XIME^{3}D: A Systematic Framework for Evaluating Explainable AI in 3D Medical Imaging under CT Image Pre-Processing Variations

Gizem Karagoz,[1] **Tanir Ozcelebi,**[1] **Nirvana Meratnia**[1]

[1] Department of Mathematics and Computer Science
Eindhoven University of Technology
Eindhoven, The Netherlands
{g.karagoz, t.ozcelebi, n.meratnia}@tue.nl

Abstract

Recent advancements in deep learning have enabled expert-level performance in medical imaging for disease classification, but their black-box decision making processes limit trust in them and their wide-spread clinical deployment. While Explainable Artificial Intelligence (XAI) methods aim to bridge this gap, studies focus on 2D data or pre-processed research datasets that overlook the role of medical imaging pre-processing operations which is an essential component of real-world 3D medical imaging workflows.

To address this limitation, we propose XIME^{3}D, a systematic and predictive model–centered framework for evaluating explainability under realistic medical pre-processing conditions for volumetric medical data. The framework integrates five volumetric pre-processing variants and ten post-hoc attribution methods, evaluated through three complementary criteria: Correctness, Contrastivity, and Completeness, which together evaluate explanation dependence on model input, internal structure, and output behavior. Across more than 300 experimental configurations, XIME^{3}D reveals that gradient-based methods, such as Integrated Gradients and Blur Integrated Gradients, provide the most consistent and model-aligned explanations, while noise-based approaches like SmoothGrad and VarGrad are less sensitive to model behavior. These findings emphasize the importance of clinically realistic evaluation pipelines for reliable explainability in 3D medical imaging.

Introduction

Deep learning has revolutionized medical imaging and achieved expert-level accuracy for disease detection, severity assessment, and prognosis across modalities such as CT and MRI scans (Aggarwal et al. 2021). Despite these advances, most predictive models are inherently non-interpretable, which prevents users from understanding how vital decisions are made directly. In high-stakes domains like healthcare, this "black-box" behavior undermines clinical trust, hinders safe deployment, and complicates compliance with emerging AI regulations (Van Kolfschooten and Van Oirschot 2024).

Explainable Artificial Intelligence (XAI) offers a potential route toward model transparency, i.e., making the internal decision behavior of trained models more understandable to end users. However, its application in 3D medical imaging is under-explored. Recent surveys reveal that while post-hoc XAI methods are widely applied in medical imaging, the vast majority target 2D modalities such as X-ray or histopathology and volumetric explainability remain largely unexplored (Nauta et al. 2023; Van der Velden et al. 2022; Bhati, Neha, and Amiruzzaman 2024).

Although deep learning pipelines in 3D medical imaging and their explainability is explored by a few recent studies, the effect of pre-processing operations, which is an essential component of real-world CT workflows, remain largely overlooked in XAI. These operations, such as denoising, isotropic resampling, Hounsfield-unit (HU) thresholding, and anatomical masking, are indispensable for diagnostic accuracy, computational efficiency, and the visual readability of volumetric data (Masoudi et al. 2021; Singh et al. 2020). Since they directly influence the voxel intensity distributions and spatial coherence, the internal feature representations learned by predictive models are also affected. Meanwhile, most existing 3D XAI studies rely on research-oriented datasets that are already pre-processed, balanced, and well-prepared for predictive models. Consequently, prior works fail to capture how these essential steps impact the behavior of both models and their explanations, which leaves a major gap between experimental settings and real clinical workflows.

To address this gap, we introduce *XIME^{3}D* (eXplainable AI Evaluation in 3D Medical Imaging), a systematic evaluation framework to assess AI explainability for realistic medical pre-processing operations on CT volumes. The framework integrates five domain-specific pre-processing variants and ten state-of-the-art post-hoc attribution methods, evaluated on both balanced and imbalanced data splits. Each method is analyzed through three predictive model–centered evaluation criteria, namely, Correctness (C_1), Contrastivity (C_2), and Completeness (C_3). Building on the taxonomy proposed by (Nauta et al. 2023), which identifies twelve quantitative characteristics for XAI evaluation, we focus on the three that directly characterize a model's major components: input dependency, internal structure sensitivity, and output responsiveness. Together, these criteria provide a unified and systematic framework to quantify how medical pre-processing affects the explanations in 3D medical imaging.

This paper makes the following key contributions:

- **XAI Evaluation Framework for 3D Data:** We propose *XIME^{3}D*, an end-to-end evaluation framework for XAI in 3D data, that is scalable, reproducible, and directly applicable to volumetric pipelines. It offers a complete predictive model centered evaluation by systematically testing their alignment with model *input*, *internal structure*, and *output*. The three evaluation criteria —*Correctness (C_1), Contrastivity (C_2), and Completeness (C_3)*— are adapted from 2D XAI studies and reformulated for volumetric data to establish a foundation for future explainability research on 3D data.
- **Pre-processing Impact Analysis:** We conduct the first[1] systematic investigation of 3D medical image pre-processing effects on XAI. Across five pre-processing operations, ten attribution XAI methods, two dataset configurations (balanced and imbalanced), and three evaluation criteria, our study reached over 300 unique experimental configurations. This provides one of the most extensive empirical evaluations of XAI method behavior under realistic volumetric medical imaging pre-processing operations.

Related Work

Explainable AI in 3D Medical Imaging

Volumetric image analysis (e.g., CT or MRI) poses unique challenges due to the three-dimensional structure of data, variable slice resolution, and complex anatomical dependencies. Several studies have applied post-hoc attribution or saliency XAI methods to models with volumetric datasets for disease classifications across diverse domain ranging from Alzheimer's disease to breast cancer and lung disorders (Song, Yoshida, and Initiative 2024; Yang, Rangarajan, and Ranka 2018; Talaat et al. 2024; Müller-Franzes et al. 2025; Brima and Atemkeng 2024). Despite the implementation of XAI for 3D medical imaging in some studies, their evaluation tends to remain qualitative (e.g., visual heatmaps) or focused on single perspective, with limited attention to systematic robustness across pre-processing pipelines or volumetric heterogeneity.

Benchmarking and Evaluation of Explainability Methods

Beyond application-specific visualizations, a few frameworks have emerged as benchmark to quantify behavior of XAI algorithms across models and datasets. Table 1 summarizes these frameworks according to their utilized XAI methods, evaluation perspectives, used metrics, and consideration of pre-processing variability. While these benchmarking and evaluation frameworks have advanced standardized evaluation across 2D and multimodal data, they typically rely on simplified or pre-processed inputs rather than full-resolution 3D medical images. This leaves open questions about how explainability behaves under clinical conditions. None of these frameworks incorporates clinically relevant pre-processing operations or assesses their influence on explainability.

[1]To the best of our knowledge based on our conducted literature review until August 2025

Positioning of *XIME^{3}D*

Our study addresses lack of comprehensive study on impact of clinically relevant pre-processing operations on 3D data on explainability of predictive models. We propose a reproducible framework that operationalizes three complementary predictive-model-centered criteria *Correctness*, *Contrastivity*, *Completeness* as a metric for assessing the alignment with the domain knowledge to evaluate the reliability of 3D explanations under varying pre-processing operations. Unlike prior benchmarks that compare methods on standardized inputs, XIME^{3}D quantifies how clinically required pre-processing operations influence explanations against the predictive model for CT data. By operating on raw-resolution volumes, rather than downsampled standardized datasets, we provide a practical baseline for future XAI studies on 3D data aiming to integrate and evaluate explanations in medical domain.

XIME^{3}D Framework

As illustrated in Figure 1, XIME^{3}D framework consists of five key modules: (1) dataset setup, (2) pre-processing operations, (3) predictive model, (4) XAI methods, and (5) evaluation.

Medical Pre-rocessing Operations

Denoising Gaussian filtering (Perona and Malik 2002) ($\sigma = 0.5$) was applied to suppress scanner and acquisition noise by improving homogeneity in low-intensity regions, while preserving sharp organ boundaries (Masoudi et al. 2021). This operation enhances signal-to-noise ratio but slightly smooths fine textural cues relevant for edge-based attributions. To check its effect on explanations, filter was subsequently applied to all CT volumes to construct the denoised dataset variant used in the benchmark experiments.

Resampling Volumes were interpolated to an isotropic voxel spacing of $2 \times 2 \times 2$ mm to normalize heterogeneous acquisition protocols across scanners (Thévenaz, Blu, and Unser 2002; Isensee et al. 2021; Wolterink et al. 2017). This harmonization ensures consistent spatial alignment for training. Therefore, each resampled scan was saved with an updated affine transformation to reflect the new isotropic spacing for generating the resampled dataset variant.

Hounsfield Unit (HU) Thresholding Intensities outside the range $[-1000, -800]$ HU were clipped to isolate air-filled lung parenchyma (Colombi et al. 2020; Ohkubo et al. 2016; Chung et al. 2018). This removes high-density structures such as ribs and the mediastinum thus, improve tissue-specific focus and model input aligns better with diagnostic anatomy. Each CT volume was converted to a binary mask by retaining voxels within the defined HU range for generating the thresholded dataset variant.

Anatomical Masking A heuristic anatomical masking approach (Rister et al. 2020) was followed to isolate the lung parenchyma. After labeling connected regions within the

Table 1: Comparison of existing benchmark or evaluation frameworks that incorporate XAI methods in 3D medical imaging or multi-modal biomedical contexts. The table summarizes each framework by applied XAI methods, data setting, and whether they include predictive model–centered evaluations (input-, output-, and model-based) (✓) or not (✗). Pre-processing considered column indicates whether the influence of pre-processing operations is explicitly analyzed (✓) or not (✗). Finally used evaluation metrics as indicated in corresponding studies is placed. Additional Abbreviations: VG = Vanilla Gradient, GxI = Gradient×Input, DLS = DeepLIFTShap, OC = Occlusion, KS = Kernel SHAP, EG = Expected Gradients.

Reference	Applied XAI Methods	Data Setting	Predictive Model–Centered Evaluation			Pre-Processing Considered	Evaluation Metrics
			Input-based (e.g., Completeness)	Output-based (e.g., Contrastivity)	Model-based (e.g., Correctness)		
(Blake et al. 2023)	GC, LIME, RS, IG, SH, Rex	Raw / Variable	✗	✗	✗	✗	DC, PDC, JI, HD
(Nielsen et al. 2023)	VG, GxI, GB, IG, SG, GC	Preprocessed	✓	✗	✗	✗	Fidelity, Robustness, Sensitivity
(Metsch and Hauschild 2025)	DC, DL, DLS, GSH, GB, IxG, KS, LRP, LM, OCC	Raw / Variable	✓	✓	✗	✗	Faithfulness, Stability, Sensitivity
(Klein et al. 2024)	OC, KS, VG, IxG, GB, GC, SC, C+, IG, EG, DL, DLS, LRP, RA, RoA, LA	Preprocessed	✓	✓	✓	✗	Faithfulness, Robustness, Complexity
XIME^{3}D (Ours)	SL, IxG, SG, VG, IG, DL, LRP, GC, GGC, BIG	**Raw / Variable**	✓	✓	✓	✓	**Correctness, Contrastivity, Completeness**

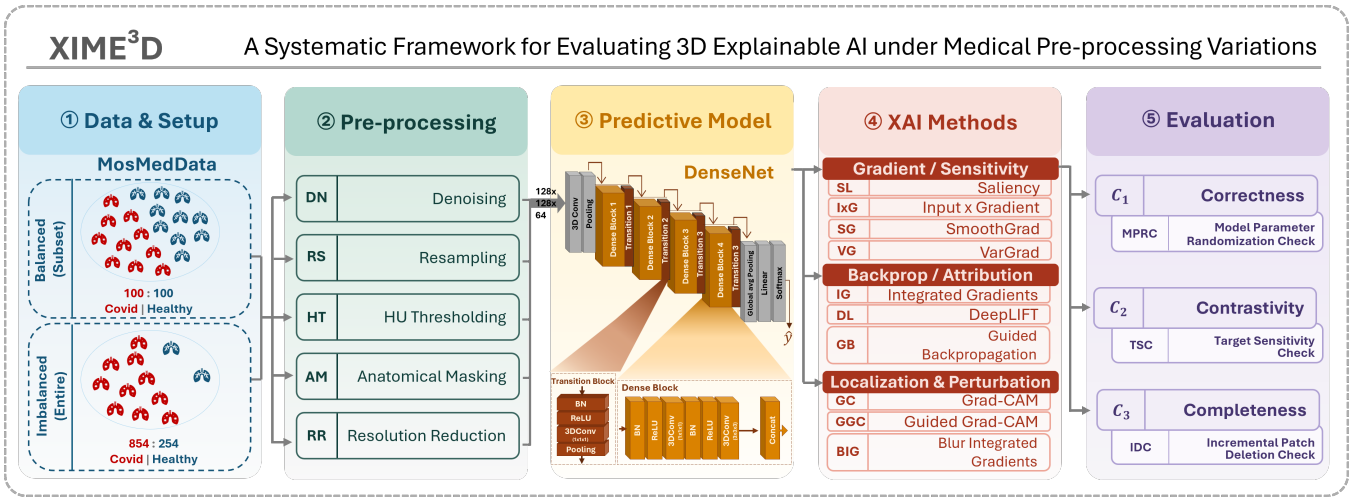


Figure 1: Overview of the proposed XIME^{3}D framework.

HU-defined lung range (-1000 to -800 HU), the largest connected component closest to the center of the volume was selected as the lung region. Components touching the outer boundaries or positioned far from the volumetric centroid were therefore discarded to avoid non-lung structures (i.e. body wall).

Resolution Reduction Prior studies have shown that spatial resolution strongly influences model saliency and localization behavior in both 2D and 3D medical imaging tasks (Sabottke and Spieler 2020). To emulate the reduced resolutions often used in research datasets, each volume was downsampled to $32 \times 32 \times 32$ voxels within the dataset and saved. This operation allows direct comparison between high-fidelity clinical pre-processing and toy-scale experimental settings, to quantify how spatial degradation influences both model performance and its explanations.

Predictive Model

Any convolutional neural network (CNN)–based architecture can be integrated into the *XIME^{3}D* framework as the predictive backbone. In this study, we utilized a 3D DenseNet (Huang et al. 2017) due to its strong generalization ability, compact feature connectivity, and consistent superiority in recent 3D medical imaging implementations (Zhou et al. 2022). From a framework perspective, the predictive model is not optimized for performance but serves as a controlled backbone whose internal feature dependencies can be systematically probed by the defined evaluation criteria (*Correctness*, *Contrastivity*, and *Completeness*).

XAI Methods

Within the *XIME^{3}D* framework, we systematically evaluate ten widely used post-hoc attribution methods, grouped into three methodological families consistent with the taxonomy:

(i) Gradient / Sensitivity-based: Saliency (SL) (Simonyan, Vedaldi, and Zisserman 2013), Input×Gradient (IxG) (Shrikumar et al. 2016), SmoothGrad (SG) (Smilkov et al. 2017), and VarGrad (VG) (Richter et al. 2020). These approaches capture local input sensitivity by computing or perturbing gradient information.

(ii) Backpropagation / Attribution-based: Integrated Gradients (IG) (Sundararajan, Taly, and Yan 2017), DeepLIFT (DL) (Shrikumar, Greenside, and Kundaje 2017), and Guided Backpropagation (GB) (Springenberg et al.

2014). They redistribute relevance through the network hierarchy to trace contribution paths from output to input.

(iii) Localization / Perturbation-based: Grad-CAM (GC) (Selvaraju et al. 2017), Guided Grad-CAM (GGC) (Selvaraju et al. 2017), and Blur Integrated Gradients (BIG) (Xu, Venugopalan, and Sundararajan 2020a). These methods spatially localize important regions by linking activations or perturbation effects to volumetric relevance maps.

Together, these three families encompass gradient sensitivity, relevance propagation, and causal perturbation perspectives. Thus, a comprehensive methodological coverage of post-hoc attribution methods for volumetric CNNs is provided.

Evaluation Criteria and Their Implementation in *XIME^{3}D*

To systematically evaluate explanations generated by XAI methods in 3D medical imaging, we define three evaluation criteria: **Correctness**, **Contrastivity**, and **Completeness**.These dimensions jointly capture how faithfully explanations reflect the model's learned parameters, decision sensitivity, and input dependency.

C_1-Correctness: Model Parameter Randomization Check (MPRC) We adopt from (Adebayo et al. 2018) to evaluate correctness of explanation methods. MPRC examines whether explanations genuinely reflect the model's learned parameters. The underlying assumption is that if model parameters are progressively randomized while keeping the input fixed, a faithful explanation should degrade accordingly.

For the 3D DenseNet backbone (see Figure 1, Lane 2), all trainable convolutional and normalization layers within each Dense Block and Transition Block were designated as randomization stages, as these layers collectively define the network's hierarchical feature representations. During randomization, all trainable parameters of Conv3D and BatchNorm3D layers were reinitialized, while non-parametric components (ReLU activations, pooling, concatenation, and dropout) remained unchanged. The randomization followed a cumulative scheme from deeper to shallower modules (`norm5` $\rightarrow$ `denseblock4` $\rightarrow$ `transition3` $\rightarrow$ `denseblock3` $\rightarrow$ `transition2` $\rightarrow$ ... $\rightarrow$ `conv0` $\rightarrow$ `norm0`) to enable a systematic analysis of how explanation faithfulness degrades as progressively larger portions of the model's representational hierarchy are disrupted.

For each randomization step, explanations were generated using multiple XAI methods and compared with their baseline (non-randomized) counterparts using the Spearman rank correlation coefficient. This metric quantifies the consistency of voxel-wise importance rankings between the baseline and the randomized models so that it captures how closely the explanation structure follows the model's learned representations.

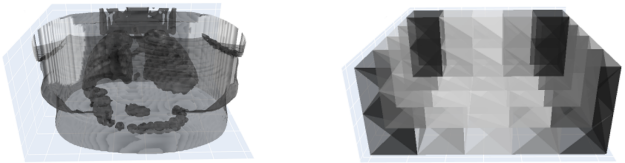
The randomization was performed cumulatively across hierarchical stages, at each step, previously randomized layers remained randomized. A correct XAI method is expected to exhibit a monotonic decline in Spearman correlation as randomization proceeds, while methods that maintain high similarity despite extensive randomization likely rely on input priors rather than the learned model parameters.

C_2-Contrastivity: Target Sensitivity Check (TSC) evaluates whether an explanation method produces distinct attribution maps for different prediction targets, reflecting contrastive faithfulness to model output changes. For each sample, explanations were generated twice (once for the predicted class and once for the alternative class) without altering the input image. The resulting attribution maps were compared using the Spearman rank correlation coefficient, which quantifies voxel-wise rank consistency between the two class-specific explanations.

A contrastive method is expected to yield low rank correlation between explanations of contrasting targets, indicating that the method responds to class-specific model reasoning rather than global input priors.

C_3-Completeness: Incremental Patch Deletion Check (IDC) evaluates the completeness of explanations by measuring how the model's confidence in its prediction degrades as the most important input regions are progressively removed. To ensure structural validity in 3D data, each CT volume was partitioned into spatially contiguous patches (16×16×8 voxels), and patch-level importance scores were computed from aggregated attributions.

At each step, the top-ranked patches were replaced with their corresponding mean intensity values rather than a fixed color to prevent global input distribution shifts (See Figure 2). Model confidence was tracked via log-odds ratio (LOR) as deletions accumulated. A complete explanation is expected to yield a rapid decline in LOR, indicating that the removed regions correspond to features critical to the model's decision-making process.



(a) Original 3D CT volume

(b) Patch-wise structure after spatial partitioning (patch size = 16×16×8).

Figure 2: 3D visualization of incremental patch deletion process. The input CT volume (a) is divided into spatially contiguous patches (b), where the highest-attributed patches were iteratively replaced with their mean intensity.

Experimental Results

All experiments were conducted on the *MosMedData: Chest CT Scans with COVID-19 Related Findings* dataset (Morozov et al. 2020), which comprises volumetric lung CT scans of healthy subjects (CT0) and patients with COVID-19 infection of varying severity levels (CT1–CT3, in increasing

order of severity). The dataset was pre-processed by DN, RS, HT, AM, and RR pre-processing operations.

As a predictive model backbone, we used a 3D DenseNet model implemented in the MONAI framework (Cardoso et al. 2022), which we trained for binary classification (non-covid vs. covid). Explanations were generated using ten post-hoc attribution methods implemented via the Captum library (Kokhlikyan et al. 2020) by utilizing default hyper-parameters to ensure comparability across methods. Blur Integrated Gradients was the only exception, following the reference implementation from its original paper (Xu, Venugopalan, and Sundararajan 2020b).

Each method was evaluated on both a balanced subset dataset (Healthy = [CT0:100], Covid = [CT1: 100]) and the entire imbalanced dataset (Healthy = [CT0: 254], Covid = [CT1: 684, CT2: 125, CT3: 45]) to assess robustness under varying class distributions and to quantify the influence of data imbalance on explanation stability. Across five pre-processing operations, ten attribution methods, two dataset settings, and three evaluation criteria, the study encompasses over **300 unique experimental configurations**, providing a comprehensive benchmarking landscape for AI explainability of 3D medical imaging.

C_1-Correctness: Model Parameter Randomization Check (MPRC)

Figure 3 shows results of the Model Parameter Randomization Check (C_1), which evaluates the *correctness* of each explanation by progressively randomizing model parameters across hierarchical layers. The randomization follows a cumulative scheme over ten sequential modules: starting from the classifier head (`norm5`, `denseblock4`) and descending through transition and dense blocks to the earliest convolutional layers (`conv0`, `norm0`). This structure allows an increasing share of the feature hierarchy to be disrupted at each randomization step.

Across most methods and pre-processing operations, a sharp performance drop occurs when the second layer is randomized (L2), corresponding to the randomization of denseblock4. Since the last convolution layer identifies the semantic representation, this result is expected and all XAI methods showed this expected behaviour.

Among individual methods, *GC* demonstrates the most pronounced and monotonic degradation, confirming its strong dependency on class-specific activations within the network's mid-to-late layers. *IG* and *IxG* exhibit similar but slightly smoother decline, reflecting that their attributions propagate gradients across all layers rather than focusing on high-level activations. *BIG* achieves a balanced behavior, as it degrades steadily and remains more stable at deeper randomization stages. This is likely because the integrated blur kernel regularizes the effect of noisy activations. In contrast, *SG* and *VG* maintain relatively flat curves throughout randomization, indicating a limited dependency of their responsiveness to model parameter changes. This can be attributed to their stochastic noise aggregation, which averages gradient noise across multiple perturbations, which makes them more robust to parameter variations but also less sensitive to structural corruption in the network. *GGC* and *GC* show notably higher standard deviations across runs, which indicates increased variability and instability of activation localization after deeper randomization, especially beyond `L2`. This variability suggests that these methods, while highly model-sensitive, are also affected by the stochastic behavior of deeper feature reactivation.

Analyzing impact of *pre-processing operations*, HT and AM yield more consistent monotonic degradation, which highlights improved model–explanation alignment once irrelevant structures are excluded. Conversely, RR and RS flatten the correlation decay, indicating that the coarse volumetric representation suppresses discriminative features, thereby diminishing layer-level sensitivity. DN produces minor shifts, suggesting that smoothing intensity noise has negligible effect on the model's parameter–explanation relationship.

Finally, comparing the *balanced subset* with the *imbalanced entire dataset* reveals that explanations from models trained on balanced data display clearer degradation patterns and smaller variance. The imbalanced setup leads to slower correlation decay, implying that class priors dominate feature attribution when data imbalance limits representational diversity. Grad-CAM, BIG, and IG emerge as the most model-faithful methods, whereas SmoothGrad, VarGrad, and Guided Backpropagation produce stable but less causally aligned explanations

C_2-Contrastivity: Target Sensitivity Check

Table 2 presents the Spearman correlation (ρ) between explanations generated for the predicted and alternative target classes. Lower ρ values indicate higher *contrastivity*, that is the explanations are sensitive to the decision direction and capture class-specific evidence. Higher ρ values indicate explanations are largely invariant to the predicted class and may instead reflect input priors. Thus, a target-aware explanation method should ideally demonstrate lower correlations between class-specific explanations.

Across both data settings, XAI methods leveraging direct gradient signals (i.e., Input×Gradient, Integrated Gradients, DeepLIFT) consistently exhibit the lowest correlations that indicates stronger target sensitivity. These methods directly compute attributions from signed gradient information, making them inherently responsive to the change in decision boundary. In contrast, *noise-aggregated methods* such as SmoothGrad and VarGrad, and Guided Backpropagation, show the highest correlations (often > 0.85), which demonstrates limited adaptation to target label changes. Especially for guided backpropagation, since it selectively propagate only positive activations through ReLU masks, the crucial sign information for distinguishing between opposing class gradients is lost. Therefore, it produces visually appealing but target-insensitive maps. Grad-CAM-based methods remain intermediate, likely due to their spatially coarse but class-conditioned activation weighting, producing moderate contrastivity across most conditions.

When comparing *pre-processing operations*, HT and AM consistently enhance contrastivity, resulting in lower ρ values across nearly all methods. This suggests that removing non-lung structures and restricting the model's attention to

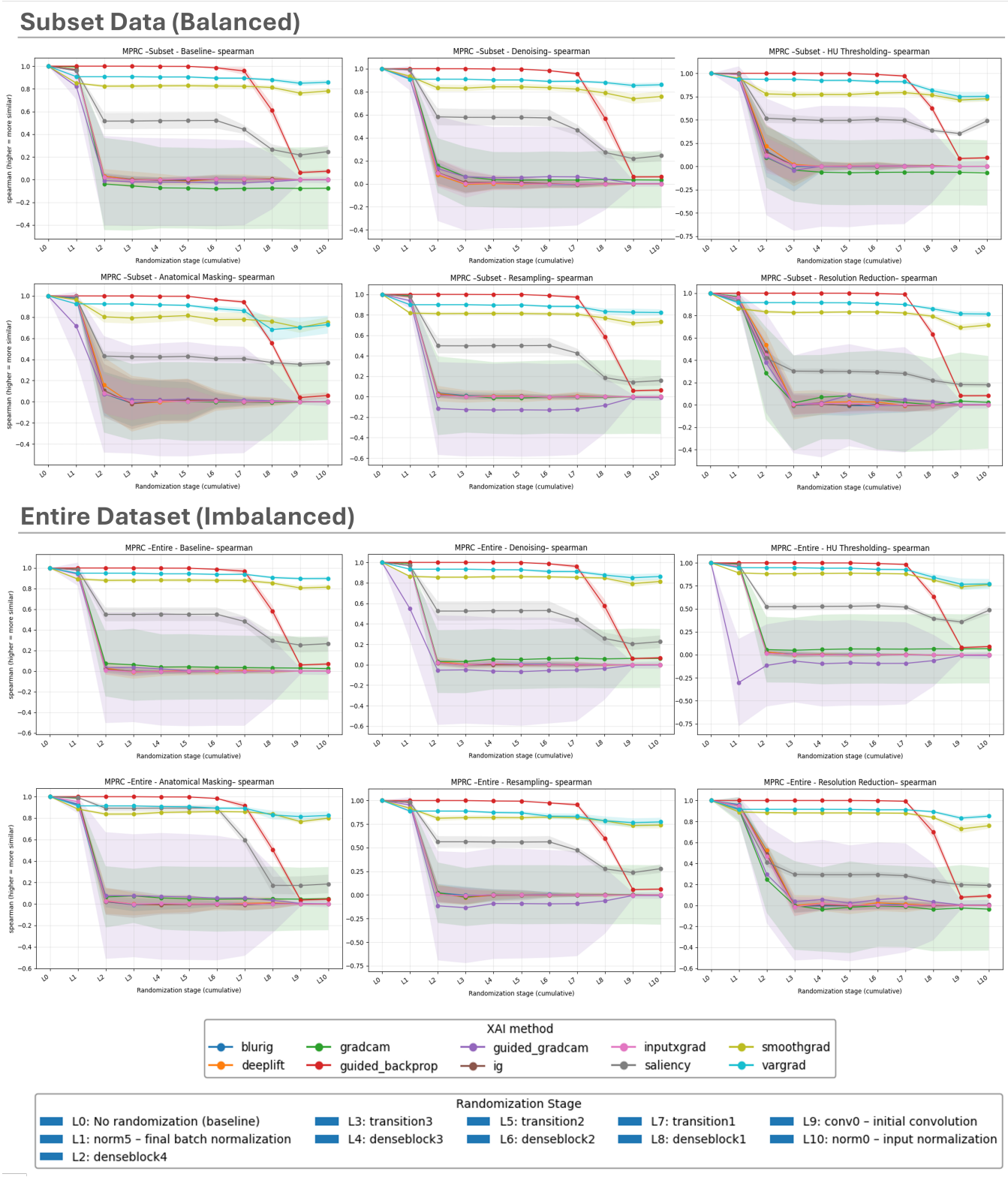


Figure 3: MPRC Results over all XAI methods for Spearman correlation coefficient.

relevant anatomical regions strengthens the discriminative focus of explanations. Conversely, RS and RR tend to elevate ρ, implying that spatial downsampling or reduced voxel granularity smooths local gradients and diminishes class-dependent feature contrast. DN shows mixed effects, slightly improving contrastivity for some gradient-based methods but reducing it for noise-averaged approaches, likely because intensity homogenization reduces sensitivity to small local changes.

Finally, comparing the balanced subset dataset and the imbalanced entire dataset reveals that class imbalance can dampen target sensitivity by biasing the explanations towards the dominant class. In the balanced subset dataset, stronger contrastivity patterns emerge with clearer separation among XAI methods, whereas in the entire dataset, high correlations persist across several methods—reflecting reduced adaptability under skewed data distribution. Overall, methods that rely on raw gradient signals capture class-specific evidence more effectively, while guided and smoothed variants prioritize visual stability over discriminative clarity. These findings confirm that target sensitivity is a valuable diagnostic for distinguishing truly class-aware explanation methods from those dominated by input-driven or architecture-induced priors.

C_3-Completeness: Incremental Patch Deletion Check

Table 3 summarizes results of the Incremental Patch Deletion Check, which evaluates each explanation's *completeness*, meaning how well the identified salient regions capture evidence critical to the model's decision. At each step, top-attributed 3D patches were replaced with their local average intensity, and the resulting confidence drop is quantified through the Log-Odds Ratio (LOR). The Area Under the Curve (AUC) of the LOR–fraction plot measures the degradation in confidence: lower AUC indicates higher completeness, meaning the method successfully highlights features most responsible for the model's prediction. Additionally, a random deletion baseline was included to ensure that confidence degradation originates from meaningful explanation regions rather than arbitrary patch removal.

BIG consistently achieves the lowest AUC values across nearly all pre-processing operations and dataset settings, indicating that it most effectively captures regions truly responsible for the model's predictions. Following BIG, *IxG* and *IG* also exhibit strong completeness, suggesting that direct gradient-based signal propagation remains the most effective strategy for faithfully identifying decision-critical features in 3D volumes. In contrast, *GB* and **GGC** consistently produce the highest AUC values—often even worse than the random deletion baseline—indicating that their highlighted regions contribute minimally to the model's actual confidence. This behavior likely arises from their gradient clipping and positivity filtering mechanisms, which yield visually sharp maps but detach attributions from the underlying causal structure of the model.

Examining *effect of pre-processing operations*, HT and AM markedly improve completeness, leading to the lowest AUCs across both balanced and imbalanced datasets. By isolating lung structures and removing irrelevant background anatomy, these steps enhance the correspondence between model saliency and diagnostic regions. In contrast, RR and RS degrade completeness, reflected by higher AUCs and weaker confidence drops, likely due to the loss of structural granularity and blurred boundary features within the coarser voxel grids. DN shows moderate improvements for attribution-based methods but inconsistent effects for localization approaches, suggesting that fine-grained intensity cues contribute meaningfully to model decision logic.

Finally, comparing the *balanced subset* and the *entire imbalanced dataset* reveals that class imbalance can hinder completeness, especially for methods that rely on averaged or smoothed gradients. The balanced subset dataset yields more pronounced and stable confidence drops, implying that equal representation improves the model's ability to rely on class-relevant regions rather than distributional priors. Overall, attribution-based methods emerge as the most complete and faithful to the model's internal reasoning, while smoothing and guided variants prioritize visual plausibility at the expense of causal fidelity. These findings reinforce the fact that completeness is a strong indicator of an explanation's alignment with the model's true decision boundary, particularly for 3D medical imaging.

Conclusion and Future Directions

In this study, we introduced *XIME^{3}D*, the first systematic framework for evaluating AI explainability for 3D medical data. By jointly analyzing ten post-hoc attribution methods across five clinical pre-processing operations and two dataset distributions, the framework provides a comprehensive view of explanation reliability from input, model, and output perspectives.

Across experiments, gradient-based methods, particularly *GC*, *IG*, and *BIG*, showed the strongest correctness, with sharp correlation declines following MPRC. In contrast, smoothing and noise-aggregation approaches (*SG*, *VG*) maintained high stability but limited sensitivity to model perturbations.

Pre-processing operations substantially modulated explanation quality. *HT* and *AM* consistently improved correctness and contrastivity by restricting attention relevant regions, while coarse operations such as *RS* and *RR* degraded both completeness and sensitivity. Denoising had negligible or method-dependent effects.

In future work, we plan to perform qualitative and expert-based evaluations to complement these quantitative findings and to fully capture explanation usability in clinical workflows that advance the integration of trustworthy 3D XAI in medical imaging.

Acknowledgments

This work has been done in the context of the ITEA 3, Privacy preserving cross-organizational data analysis in the healthcare sector (Secure e-Health) project.

Table 2: Target Sensitivity Check (C_2–Contrastivity) results across pre-processing operations and XAI methods for both Subset and Entire datasets. Values represent the mean Spearman correlation (ρ) between predicted and alternative class explanations. Lower ρ indicates higher contrastivity (i.e., greater difference between explanations of different targets), while higher ρ indicates lower contrastivity (i.e., explanations are less sensitive to the predicted class). The **green cells** denote the *Top 3 most contrastive methods (lowest ρ)*, and the **red cells** denote the *Bottom 3 least contrastive methods (highest ρ)* for each pre-processing operation.

Dataset	PPO	BIG	IxG	IG	DL	SG	SL	VG	GGC	GB	GC
Subset	Baseline	-0.024	-0.019	0.008	**-0.031**	**0.804**	0.527	**0.907**	**-0.244**	**1.000**	**-0.319**
	DN	**-0.833**	**-0.840**	-0.812	-0.691	**0.924**	0.865	**0.907**	**-0.871**	**1.000**	-0.803
	HT	**-0.965**	-0.912	-0.923	-0.942	**0.945**	**0.939**	0.928	**-0.992**	**0.999**	-0.929
	AM	-0.797	-0.798	**-0.901**	**-0.911**	**0.958**	0.747	0.924	**-0.964**	**0.999**	-0.737
	RS	**-0.491**	**-0.481**	-0.338	-0.257	**0.808**	0.592	**0.899**	**-0.584**	**1.000**	-0.363
	RR	**-0.162**	**-0.113**	-0.053	**-0.144**	**0.838**	0.320	**0.916**	0.021	**1.000**	0.027
Entire	Baseline	**-0.508**	**-0.542**	-0.380	-0.358	**0.874**	0.678	**0.949**	**-0.689**	**1.000**	-0.200
	DN	-0.180	**-0.220**	-0.037	-0.082	**0.857**	0.558	**0.933**	**-0.626**	**1.000**	**-0.357**
	HT	**-0.621**	**-0.585**	-0.523	-0.582	**0.888**	0.688	**0.948**	**-0.671**	**1.000**	-0.250
	AM	-0.378	-0.315	-0.550	**-0.649**	**0.871**	**0.913**	**0.913**	**-0.828**	**1.000**	**-0.645**
	RS	**-0.677**	**-0.679**	-0.534	-0.415	**0.887**	0.746	**0.887**	**-0.594**	**1.000**	-0.345
	RR	**-0.267**	**-0.225**	-0.133	**-0.214**	**0.884**	0.326	**0.916**	0.105	**1.000**	0.015

Lower ρ = Higher Contrastivity (Top 3) **Higher ρ = Lower Contrastivity (Bottom 3)**

Table 3: Incremental Patch Deletion Check (C_3-Completeness) results across pre-processing operations and XAI methods. AUC values computed from Log-Odds Ratio (LOR) curves, which quantify the change in model confidence as top-attributed patches are progressively removed. Lower AUC indicates stronger completeness (larger confidence drop after salient region deletion). The green cells represent the *Top 3 strongest-performing methods (lowest AUC values)*, while the red cells represent the *Bottom 3 weakest-performing methods (highest AUC values)* for each pre-processing operation.

Dataset	PPO	BIG	IxG	IG	DL	SG	SL	VG	GGC	GB	GC	Rand
Subset	Baseline	**-0.035**	**0.070**	**0.136**	0.188	0.273	0.277	0.295	0.306	**0.308**	**0.313**	**0.327**
	DN	**-1.693**	**-0.373**	**-0.193**	0.246	0.146	0.167	0.240	**1.318**	**0.806**	-0.178	**0.617**
	HT	**-8.566**	-5.248	**-7.194**	**-8.333**	2.890	2.773	3.013	**4.968**	**5.208**	3.580	**4.783**
	AM	**-2.992**	**-1.346**	**-1.105**	-0.712	1.161	1.131	1.362	**1.562**	**1.716**	0.425	**1.878**
	RS	**4.787**	**6.152**	**6.108**	6.736	6.648	6.691	6.749	**7.148**	6.834	**7.061**	**7.131**
	RR	**-0.023**	**0.209**	**0.213**	**0.494**	0.277	0.218	0.282	0.313	**0.617**	**0.459**	0.266
Entire	Baseline	**0.391**	**0.765**	**0.876**	1.157	1.206	**1.243**	1.237	1.238	1.234	**1.248**	**1.347**
	DN	**0.122**	**0.328**	**0.425**	0.515	**0.583**	0.534	**0.606**	0.527	0.543	0.538	**0.765**
	HT	**-2.208**	-1.373	**-1.893**	**-2.051**	-0.930	-1.064	-0.828	**-0.239**	**-0.306**	-0.322	**-0.207**
	AM	**-0.043**	**0.377**	0.638	0.578	0.512	**0.452**	0.538	**1.170**	**1.146**	0.538	**0.999**
	RS	**-12.697**	**-6.413**	**-7.709**	-2.145	1.113	1.138	**1.236**	0.711	**1.349**	-0.5897	**1.149**
	RR	**-0.435**	-0.099	-0.111	-0.102	**-0.139**	-0.095	**-0.115**	0.045	**0.057**	**0.060**	**0.107**

Lower AUC = Higher Completeness (Top 3) **Higher AUC = Lower Completeness (Bottom 3)**

References

Adebayo, J.; Gilmer, J.; Muelly, M.; Goodfellow, I.; Hardt, M.; and Kim, B. 2018. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31.

Aggarwal, R.; Sounderajah, V.; Martin, G.; Ting, D. S.; Karthikesalingam, A.; King, D.; Ashrafian, H.; and Darzi, A. 2021. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. *NPJ digital medicine*, 4(1): 65.

Bhati, D.; Neha, F.; and Amiruzzaman, M. 2024. A survey on explainable artificial intelligence (xai) techniques for visualizing deep learning models in medical imaging. *Journal of Imaging*, 10(10): 239.

Blake, N.; Chockler, H.; Kelly, D. A.; Pena, S. C.; and Chanchal, A. 2023. MRxaI: Black-Box Explainability for Image Classifiers in a Medical Setting. *CoRR*, abs/2311.14471.

Brima, Y.; and Atemkeng, M. 2024. Saliency-driven explainable deep learning in medical imaging: bridging visual explainability and statistical quantitative analysis. *BioData mining*, 17(1): 18.

Cardoso, M. J.; Li, W.; Brown, R.; Ma, N.; Kerfoot, E.; Wang, Y.; Murrey, B.; Myronenko, A.; Zhao, C.; Yang, D.; et al. 2022. Monai: An open-source framework for deep learning in healthcare. *arXiv preprint arXiv:2211.02701*.

Chung, M. H.; Gil, B. M.; Kwon, S. S.; Park, H.-i.; Song, S. W.; Jung, N. Y.; and Yoo, W. J. 2018. Computed tomographic thoracic morphologic indices in normal subjects and patients with chronic obstructive pulmonary disease: Comparison with spiral CT densitometry and pulmonary function tests. *European Journal of Radiology*, 100: 147–153.

Colombi, D.; Bodini, F. C.; Petrini, M.; Maffi, G.; Morelli, N.; Milanese, G.; Silva, M.; Sverzellati, N.; and Michieletti, E. 2020. Well-aerated lung on admitting chest CT to predict adverse outcome in COVID-19 pneumonia. *Radiology*, 296(2): E86–E96.

Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708.

Isensee, F.; Jaeger, P. F.; Kohl, S. A.; Petersen, J.; and Maier-Hein, K. H. 2021. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2): 203–211.

Klein, L.; Lüth, C.; Schlegel, U.; Bungert, T.; El-Assady, M.; and Jäger, P. 2024. Navigating the Maze of Explainable AI: A Systematic Approach to Evaluating Methods and Metrics. In Globerson, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J.; and Zhang, C., eds., *Advances in Neural Information Processing Systems*, volume 37, 67106–67146. Curran Associates, Inc.

Kokhlikyan, N.; Miglani, V.; Martin, M.; Wang, E.; Alsallakh, B.; Reynolds, J.; Melnikov, A.; Kliushkina, N.; Araya, C.; Yan, S.; et al. 2020. Captum: A unified and generic model interpretability library for pytorch. *arXiv preprint arXiv:2009.07896*.

Masoudi, S.; Harmon, S. A.; Mehralivand, S.; Walker, S. M.; Raviprakash, H.; Bagci, U.; Choyke, P. L.; and Turkbey, B. 2021. Quick guide on radiology image pre-processing for deep learning applications in prostate cancer research. *Journal of Medical Imaging*, 8(1): 010901–010901.

Metsch, J. M.; and Hauschild, A.-C. 2025. BenchXAI: Comprehensive benchmarking of post-hoc explainable AI methods on multi-modal biomedical data. *Computers in Biology and Medicine*, 191: 110124.

Morozov, S. P.; Andreychenko, A. E.; Pavlov, N. A.; Vladzymyrskyy, A.; Ledikhova, N. V.; Gombolevskiy, V. A.; Blokhin, I. A.; Gelezhe, P. B.; Gonchar, A.; and Chernina, V. Y. 2020. Mosmeddata: Chest ct scans with covid-19 related findings dataset. *arXiv preprint arXiv:2005.06465*.

Müller-Franzes, G.; Khader, F.; Siepmann, R.; Han, T.; Kather, J. N.; Nebelung, S.; and Truhn, D. 2025. Medical slice transformer for improved diagnosis and explainability on 3D medical images with DINOv2. *Scientific Reports*, 15(1): 23979.

Nauta, M.; Trienes, J.; Pathak, S.; Nguyen, E.; Peters, M.; Schmitt, Y.; Schlötterer, J.; Van Keulen, M.; and Seifert, C. 2023. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *ACM Computing Surveys*, 55(13s): 1–42.

Nielsen, I. E.; Ramachandran, R. P.; Bouaynaya, N.; Fathallah-Shaykh, H. M.; and Rasool, G. 2023. EvalAttAI: A Holistic Approach to Evaluating Attribution Maps in Robust and Non-Robust Models. *IEEE Access*, 11: 82556–82569.

Ohkubo, H.; Kanemitsu, Y.; Uemura, T.; Takakuwa, O.; Takemura, M.; Maeno, K.; Ito, Y.; Oguri, T.; Kazawa, N.; Mikami, R.; et al. 2016. Normal lung quantification in usual interstitial pneumonia pattern: the impact of threshold-based volumetric CT analysis for the staging of idiopathic pulmonary fibrosis. *PloS one*, 11(3): e0152505.

Perona, P.; and Malik, J. 2002. Scale-space and edge detection using anisotropic diffusion. *IEEE Transactions on pattern analysis and machine intelligence*, 12(7): 629–639.

Richter, L.; Boustati, A.; Nüsken, N.; Ruiz, F.; and Akyildiz, O. D. 2020. Vargrad: a low-variance gradient estimator for variational inference. *Advances in Neural Information Processing Systems*, 33: 13481–13492.

Rister, B.; Yi, D.; Shivakumar, K.; Nobashi, T.; and Rubin, D. L. 2020. CT-ORG, a new dataset for multiple organ segmentation in computed tomography. *Scientific Data*, 7(1): 381.

Sabottke, C. F.; and Spieler, B. M. 2020. The effect of image resolution on deep learning in radiography. *Radiology: Artificial Intelligence*, 2(1): e190015.

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

Shrikumar, A.; Greenside, P.; and Kundaje, A. 2017. Learning important features through propagating activation dif-

ferences. In *International conference on machine learning*, 3145–3153. PMlR.

Shrikumar, A.; Greenside, P.; Shcherbina, A.; and Kundaje, A. 2016. Not just a black box: Learning important features through propagating activation differences. *arXiv preprint arXiv:1605.01713*.

Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Singh, S. P.; Wang, L.; Gupta, S.; Goli, H.; Padmanabhan, P.; and Gulyás, B. 2020. 3D deep learning on medical images: a review. *Sensors*, 20(18): 5097.

Smilkov, D.; Thorat, N.; Kim, B.; Viégas, F.; and Wattenberg, M. 2017. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*.

Song, B.; Yoshida, S.; and Initiative, A. D. N. 2024. Explainability of three-dimensional convolutional neural networks for functional magnetic resonance imaging of Alzheimer's disease classification based on gradient-weighted class activation mapping. *Plos one*, 19(5): e0303278.

Springenberg, J. T.; Dosovitskiy, A.; Brox, T.; and Riedmiller, M. 2014. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*.

Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, 3319–3328. PMLR.

Talaat, F. M.; Gamel, S. A.; El-Balka, R. M.; Shehata, M.; and ZainEldin, H. 2024. Grad-CAM enabled breast cancer classification with a 3D inception-ResNet V2: Empowering radiologists with explainable insights. *Cancers*, 16(21): 3668.

Thévenaz, P.; Blu, T.; and Unser, M. 2002. Interpolation revisited [medical images application]. *IEEE Transactions on medical imaging*, 19(7): 739–758.

Van der Velden, B. H.; Kuijf, H. J.; Gilhuijs, K. G.; and Viergever, M. A. 2022. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. *Medical image analysis*, 79: 102470.

Van Kolfschooten, H.; and Van Oirschot, J. 2024. The EU artificial intelligence act (2024): implications for healthcare. *Health Policy*, 149: 105152.

Wolterink, J. M.; Leiner, T.; Viergever, M. A.; and Išgum, I. 2017. Generative adversarial networks for noise reduction in low-dose CT. *IEEE transactions on medical imaging*, 36(12): 2536–2545.

Xu, S.; Venugopalan, S.; and Sundararajan, M. 2020a. Attribution in Scale and Space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xu, S.; Venugopalan, S.; and Sundararajan, M. 2020b. Attribution in scale and space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9680–9689.

Yang, C.; Rangarajan, A.; and Ranka, S. 2018. Visual explanations from deep 3D convolutional neural networks for Alzheimer's disease classification. In *AMIA annual symposium proceedings*, volume 2018, 1571.

Zhou, T.; Ye, X.; Lu, H.; Zheng, X.; Qiu, S.; and Liu, Y. 2022. Dense convolutional network and its application in medical image analysis. *BioMed Research International*, 2022(1): 2384830.