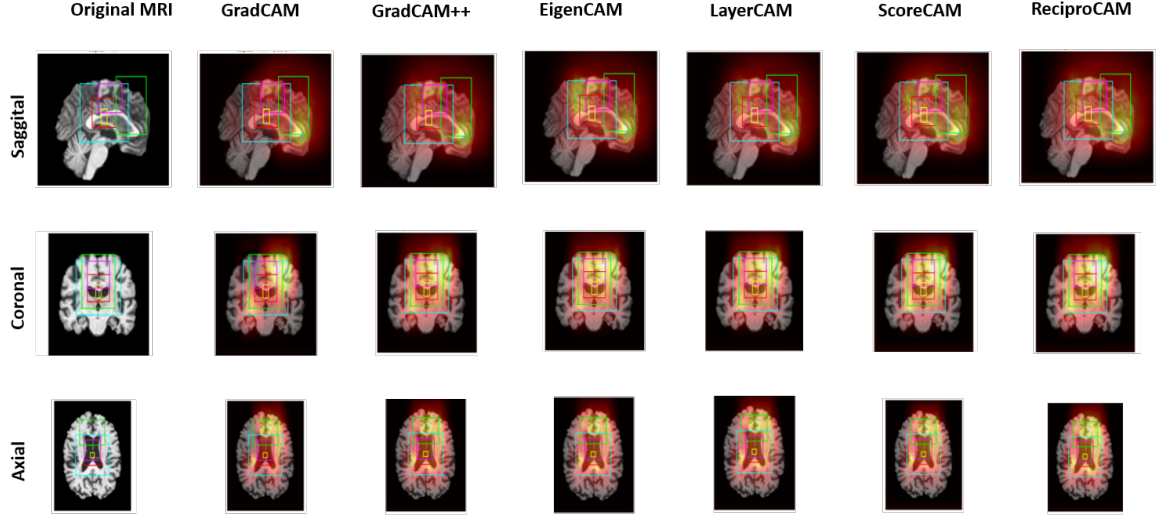


Figure 2: Qualitative comparison of saliency maps generated by six CAM methods for a representative ADNI test subject. Anatomical ROIs are visualized using the following color scheme: Hippocampus (red), Temporal Lobe (cyan), Ventricles (yellow), Frontal Lobe (lime), Parietal Lobe (magenta).

$$S_R = \sum_{x \in R} C(x), \quad (1)$$

$$F_R = \frac{\sum_{x \in R} C(x)}{\sum_{x \in B} C(x) + \varepsilon}, \quad (2)$$

$$M_R = \frac{\sum_{x \in R} C(x)}{|R| + \varepsilon}, \quad (3)$$

where:

- $C(x)$ is the CAM activation value at voxel x .
- R is the voxel set corresponding to a specific anatomical region.
- B is the union of all anatomical regions, used to restrict CAM energy to the brain.
- $|R|$ denotes the number of voxels in region R .
- ε is a small numerical constant to prevent division by zero.

Interpretation. The three metrics offer complementary viewpoints on how attention is distributed within an anatomical region using a CAM method. The total amount of saliency

accumulated within that region is measured by the quantity S_R . Comparisons between subjects and methods are made possible by the metric F_R , which expresses this value as a percentage of the total CAM energy in the entire brain volume. Lastly, M_R represents the average CAM intensity per voxel in the region, showing whether the activation is concentrated in a smaller subregion or distributed uniformly.

Findings. According to Table 1, the regional contribution values (F_R) indicate that all six saliency methods converge on attribution patterns that are clinically significant and generally similar. With fractions ranging from roughly 0.74 to 0.79, the temporal lobe consistently captures the greatest amount of total CAM energy across all techniques. The well-known involvement of medial-temporal regions in Alzheimer’s disease is consistent with this dominant focus. An important early indicator of neurodegeneration is highlighted by the hippocampus, which makes up an additional 10–11% of the saliency distribution.

Although the precise values differ slightly between methods, the frontal and parietal regions show moderate CAM fractions. While EigenCAM tends to distribute saliency more widely, leading to lower frontal values, GradCAM assigns the highest frontal contribution (0.42). Parietal fractions, which fall between 0.11 and 0.12, are relatively constant across methods. Only a small amount of activation (about 0.003) is received by ventricular regions, which is consistent with their low expected relevance to AD-specific structural alterations.

Table 1: Mean regional CAM fraction (F_R) across five anatomical regions for six saliency methods. Values represent the proportion of total brain CAM energy attributed to each ROI.

Region	GradCAM	GradCAM++	EigenCAM
Hippocampus	0.106061	0.107929	0.108327
Temporal Lobe	0.735976	0.777555	0.789272
Frontal Lobe	0.421915	0.353569	0.318243
Parietal Lobe	0.118317	0.112231	0.109431
Ventricles	0.003284	0.003089	0.003001
Region	LayerCAM	ScoreCAM	ReciproCAM
Hippocampus	0.108164	0.108914	0.106548
Temporal Lobe	0.780756	0.747297	0.774862
Frontal Lobe	0.348298	0.367127	0.357564
Parietal Lobe	0.111464	0.115654	0.109557
Ventricles	0.003091	0.003176	0.003095

The sharpness and concentration of saliency patterns across methods are further differentiated by the mean regional intensity metric (M_R), which is summarized in Table 2. Strong, localized activations within this important structure are indicated by the highest hippocampal intensities for LayerCAM, ScoreCAM, and ReciproCAM (ranging from 0.36 to 0.42). The hippocampal activations produced by GradCAM and GradCAM++ are moderately intense (0.33–0.36), whereas EigenCAM exhibits the weakest hippocampal intensity (0.29), which is consistent with its less focused and broader attribution maps.

The majority of methods show relatively high intensities in temporal and parietal regions, with parietal intensities being especially high for ScoreCAM and ReciproCAM (0.44 and 0.41), indicating that even though these regions contain a smaller fraction of total CAM energy, their activated voxels are strongly highlighted. Similar trends are seen in frontal lobe intensities: ScoreCAM and ReciproCAM produce stronger activations (> 0.36), while gradient-based techniques produce moderate values. The anatomical specificity of the CAM responses is further supported by the consistently low ventricular intensities across all techniques.

Table 2: Mean regional CAM intensity (M_R) across anatomical ROIs for six saliency methods. Values represent the average CAM activation per voxel inside each ROI.

Region	GradCAM	GradCAM++	EigenCAM
Hippocampus	0.333435	0.365186	0.294041
Temporal Lobe	0.281928	0.325150	0.265444
Frontal Lobe	0.352446	0.325452	0.237036
Parietal Lobe	0.370075	0.381657	0.298856
Ventricles	0.336553	0.337166	0.263315
Region	LayerCAM	ScoreCAM	ReciproCAM
Hippocampus	0.358444	0.417623	0.400528
Temporal Lobe	0.320234	0.356476	0.361360
Frontal Lobe	0.313201	0.378767	0.361193
Parietal Lobe	0.371090	0.443598	0.413073
Ventricles	0.330109	0.393265	0.373536

Overall, the region-level analysis shows that, despite systematic variations in anatomical precision, all saliency techniques highlight clinically significant AD-related regions. While ScoreCAM and ReciproCAM generate sharper but more spatially distributed activations, gradient-based methods (GradCAM and GradCAM++) show the strongest region-specific concentration. With diffuse attributions and lower regional intensities, EigenCAM exhibits the least anatomical specificity. These findings support the necessity of systematic anatomical assessment when contrasting saliency techniques in neuroimaging.

4.4. Faithfulness Evaluation of Saliency Methods

To assess whether saliency maps truly reflect the features driving model predictions, we evaluated all CAM methods using three perturbation-based faithfulness metrics: Deletion AUC, Insertion AUC (Petsiuk et al., 2018) and Remove-And-Debias (ROAD) Metric (Rong et al., 2022). A faithful explanation should cause a sharp confidence drop when the most important voxels are removed (low deletion AUC), a rapid confidence increase when salient voxels are added (high insertion AUC), and a substantial confidence degradation when the top 20% regions are masked (high ROAD drop).

Table 3 reports the quantitative results for 308 test subjects. Across all metrics, the gradient based Grad-CAM and Grad-CAM++ exhibit the strongest causal alignment with

model behavior, achieving the lowest deletion AUC (0.131) and the highest insertion AUC (0.498), together with the largest ROAD drop (0.710). These values indicate that gradient-based CAM methods highlight voxels that the classifier relies on most strongly for its final decision.

LayerCAM and ReciPro-CAM demonstrate intermediate faithfulness, with moderately higher deletion AUCs (≈ 0.18) and reduced ROAD drops (≈ 0.59). Although still informative, their saliency tends to be more spatially diffuse, reducing the causal specificity of the highlighted regions. ScoreCAM also exhibits weaker causal behavior, achieving the lowest insertion AUC (0.461), consistent with its smoother, mask-based attribution mechanism.

EigenCAM performs the weakest overall, yielding the highest deletion AUC (0.193) and reduced ROAD drop (0.582), indicating that its PCA-based projections do not reliably capture the discriminative features used by the classifier. Figure ?? shows the deletion and insertion trajectories for all CAM methods. Gradient-based approaches Grad-CAM and Grad-CAM++ produce the steepest decline in confidence during deletion and the fastest recovery during insertion, indicating strong causal fidelity. In contrast, LayerCAM, ScoreCAM, ReciproCAM, and especially EigenCAM exhibit flatter curves, suggesting weaker correspondence between their saliency maps and the model’s decision. These trends align with the AUC results and confirm the superior faithfulness of gradient-weighted CAM variants.

Overall, these results demonstrate that gradient-based saliency methods provide the most faithful explanations for 3D MRI classification in Alzheimer’s disease, while gradient-free methods—despite producing visually smooth maps—tend to be less aligned with true model reasoning.

Table 3: Faithfulness evaluation across CAM methods on the ADNI test set ($n = 308$). Lower deletion AUC and higher insertion AUC / ROAD indicate more faithful explanations.

CAM Method	Deletion AUC \downarrow	Insertion AUC \uparrow	ROAD Drop (20%) \uparrow
Grad-CAM	0.131	0.498	0.710
Grad-CAM++	0.131	0.498	0.710
Layer-CAM	0.181	0.485	0.589
Score-CAM	0.184	0.461	0.570
Eigen-CAM	0.193	0.470	0.582
ReciPro-CAM	0.181	0.487	0.589

5. Conclusion

Deep learning has emerged as a powerful tool for automated Alzheimer’s disease (AD) assessment from structural MRI, yet the opacity of modern neural networks continues to limit their clinical adoption. In this work, we presented a systematic evaluation of six widely used class activation mapping (CAM) methods for explaining the predictions of a 3D ResNet-based AD classifier trained on the ADNI dataset. Our analysis combined qualitative visu-

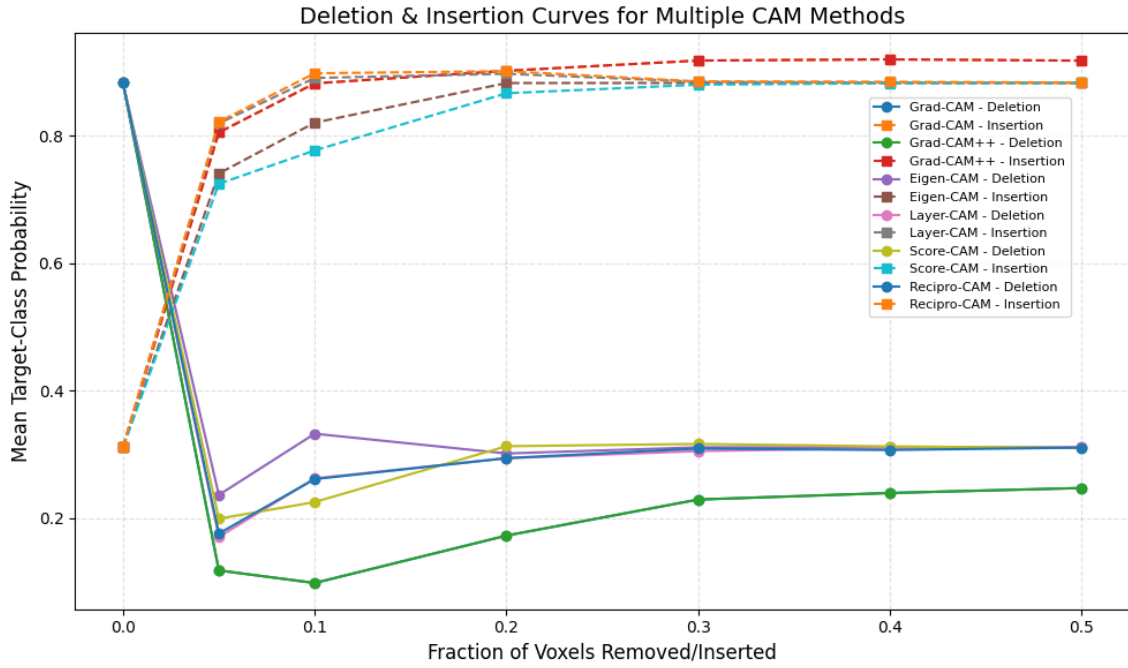


Figure 3: Comparison of deletion and insertion faithfulness curves across CAM methods on the ADNI test set. Lower deletion curves and higher insertion curves indicate more faithful explanations.

alization, region-wise anatomical attribution, and perturbation-based faithfulness metrics to provide a comprehensive characterization of saliency behavior in 3D neuroimaging.

Across these complementary evaluations, gradient-weighted approaches such as Grad-CAM and Grad-CAM++ consistently demonstrated higher anatomical specificity, concentrating saliency in regions known to be affected early in AD, including the hippocampus and temporal lobe. Gradient-free methods such as EigenCAM produced more diffuse attributions, while methods like LayerCAM, ScoreCAM, and ReciproCAM yielded intermediate behavior. Perturbation-based metrics deletion AUC, insertion AUC, and ROAD further confirmed that gradient-based methods generate the most causally faithful explanations, resulting in sharper declines or increases in model confidence under voxel perturbation.

Taken together, our findings highlight the importance of rigorous, anatomy-aware evaluation when deploying explainability tools for neuroimaging-based AD diagnosis. This work provides practical guidance for researchers and clinicians seeking to interpret 3D CNN models and lays the foundation for future studies integrating more advanced causal, multimodal, or anatomically constrained explainability frameworks. In future work, we aim to extend this benchmarking effort to alternative architectures, longitudinal prediction settings, and large-scale multi-site datasets to further strengthen the reliability and clinical utility of saliency-based explanations in neuroimaging.

Acknowledgments

Shubham Joshi acknowledge the Ministry of education, Govt. of India for providing the Prime Minister’s Research Fellowship.

References

- adni. Alzheimer’s disease neuroimaging initiative (adni). <https://adni.loni.usc.edu/>, 2025. Accessed: 2025-10-29.
- Muhammad Muneeb Afzal, Muhammad Osama Khan, and Yi Fang. A comprehensive benchmark of supervised and self-supervised pre-training on multi-view chest x-ray classification. In *Proceedings of The 7nd International Conference on Medical Imaging with Deep Learning*, volume 250 of *Proceedings of Machine Learning Research*, pages 1–16. PMLR, 03–05 Jul 2024.
- G. Baselli, M. Codari, and F. Sardanelli. Opening the black box of machine learning in radiology: can the proximity of annotated cases be a way? *European Radiology Experimental*, 4(30), 2020. doi: 10.1186/s41747-020-00159-0. URL <https://doi.org/10.1186/s41747-020-00159-0>.
- Seok-Yong Byun and Wonju Lee. Reciprocam: Lightweight gradient-free class activation map for post-hoc explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 8364–8370, June 2024.
- Aditya Chattopadhyay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 839–847, 2018. doi: 10.1109/WACV.2018.00097.
- Amir Ebrahimi, Suhuai Luo, and Raymond Chiong. Introducing transfer learning to 3d resnet-18 for alzheimer’s disease detection on mri images. In *2020 35th International Conference on Image and Vision Computing New Zealand (IVCNZ)*, pages 1–6, 2020. doi: 10.1109/IVCNZ51579.2020.9290616.
- Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Learning spatio-temporal features with 3d residual networks for action recognition. In *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 3154–3160, 2018.
- Peng-Tao Jiang, Chang-Bin Zhang, Qibin Hou, Ming-Ming Cheng, and Yunchao Wei. Layercam: Exploring hierarchical class activation maps for localization. *IEEE Transactions on Image Processing*, 30:5875–5888, 2021. doi: 10.1109/TIP.2021.3089943.
- Hyungsik Jung and Youngrock Oh. Towards better explanations of class activation mapping. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1336–1344, October 2021.
- Arshdeep Kaur, Meenakshi Mittal, Jasvinder Singh Bhatti, Suresh Thareja, and Satwinder Singh. A systematic literature review on the significance of deep learning and machine

- learning in predicting alzheimer’s disease. *Artificial Intelligence in Medicine*, 154:102928, 2024. ISSN 0933-3657. doi: <https://doi.org/10.1016/j.artmed.2024.102928>.
- Mohammed Bany Muhammad and Mohammed Yeasin. Eigen-cam: Class activation map using principal components. In *2020 International Joint Conference on Neural Networks (IJCNN)*, page 1–7. IEEE, July 2020. doi: 10.1109/ijcnn48605.2020.9206626. URL <http://dx.doi.org/10.1109/IJCNN48605.2020.9206626>.
- Kwanseok Oh, Jee Seok Yoon, and Heung-Il Suk. Learn-explain-reinforce: Counterfactual reasoning and its guidance to reinforce an alzheimer’s disease diagnosis model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4843–4857, 2023. doi: 10.1109/TPAMI.2022.3197845.
- Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018.
- Z. Rafati, M. Hoseyni, J. Khoramdel, and A. Nikoofard. Benchmarking class activation map methods for explainable brain hemorrhage classification on hemorica dataset, 2025. URL <https://arxiv.org/abs/2508.17699>.
- Yao Rong, Tobias Leemann, Vadim Borisov, Gjergji Kasneci, and Enkelejda Kasneci. A consistent and efficient evaluation strategy for attribution methods. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 18770–18795. PMLR, 17–23 Jul 2022. URL <https://proceedings.mlr.press/v162/rong22a.html>.
- Francis Sam, Zhiguang Qin, Collins Sey, Joseph Roger Arhin, Daniel Addo, Linda Delali Fiasam, Williams Ayivi, and Gladys Wavinya Muoka. Multisite t1-weighted mri classification of alzheimer’s disease using 3d-cnn-hscam architecture with contrastive domain adaptation. *Biomedical Signal Processing and Control*, 112:108686, 2026. ISSN 1746-8094. doi: <https://doi.org/10.1016/j.bspc.2025.108686>.
- Adriel Saporta, Xuan Gui, Ankit Agrawal, Anirudh Pareek, Jashvant Seekins, Arne Blattmann, Awni Khalafallah, David A. Mong, Maya Galperin-Aizenberg, and Paras Lakhani. Benchmarking saliency methods for chest x-ray interpretation. *Nature Machine Intelligence*, 4(9):867–878, 2022. doi: 10.1038/s42256-022-00536-x.
- Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017. doi: 10.1109/ICCV.2017.74.
- Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 111–119, 2020. doi: 10.1109/CVPRW50498.2020.00020.

Amin Zarei, Ahmad Keshavarz, Esmail Jafari, Reza Nemati, Akram Farhadi, Ali Gholam-rezanezhad, Habib Rostami, and Majid Assadi. Automated classification of alzheimer's disease, mild cognitive impairment, and cognitively normal patients using 3d convolutional neural network and radiomic features from t1-weighted brain mri: A comparative study on detection accuracy. *Clinical Imaging*, 115:110301, 2024. ISSN 0899-7071.