Navigating the Maze of Explainable AI: A Systematic Approach to Evaluating Methods and Metrics

Anonymous authors
Paper under double-blind review

Abstract

Explainable AI (XAI) is a rapidly growing domain with a myriad of proposed methods as well as metrics aiming to evaluate their efficacy. However, current literature is often of limited scope, examining only a handful of XAI methods and ignoring underlying design parameters for performance, such as the model architecture or the nature of input data. Moreover, they often rely on one or a few metrics, neglecting thorough validation and increasing the risk of selection bias. These shortcomings leave practitioners confused about which method to choose for their problem. In response, we introduce LATEC, a large-scale benchmark that critically evaluates 17 prominent XAI methods using 20 distinct metrics. We systematically incorporate vital design parameters like varied architectures and diverse input modalities, resulting in 7,560 examined combinations. Through LATEC, we first showcase the high risk of conflicting metrics leading to unreliable rankings, and propose a robust evaluation scheme. Further, we comprehensively evaluate various XAI methods to assist practitioners in selecting appropriate methods aligning with their needs. Curiously, the emerging top-performing method, Expected Gradients, has not been examined in relevant related studies before. LATEC reinforces its role in future XAI research by publicly releasing all auxiliary data, including model weights, over 326k saliency maps, and 378k metric scores as a dataset. The benchmark is hosted at: `https://github.com/kjdhfg/LATEC`.

1 Introduction

Explainable AI (XAI) methods have become essential tools in numerous domains, allowing for a better understanding of complex machine learning decisions. The most prevalent XAI methods originate from the domain of saliency maps (Simonyan et al., 2013). As the diversity and abundance of proposed saliency XAI methods expand alongside their growing popularity, ensuring their reliability becomes paramount (Adebayo et al., 2018). Given that there is no clear "ground truth" for individual explanations (e.g., discussed in Adebayo et al. (2020)), the trustworthiness of XAI methods is typically determined by examining three key criteria: their accuracy in reflecting a model's reasoning ("faithfulness") (Bach et al., 2015; Samek et al., 2017), their stability under small changes ("robustness") (Yeh et al., 2019; Alvarez Melis & Jaakkola, 2018), and the understandability of their explanations ("complexity") (Chalasani et al., 2020; Bhatt et al., 2021). Beyond qualitative assessments such as in Doshi-Velez & Kim (2017); Ribeiro et al. (2016); Shrikumar et al. (2017), which can be influenced by human biases and don't always scale well (as shown by Wang et al. (2019); Rosenfeld (2021)), a wide array of metrics have been introduced to quantitatively evaluate XAI methods based on these three criteria. These metrics are deployed in several studies (see Table 1) to determine *"What XAI method should I (not) use for my problem?"*.

However, this current work in validating XAI methods has two major shortcomings:

Shortcoming 1: Gaps and inconsistencies in XAI evaluation. Many studies restrict their analyses to a limited set of design parameters such as modalities, (toy-)datasets, model architectures, methods, and metrics, which all directly impact the performance of XAI methods (we define the first three parameters as *underlying* design parameters, as they directly influence the XAI method). Table 1 demonstrates this fragmented landscape specifically for the domain of computer vision, including discrepancies found across

Studies: \ XAI Method:	Attribution														Attention		
	OC	LIME	KS	VG	IxG	GB	GC	SC	C+	IG	EG	DL	DLS	LRP	RA	RoA	LA
Adebayo et al. (2018) ($n_A = 5$ (8), $n_E = 2$)				F	F	F	F			F							
Nie et al. (2019) ($n_A = 2$ (3), $n_E = 0$)				F		F											
Kindermans et al. (2019) ($n_A = 4$ (8), $n_E = 1$)					F					F				F			
Ghorbani et al. (2019) ($n_A = 3$, $n_E = 3$)				R						R		R					
Hooker et al. (2019) ($n_A = 3$ (12), $n_E = 1$)				F		F				F							
Yang & Kim (2019) ($n_A = 6$ (9), $n_E = 3$)				F	F	F	F			F							
Yeh et al. (2019) ($n_A = 4$ (6), $n_E = 2$)			R	R		R				R							
Nguyen & Martinez (2020) ($n_A = 3$, $n_E = 4$)				C	C					C							
Chefer et al. (2020) ($n_A = 5$, $n_E = 1$)							F							F	F	F	F
Bhatt et al. (2020) ($n_A = 6$, $n_E = 3$)			F/R	F/R	F/R					F/R		F/R					
Arun et al. (2021) ($n_A = 5$ (8), $n_E = 4$)				F		F	F			F							
Singh et al. (2021) ($n_A = 9$ (12), $n_E = 0$)	C			C	C	C				C		C	C	C			
Kakogeorgiou et al. (2021) ($n_A = 8$ (12), $n_E = 3$)	F/R/C	F/R/C		F/R/C	F/R/C	F/R/C	F/R/C			F/R/C		F/R/C					
Dombrowski et al. (2022) ($n_A = 5$, $n_E = 1$)				R	R	R				R				R			
Arras et al. (2022) ($n_A = 6$ (12), $n_E = 2$)				F	F	F	F			F				F			
Hesse et al. (2023) ($n_A = 11$, $n_E = 3$)		F			F		F			F						F	F
Ours * ($n_A = 17$, $n_E = 20$)	F/R/C	F/R/C	F/R/C	F/R/C	F/R/C	F/R/C	F/R/C	F/R/C	F/R/C	F/R/C	F/R/C	F/R/C	F/R/C	F/R/C	F/R/C	F/R/C	F/R/C

Evaluated Criteria: **F** Faithfulness, **R** Robustness, **C** Complexity

* Results for image modality

Relative assessment to other XAI methods in study: Positive Neutral Negative

Table 1: Showing gaps and inconsistencies between 16 relevant related studies evaluating XAI methods. Colors coincide with their ranking inside the study depending on the evaluation criteria. n_A: Amount of distinct XAI methods. n_E: Number of evaluation metrics. If $n_E = 0$, the study was conducted qualitatively.

studies, with some methods, such as GradCAM (GC) (Selvaraju et al., 2017), receiving contradictory assessments depending on the evaluation setup. As a consequence, our current understanding of XAI performance is limited, making it challenging for practitioners to determine a reliable XAI method for their specific use case.

Shortcoming 2: Individual XAI metrics lack trustworthiness. Recently, numerous metrics have been proposed to approximate evaluation criteria. However, a metric is always only a proxy for a criterion, reflecting the diversity of perspectives, and they are distinguishable solely in their mathematical formulation. Generally, studies apply one or two metrics to address a criterion (see n_E in Table 2 for the total amount of metrics used for evaluation in each study). Arguably, this is not a reliable measure of success, as these limited subsets can lead to selection bias, overfitting to one metric and one perspective on the problem. Consequently, rankings and scores of current work lack trustworthiness.

In response to these shortcomings, we developed LATEC: the first comprehensive benchmark tailored for large-scale attribution & attention evaluation in computer vision. LATEC encompasses 17 of the most widely-used saliency XAI methods, including attention-based methods, and evaluates them using 20 distinct metrics (see Figure 1). Notably, LATEC integrates a variety of model architectures, and, to extend the evaluation spectrum beyond traditional 2D images, we included point cloud and volume data, adapting XAI methods and metrics as necessary to suit these modalities. In total, LATEC assesses 7,560 unique combinations.

As LATEC systematically incorporates all vital underlying design parameters affecting XAI methods, we can quantify their effect on XAI methods and perform a more trustworthy and generalizing benchmark, answering Shortcoming 1. However, we can also use LATEC for metric analysis (also called meta-evaluation) to, a priori of the evaluation of the methods, quantitatively validate ranking behavior of metrics, and determine a robust evaluation mechanism, answering Shortcoming 2. Moreover, in support of future research, we've made all intermediate data, including 326,790 saliency maps and 378,000 evaluation scores, as well as the benchmark publicly accessible.

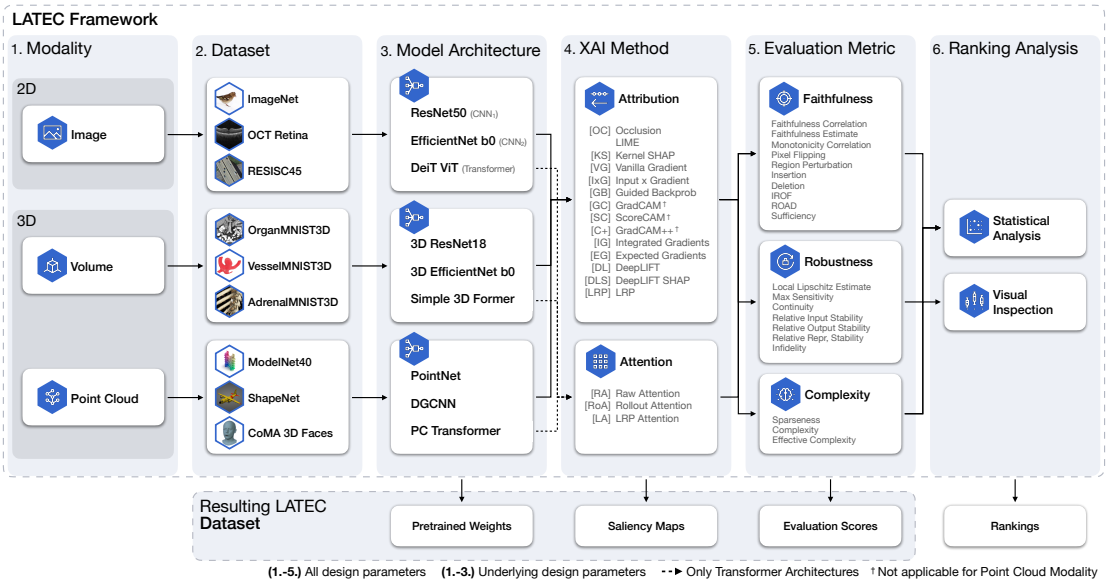


Figure 1: Structure and design parameters of the LATEC framework with the output data for the LATEC dataset of each stage.

2 The LATEC benchmark

The LATEC benchmark includes a framework and a dataset with the rankings as the final output. The framework allows for for diverse large-scale studies, structuring the experiments in six stages (see Figure 1), and the dataset provides reference data for evaluation and exploration. As the benchmark is easily extendable and leverages the high-quality dataset for standardized evaluation, it also serves as a foundation for future benchmarking of new XAI methods and metrics.

Utilized datasets For the image modality, we use ImageNet (IMN) (Deng et al., 2009), UCSD OCT retina (OCT) (Kermany et al., 2018) and RESISC45 (R45) (Cheng et al., 2017), the volume modality the Adrenal-(AMN), Organ-(OMN) and VesselMedMNIST3D (VMN) datasets (Yang et al., 2023), and the point cloud modality the CoMA (CMA) (Ranjan et al., 2018), ModelNet40 (M40) (Wu et al., 2014) and ShapeNet (SHN) (Chang et al., 2015) datasets.

Model architectures On each utilized dataset except IMN, where we take pretrained models, we train three models to achieve the architecture-dependent SOTA performance on the designated test set (if available, see Appendix A for a detailed description of all model trainings and hyperparameters). For the image modality, we use the ResNet50, EfficientNetb0, and DeIT ViT (Touvron et al., 2022) architectures, for the volume modality the 3D ResNet18, 3D EfficientNetb0, and Simple3DFormer (Wang et al., 2022) architectures, and for the point cloud modality the PointNet, DGCNN and PC Transformer (Guo et al., 2021) architectures. The first two architectures are always CNNs (CNN_1 and CNN_2), and the third is a Transformer.

XAI methods In total, we include 17 XAI methods, 14 attribution methods: Occlusion (OC) (Zeiler & Fergus, 2013), LIME (on feature masks) (Ribeiro et al., 2016), Kernel SHAP (KS, on feature masks) (Lundberg & Lee, 2017), Vanilla Gradient (VG) (Simonyan et al., 2013), Input x Gradient (IxG) (Shrikumar et al., 2017), Guided Backprob (GB) (Springenberg et al., 2015), GC, ScoreCAM (SC) (Wang et al., 2020), GradCAM++ (C+) (Chattopadhay et al., 2018), Integrated Gradients (IG) (Sundararajan et al., 2017), Expected Gradients (EG, also called Gradient SHAP) (Erion et al., 2020), DeepLIFT (DL) (Shrikumar et al., 2017), DeepLIFT SHAP (DLS) (Lundberg & Lee, 2017), LRP (with ϵ-,γ- and 0^+-rules depending on the model architecture) (Binder et al., 2016), and three attention methods: Raw Attention (RA) (Dosovitskiy et al., 2021), Rollout Attention (RoA) (Abnar & Zuidema, 2020) and LRP Attention (LA) (Chefer et al., 2021). While the attribution methods are applied to all model architectures, the attention methods can only

be applied to the Transformer-based architectures. For comparison reasons, we only consider the original methods without adaptations, as several other works (Hooker et al., 2019; Yang & Kim, 2019) already showed that advancing methods by VarGrad (Adebayo et al., 2018) or SmoothGrad(-squared) (Smilkov et al., 2017) can, in general, improve results. We qualitatively tuned the XAI hyperparameters per dataset (see Appendix B), as also commonly done in practice, to not bias the quantitative evaluation results (see Appendix C for all hyperparameters). However, we observe that most hyperparameters transfer well over all datasets within a modality.

Evaluation metrics The evaluation metrics are grouped into three criteria: faithfulness (is the explanation following the model behavior?), robustness (is the explanation stable?), and complexity (is the explanation concise and human understandable?). We utilize in total 20 evaluation metrics, of which 10 evaluate faithfulness, seven robustness, and three complexity (see Appendix D for a description of every metric). We tune the hyperparameters per dataset as some depend on dataset properties (see Appendix subsection D.2 for all parameters).

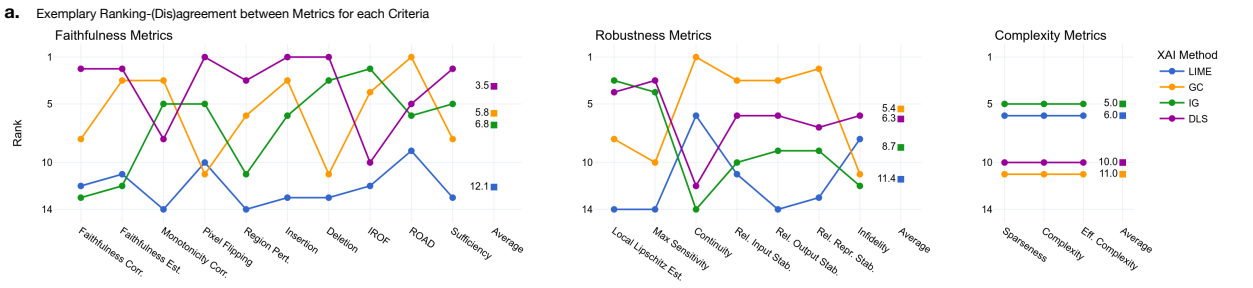
3D adaption While many XAI methods and metrics in LATEC are independent of the input space dimensions, we had to extend many of them to 3D volume and point cloud data, building upon the implementations for image data by Kokhlikyan et al. (2020) and Hedström et al. (2023). We describe the adaption process for all respective XAI methods and metrics in Appendix E and show illustrative saliency maps. All adaptations were tested for their coherency.

Ranking computation As the nominal evaluation scores have no semantic meaning and their scales differ between utilized datasets, we analyze the XAI methods based on their ranking. To this end, we compute the median evaluation score over all observations on the dataset level and rank the methods accordingly, achieving more robust results. See Appendix G for a detailed flow chart of how we get from evaluation scores to rankings and Appendix N for the distribution of all scores before ranking. We utilize these rankings in the subsequent metric analysis and benchmark and refer to Appendix C for all implementation details and Appendix B for more information about the dataset and ranking computation.

3 Metrics analysis

How severe is the risk of metric selection bias in XAI evaluation? In Shortcoming 2, we describe a risk of selection bias due to evaluating one or a few metrics. In our metrics analysis, we first aim to provide empirical evidence for this risk. A first exploratory analysis quickly supports the hypothesis, as we encounter strong ranking disagreement for various combinations of underlying design parameters. Figure 2 (a.) demonstrates the ranking behavior of four selected XAI methods for one selection of underlying design parameters. The line charts show the agreement and disagreement in ranking between metrics, with the average aggregated rank to the right. For faithfulness, we observe high disagreement between metrics in the case of GC and IG and one strong outlier for DLS. However, the rankings of other XAI methods, e.g., LIME in this case, are reasonably stable. The inquiry emerges as to whether the risk of selection bias is generally present in certain combinations of underlying design parameters or is uniformly distributed across them. To this end, we computed the standard deviation between metric rankings mean aggregated for each underlying design parameter (see results for volume and point cloud modalities in Appendix H). Figure 2 (b.) shows that the average standard deviation is generally stable within each evaluation criterion. Thus, we can conclude that there is no single model architecture, modality, or dataset choice that has a substantial effect on the disagreement between metric rankings of all XAI methods.

Now that we can rule out the general influence of underlying design parameters, we quantify how strong the risk of selection bias, i.e., disagreement, actually is and if there is a difference between XAI methods. To this end, we utilize a one-sided Levene's Test (Levene, 1960), testing if the rank-variance of a set of metrics is significant ($\alpha = 0.1$) lower than the variance of a random rank distribution, which can be analytically inferred. We compute this test for every possible combination of design parameters and show the average acceptance rate of the test per XAI method in Figure 2 (c.). By computing the weighted average proportion at the bottom, we indeed observe strong variations between the XAI methods. Specifically for KS, EG, and DLS, in a large majority of cases, metrics agree, while for OC, IxG, SC, and DL only in about $\sim 27\%$ of the



b. Avg. Standard Deviation for Image Model Architectures and Datasets

Evaluation Criteria:	Model Architectures			Utilized CV Datasets		
	ResNet50	EffNetb0	DeiT ViT	IMN	OCT	R45
Faithfulness	3.35	3.31	3.45	3.43	3.25	3.43
Robustness	3.19	3.2	2.96	3.25	3.09	3.01
Complexity	0.43	0.48	0.47	0.38	0.63	0.37
Weighted Average	2.86	2.85	2.83	2.91	2.80	2.82

c. Proportion of smaller Metric Ranking Variance than Random Ranking Variance based on one-sided Levene Test over all Image Datasets and Model Architectures

Attribution Method: / Evaluation Criteria:	OC	LIME	KS	VG	IxG	GB	GC	SC	C+	IG	EG	DL	DLS	LRP	RA	RoA	AL	*Average*
Faithfulness	0.0	0.56	0.78	0.67	0.22	0.22	0.56	0.11	0.56	0.33	1.0	0.22	0.67	0.22	0.17	0.17	0.17	0.44
Robustness	0.33	0.56	1.0	0.56	0.0	0.56	0.44	0.22	0.33	0.11	0.33	0.0	0.89	0.33	0.33	0.5	0.45	0.4
Complexity	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Weighted Average	0.27	0.63	0.89	0.68	0.26	0.46	0.58	0.28	0.55	0.35	0.77	0.26	0.80	0.38	0.35	0.41	0.41	0.51

> 0.7 | < 0.3 | Attention Methods only on Transformer Architectures

Figure 2: **a.** Distribution of ranks based on the different metrics for each criterion. Average rank to the right of each plot. **b.** Average standard deviation per model architectures and utilized datasets. Weighted average is based on the number of metrics per criterion. **c.** Proportion of accepted one-sided Levene-Tests for significant ($\alpha = 0.1$) smaller ranking variance compared to the variance of an entire random ranking. Higher values show higher agreement between metrics. Weighted average is based on the number of metrics per criterion.

cases variance in metric ranking is significantly lower than random ranking. Concluding, our findings reveal that metrics disagree and agree in varying degrees depending on the XAI method.

How can we reliably benchmark and find robust trends within agreeing and disagreeing metrics? We argue that possible disagreement between metrics provides essential insights for XAI evaluation, as it also reflects the diversity of perspectives and implementations of the criteria. Subsequently, it should not be discarded but included in the benchmark. Thus, we propose the simple but effective approach of mean aggregation over several metric ranks to get a robust estimator for each criterion invariant to the bias of individual implementations. In the event of a lack of consensus among metrics, the average rank is expected to converge towards an uninformative mean rank (as the "Average" faithfulness for OC and IG shows in Figure 2 (a.)), and only robust and meaningful trends are detected (such as in the faithfulness case of LIME and DLS). At the same time, we retain information about ambiguity through the standard error of the aggregation.

In scenarios of high disagreement among metrics, potential biases could arise when considering only a limited subset of metrics, a concern we have addressed in contrast to other studies via our diverse pool of metrics and aggregation schema. To show that our pool of metrics and aggregation schema is unbiased from such highly (dis-)similar ranking metric subgroups, we compute the average Euclidean distance (δ) between the ranking of two metrics for the faithfulness and robustness criteria (see Appendix I). While the mean rank distance between two metrics is around $\bar{\delta}_F = 4.45$ and $\bar{\delta}_R = 4.14$ respectively, we observe either substantially higher agreement or disagreement on this aggregation level for 5 outlier pairs out of 40 for faithfulness ($CI_{0.9}(3.1, 5.8)$) and 3 out of 21 for robustness ($CI_{0.9}(1.5, 6.7)$). By first selecting a favorable metric and then pairing it with strongly similar ranking metrics based on this distance matrix, it is in theory possible to selectively pair similar ranking metrics to deliberately skew results towards a favorable outcome (dissimilar metrics for an unfavorable outcome). However, this is only possible for a very small subset of metrics. The likelihood of any such "extremist subgroups" unduly and consistently influencing our large-scale study is small. The closest relative to such a subgroup would be the Relative Stability metrics for robustness. To ensure certainty beyond doubt, ranking independence from subgroup influence could also be ensured by using the computed ranking distance matrices between the metrics for weighted aggregation.

Evaluation Criteria:		Faithfulness ($n_{(A,M)} = 90$)			Robustness ($n_{(A,M)} = 63$)			Complexity ($n_{(A,M)} = 27$)		
	Modality:	Image	Volume	Point Cloud	Image	Volume	Point Cloud	Image	Volume	Point Cloud
Attribution	OC	11.5	9	12	12.5	9.5	12	9	7.5	6.5
	LIME	16.5	13	2	16	15	14	7	9.5	9
	KS	16.5	12	3	17	17	12	12.5	12	10
	VG	13	14	8.5	4	4	6.5	11	14	8
	IxG	9.5	5	5	14.5	11.5	9.5	4	2	3
	GB	7.5	10	1	6.5	8	12	5	6	5
	GC	6	17	-	2.5	13	-	14	9.5	-
	SC	4.5	11	-	10	15	-	10	16.5	-
	C+	11.5	15.5	-	5	11.5	-	15.5	16.5	-
	IG	4.5	2	4	10	4	9.5	3	3.5	4
	EG	1	3	7	2.5	1	6.5	17	7.5	11
	DL	7.5	5	6	12.5	7	8	2	3.5	2
	DLS	2	5	12	6.5	9.5	5	6	5	6.5
	LRP	15	15.5	10	14.5	15	2	1	1	1
Attention	RA	14	8	8.5	1	6	3.5	8	11	12
	RoA	9.5	7	12	8	2	3.5	15.5	13	14
	LA	3	1	14	10	4	1	12.5	15	13

Per modality: Top 1 | Top 2-4 | Bottom 2-4 | Bottom 1

Table 2: Average rank over model architectures, datasets, and all evaluation metrics of the respective criteria for each XAI method and modality (i.e. the rank of OC on image is based on $3 * 3 * 10 = 90$ ranks). Coloring coincides with top and bottom positions as point cloud rankings are of length 14 and all others of length 17.

Encountered pitfalls in XAI metric application. During our analysis we encountered several critical pitfalls of current metric application in XAI, to our knowledge not discussed before. Suspiciously, Figure 2 (c.) indicates almost no disagreement between complexity rankings. By investigating complexity metrics further, we observed that gradient and Deep Taylor Decomposition principle-based methods (IxG, IG, DL, and LRP, see Table 2) are ranked as significantly less complex compared to the CAM and attention methods. In our opinion, this observation is counterintuitive when comparing the complexity rankings to the saliency maps in Appendix B, based on which we would classify CAM and attention methods as more clearly arranged and less noisy. While all three complexity metrics were also explicitly proposed for image data, we notice that they all treat each pixel, voxel, or point independently of each other, ignoring locality and favoring methods that attribute to the smallest set of single pixels. As this approach possibly transfers to low dimensional images such as MNIST (Lecun et al., 1998) or CIFAR-10 (Krizhevsky, 2009), the image datasets the three metrics were originally presented on, we hypothesize that it may not be effective with higher-dimensional inputs as observed in our study. Consequently, it is expected that techniques such as LRP would be highly regarded due to their emphasis on filtering the significance of individual pixels, in contrast to CAM methods (GC, SC, C+) that attribute importance to broader local regions. Whether these XAI methods are less complex and more human-understandable on computer vision modalities is debatable and *subsequent complexity evaluation results should be interpreted with caution*. While all metrics are theoretically very well founded, we encountered several smaller pitfalls during our evaluation, which can be found in Appendix K. However, even if a metric fails for a specific combination of design parameters, we still include them in the benchmark, as the determination of such failure cases is pivotal to future work in metric analysis.

4 XAI benchmark results

Upon establishing a robust framework for evaluation, we address Shortcoming 1 by decomposing it into four essential sub-questions. The first two questions aim to bridge gaps in research due to limited design parameters, while the last two focus on resolving inconsistencies by identifying key trends in XAI method performance. Every sub-question includes a set of **four main findings (1-4)**, each of which consists of an *observation* followed by a *recommendation*. For more fine-grained assistance in selecting the most appropriate XAI methods, we refer to detailed ranking tables (see also Appendix J) and the LATEC dataset.

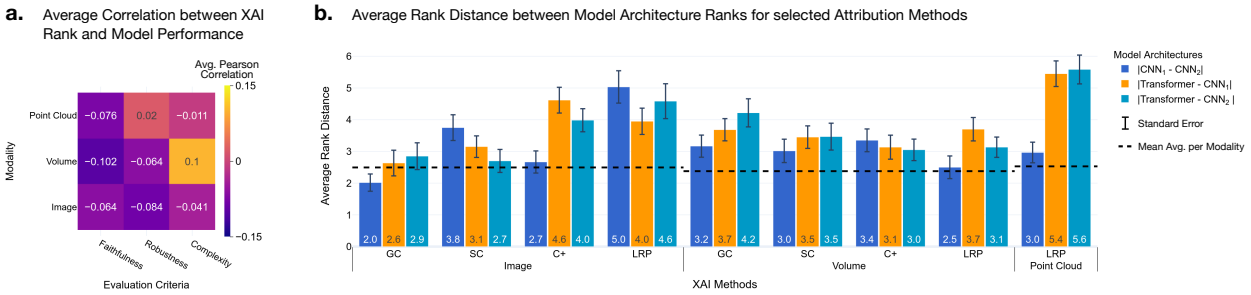


Figure 3: **a.** Average distance between ranks on different model architectures for CAM methods and LRP, as their distance is well above the average. **b.** Correlation between ranks of XAI methods and model F1 test scores, aggregated for each modality/criterion combination.

How is XAI method performance dependent on underlying modalities and datasets? Although, modalities in computer vision share attributes like locality, spatial structure, and associated feature descriptors, they differ in aspects such as dimensionality and representation forms (e.g., grid vs. points). If these differences indeed influence XAI performance, the resulting inconsistencies are largely overlooked in the current literature. To quantify such an effect on XAI methods, we average ranks of XAI methods within modalities and evaluation criteria in Table 2, revealing new patterns in the ranking of various methods. **(1)** Between modalities, only a few methods achieve generalizing results over all modalities, even for one criterion. We generally recommend that users should make their XAI method selection depending on the modality. **(2)** The extended table in Appendix J shows that ranking disparities between datasets within individual modalities are minimal. This suggests that a method selected for one dataset can transfer well to others if dimensionality and characteristics are not too distinct. **(3)** For some methods, we observe a high correlation between performance on the image and volume modality. In particular, both linear surrogate methods (LIME, KS) underperform on image and volume compared to the lower dimensional point cloud modality. On these modalities, their performance also strongly depends on the suitability of the feature mask computed via a grid or super-pixels, which is very time-consuming to fine-tune for single observations. Concluding, we advise against using them for high-dimensional and complex relationships. **(4)** CAM methods always perform better on image than on volume data, which we attribute to less accurate latent model representations and subsequent up-sampling in 3D compared to 2D space, subsequently not recommending them for volume data. *In summary, performance relies generally on the modality but not on the dataset.*

How does XAI method performance depend on the underlying classifier? We investigate this frequently neglected design parameter by focusing on the aspects of model performance and architecture type. Model performance evaluation might be underemphasized in current research if, even after adequate training, XAI performance remains dependent on model performance. Thus, we compute in Figure 3 (a.) the Pearson correlation between the aggregated ranks per criteria and the F1 test scores of the models per dataset. **(1)** Even for well-trained models, we observe a small trend that a better F1 score correlates with better faithfulness and robustness on image and volume modalities. The observed weak correlation raises questions about whether optimizing model performance to a high degree of precision would yield significant alterations in the saliency map. Thus, in tasks where several models with high performance but different properties are available for XAI analysis (e.g., for knowledge/scientific discovery), we would additionally recommend taking into account other criteria, such as the use of attention methods or operational limitations.

Recent research around Vision-Transformers has shown distinct differences in their learning dynamics (Raghu et al., 2022; Park & Kim, 2022), robustness (Zhou et al., 2022) or latent representations (Wang et al., 2023) compared to classical CNNs. As many of their mechanisms (e.g., global processing, lack of inductive biases, (self-)attention, negative activations) can theoretically also affect attribution methods, we want to test if similar distinctions between Transformer and CNN architectures can also be detected for the performance of attribution methods. To this end, we analyze the distance of ranks between each model architecture for each attribution method (see Figure 3 (b.) and Appendix L for all attribution methods and modalities and

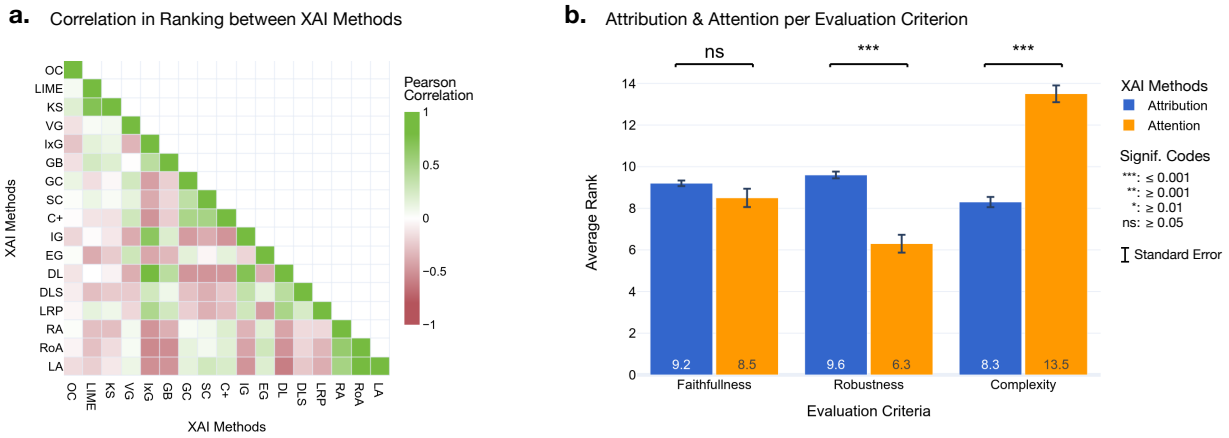


Figure 4: **a.** Correlation in ranking between XAI methods. **b.** Average attribution (across all architectures, but across Transformers-only is very alike) and attention (Transformers-only) ranks.

Appendix M for differences in ranking order). **(2)** The average rank distance does not change substantially between CNN and Transformer architectures for most attribution methods (except for CAM methods and LRP), which is mainly centered around the modality mean. This indicates that almost all attribution methods do not receive significantly different evaluation scores depending on the underlying architecture and we can not support the hypothesis that attribution methods behave fundamentally differently on Transformer architectures compared to CNNs. **(3)** CAM methods generally show a higher rank distance between architectures, which could be attributed to differences in latent representations of the models, as the semantics captured in the last convolutional or cls-token layers do not have to coincide between models. Thus, we recommend increasing the robustness by averaging the activation map of several hidden layers, which has shown effective in application (Gildenblat, 2024), but can lead to less accurate saliency maps. **(4)** LRP shows additionally high dissimilarity between CNN and Transformer architectures, especially for the 3D modalities. We use the recommended γ- and ϵ-rules in LRP for the CNN models. However, on Transformer architectures, LRP does not preserve the conservation rule and only works with the 0^+-rule (see Chefer et al. (2021) for a detailed explanation). Both implemented changes to LRP bias the relevance computation, which consequentially impacts its performance on Transformer architectures. Thus, we recommend using LA instead of LRP as a relevance-based method on Transformer architectures, as it leverages the Transformer-inherent attention and performs much better regarding faithfulness and robustness. *Summarizing, XAI performance just weakly depends on the model performance and the model architecture only significantly influences CAM and LRP methods.*

What behavioral similarities exist among XAI methods? To resolve inconsistencies in current research for method selection, our analysis of XAI behavior focuses on two key aspects: similarities among methods and distinct performance trends. Similarity is important in method selection because choosing a heterogeneous set of XAI methods includes different perspectives on the explanation, which is often advantageous in application. In specific, we analyze the similarity between single methods and the subgroups of attention and attribution methods, obtaining findings 1-4, answering our main question. Figure 4 (a.) shows the correlation in ranking between XAI methods, indicating their relative similarity. **(1)** We observe that methods belonging to methodological similar groups are positively correlated: Linear surrogate methods (LIME, KS), CAM methods (GC, SC, C+), and attention methods (RA, RoA, LA). Also, CAM and attention methods are slightly positively correlated, indicating their similar attributing to local regions. We would advise not restricting access to such methodological subgroups to preserve method diversity in application. **(2)** Contrarily to other method subgroups, the Shapely value approximating SHAP methods (EG, KS, and DLS) are not correlated. Also, their performances in Table 2 differ extensively. This observation is consistent with the results of Molnar et al. (2022), which are, however, not in the context of XAI evaluation. Therefore, it is advisable not to select a single SHAP method with the expectation of achieving similar results to others but rather to employ multiple such methods. **(3)** CAM and attention methods negatively correlate with

IxG, GB, IG, and DL, which contrarily attribute to single pixels, resulting in more fine-grade saliency maps. Interestingly, we observe a very strong positive correlation between IG/IxG, DL/IxG, and IG/DL, indicating very homogeneous behavior between the methods, even though they are based on different mathematical mechanisms. We would strongly recommend mixing such single-pixel and local-region attributing methods, not only for the diversity in visualization but also because of their different performance in evaluation.

Due to the success of Transformers, attention methods are one of the most emerging subgroups of XAI methods. This raises a pressing question for users: should they exclusively use Transformer-based models for attention methods, or can architecture-independent attribution methods still provide equal or superior explanations? **(4)** When comparing the ranking between both groups, we observe in Figure 4 (b.) a large difference in complexity and a smaller difference in robustness while the difference in faithfulness is insignificant. The comparatively high robustness of attention methods extends across all methods and modalities, as can be seen from Table 2. Considering our concerns about the complexity metrics, we would subsequently generally advocate for prioritizing attention methods over attribution methods if Transformers are suitable for the problem.

What are prominent trends for individual XAI methods? The comprehensive evaluation further revealed distinct findings (1-4) relevant for selecting individual methods. **(1)** The most reliable performing method in terms of faithfulness and robustness is EG. We would recommend EG as an initial approach in various situations due to its weak dependence on hyperparameters and input modalities. This is especially important for data with non-triviality to select baseline values. **(2)** We observe a high variance in performance between methods that rely heavily on the gradient (VG, IxG, GB), with only the raw gradient VG being robust but not faithful, which can be attributed to gradient shattering (Balduzzi et al., 2017). We would therefore advise against these methods or advance them through VarGrad or SmoothGrad. **(3)** Relevance-filtered attention (LA) consistently scores better than non-filtered raw attention. Thus we would always prefer it. Also, it allows to visualize input features that attribute to a specific outcome and are not only detected by the model in general, making it much more versatile. **(4)** For some methods (EG, LRP, LA) we observe a trade-off between faithfulness and complexity. For LRP, its tendency to attribute to a very small set of input features likely explains this observation: Faithfulness is low due to the absence of important input features in the set, robustness is low as the relative change in this set can occur fast, but complexity, as evaluated in our metrics, is also low due to the small set size. We would argue that, theoretically, there can be a trade-off between faithfulness and complexity. However, we advise against overinterpreting these findings in light of our reservations regarding the complexity metrics and the subjectiveness of such a trade-off.

4.1 Main insights & takeaways

Our benchmark study presents key findings to bridge research gaps and inconsistencies. More specific findings are obtainable from our full ranking tables and the LATEC dataset. The most critical outcomes of our study can be condensed into the following main insights for evaluation and practical takeaways for XAI application:

Insights for XAI evaluation

1. For the same XAI method, there is frequently a high risk of conflicting metrics, leading to unreliable rankings. However possible disagreement between metrics can provide essential insights for XAI evaluation.
2. High disagreement among metrics can lead to biases when deliberately pairing similar-ranking metrics. However, in a large-scale setting, the actual risk of undue influence from such subgroups is minimal.
3. Results in complexity seem counterintuitive, indicating that the evaluation objective of complexity metrics in computer vision does not always has to match the perception of low complexity.

Takeaways for XAI application

1. No XAI method ranks consistently high on all evaluation criteria. Curiously, EG performs most reliable, even if it was never examined in any relevant related study before.
2. Attention methods are mainly more robust compared to attribution methods, with relevance-filtered attention (LA) consistently scoring higher than non-filtered raw attention.
3. Rankings of XAI methods typically generalize well over datasets given not too distinct size and characteristics but can highly depend on modality (especially for CAM and linear surrogate methods).

5 Comparison with related work

As we previously stated, the significance of current research is limited due to small and varying subsets in XAI methods, metrics, and criteria used for evaluation, which consequently lead to contradicting outcomes. Due to the comprehensiveness and more robust metric aggregation of our study, we argue that the validity is significantly increased compared to previous work and that our findings offer a more definitive resolution to the inconsistencies observed in earlier, smaller-scale investigations. The difference in scale is particularly evident from Table 1, which presents a summary of 16 related relevant studies (all on image data as there are none for volume or point cloud), with an indication of what criteria were evaluated and how XAI methods performed relative to other methods or baselines analyzed in the study, based on our assessment. We present our results for the image modality indicated at the bottom, which do not necessarily transfer to other modalities as Table 1 shows. While some "evergreen" XAI methods, i.e., VG, IxG, GB, and IG, stand out, Table 1 visualizes how sparse the field of XAI evaluation is, especially for attention methods. Surprisingly, our best-evaluated method in terms of faithfulness and robustness, EG, was never evaluated in any related study.

Back-referencing to Table 1, we observe in several cases similar results to other studies on image data: low faithfulness of GB, IxG (as both methods perform partial input recovery), LIME or LRP by Chefer et al. (2021). High faithfulness of IG by Arras et al. (2022) and Hesse et al. (2023) (depends highly on the selected baseline) and LA (only two studies including attention methods). Regarding the conflicting outcomes reported for GC, our results show average faithfulness but high robustness on image data (but can depend on the underlying model, as our work suggests). On the contrary, our results contradict the findings on high faithfulness of LRP by Arras et al. (2022), low faithfulness of IG, and high robustness of KS (KS studies used lower dimensional image data). However, these results can differ between modalities, as GC, for example, obtains very low scores in faithfulness and robustness on volume data.

Most work in complexity and human understandability is qualitative and only quantitative work includes papers that present a metric. We consider the high fluctuation between quantitative and especially qualitative complexity evaluation outcomes by Singh et al. (2021) as further support for our hypothesis that there is a gap between the aim of the metrics and human conception of low complexity. As also the human conception of low complexity can be very subjective, we strongly recommend the development of either new metrics or falling back to robust qualitative user studies, even if they are more difficult and resource-consuming.

6 Conclusion and discussion

Although our benchmark is one of the most comprehensive in the field, we restricted us to the modalities with the, in our opinion, most unique and not overlapping characteristics, ignoring e.g. video data, and did not include more unconventional post-hoc XAI methods such as symbolic representations or metamodels. We also did not include the evaluation criteria of localization and axiomatic properties as they either require ground-truth bounding boxes or can not be applied to all XAI methods. Further, our benchmark focuses on the comparison between methods, not on the evaluation of whether a method is faithful or robust in general, thus ignoring e.g. synthetic baselines. Our findings demonstrate high generalizability, yet it's crucial to contextualize them within the benchmark, considering possible variance introduced by real-world and large-scale scenarios.

Our results demonstrate vividly the need for rethinking the application of evaluation metrics and the risks of inconsistent benchmarking for practitioners and researchers. As a solution, we offer practitioners concise, practical takeaways for applying and selecting XAI methods derived from our large-scale benchmark. This includes the most all-encompassing answer to *"What XAI method should I (not) use for my problem?"* to date, based on the extensive evidence in our provided result's tables and the LATEC dataset. Through these detailed result's tables, practitioners can evaluate which method is most applicable or should be avoided in their specific setting with regard to different criteria. After method selection, practitioners can additionally leverage LATEC to generate diverse and large-scale sets of saliency maps including different computer vision modalities. For researchers, we propose a robust evaluation mechanism, addressing the risk of conflicting metrics, and introduce LATEC as a platform for standardized benchmarking of methods and metrics in

XAI. LATEC offers researchers the opportunity to explore and answer numerous critical questions regarding trustworthiness of XAI, thereby playing a pivotal role in the advancement of the field. We hope that our systematic approach to addressing significant shortcomings in existing research will provide a clearer and more reliable path through the intricate maze of methods and metrics in XAI.

References

Samira Abnar and Willem Zuidema. Quantifying Attention Flow in Transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4190–4197, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.385. URL https://aclanthology.org/2020.acl-main.385.

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity Checks for Saliency Maps. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://papers.nips.cc/paper_files/paper/2018/hash/294a8ed24b1ad22ec2e7efea049b8737-Abstract.html.

Julius Adebayo, Michael Muelly, Ilaria Liccardi, and Been Kim. Debugging Tests for Model Explanations. In *Advances in Neural Information Processing Systems*, volume 33, pp. 700–712. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/075b051ec3d22dac7b33f788da631fd4-Abstract.html.

Chirag Agarwal, Nari Johnson, Martin Pawelczyk, Satyapriya Krishna, Eshika Saxena, Marinka Zitnik, and Himabindu Lakkaraju. Rethinking Stability for Attribution-based Explanations, March 2022. URL http://arxiv.org/abs/2203.06877. arXiv:2203.06877 [cs].

David Alvarez Melis and Tommi Jaakkola. Towards Robust Interpretability with Self-Explaining Neural Networks. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/hash/3e9f0fc9b2f89e043bc6233994dfcf76-Abstract.html.

Leila Arras, Ahmed Osman, and Wojciech Samek. CLEVR-XAI: A benchmark dataset for the ground truth evaluation of neural network explanations. *Information Fusion*, 81:14–40, May 2022. ISSN 1566-2535. doi: 10.1016/j.inffus.2021.11.008. URL https://www.sciencedirect.com/science/article/pii/S1566253521002335.

Nishanth Arun, Nathan Gaw, Praveer Singh, Ken Chang, Mehak Aggarwal, Bryan Chen, Katharina Hoebel, Sharut Gupta, Jay Patel, Mishka Gidwani, Julius Adebayo, Matthew D. Li, and Jayashree Kalpathy-Cramer. Assessing the Trustworthiness of Saliency Maps for Localizing Abnormalities in Medical Imaging. *Radiology. Artificial Intelligence*, 3(6):e200267, November 2021. ISSN 2638-6100. doi: 10.1148/ryai.2021200267.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLOS ONE*, 10(7):e0130140, July 2015. ISSN 1932-6203. doi: 10.1371/journal.pone.0130140. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0130140. Publisher: Public Library of Science.

David Balduzzi, Marcus Frean, Lennox Leary, J. P. Lewis, Kurt Wan-Duo Ma, and Brian McWilliams. The Shattered Gradients Problem: If resnets are the answer, then what is the question? In *Proceedings of the 34th International Conference on Machine Learning*, pp. 342–350. PMLR, July 2017. URL https://proceedings.mlr.press/v70/balduzzi17b.html. ISSN: 2640-3498.

Umang Bhatt, Adrian Weller, and José M. F. Moura. Evaluating and Aggregating Feature-based Model Explanations. In Christian Bessiere (ed.), *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pp. 3016–3022. International Joint Conferences on Artificial Intelligence Organization, July 2020. doi: 10.24963/ijcai.2020/417. URL https://doi.org/10.24963/ijcai.2020/417.

Umang Bhatt, Adrian Weller, and José M. F. Moura. Evaluating and aggregating feature-based model explanations. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20, pp. 3016–3022, Yokohama, Yokohama, Japan, January 2021. ISBN 978-0-9992411-6-5.

Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, Klaus-Robert Müller, and Wojciech Samek. Layer-Wise Relevance Propagation for Neural Networks with Local Renormalization Layers. In Alessandro E.P. Villa, Paolo Masulli, and Antonio Javier Pons Rivero (eds.), *Artificial Neural Networks and Machine Learning – ICANN 2016*, Lecture Notes in Computer Science, pp. 63–71, Cham, 2016. Springer International Publishing. ISBN 978-3-319-44781-0. doi: 10.1007/978-3-319-44781-0_8.

Prasad Chalasani, Jiefeng Chen, Amrita Roy Chowdhury, Xi Wu, and Somesh Jha. Concise Explanations of Neural Networks using Adversarial Training. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 1383–1391. PMLR, November 2020. URL https://proceedings.mlr.press/v119/chalasani20a.html. ISSN: 2640-3498.

Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An Information-Rich 3D Model Repository. Technical Report arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago, 2015.

Aditya Chattopadhay, 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)*, pp. 839–847, March 2018. doi: 10.1109/WACV.2018.00097.

Hila Chefer, Shir Gur, and Lior Wolf. Transformer Interpretability Beyond Attention Visualization. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 782–791, Nashville, TN, USA, June 2021. IEEE. ISBN 978-1-66544-509-2. doi: 10.1109/CVPR46437.2021.00084. URL https://ieeexplore.ieee.org/document/9577970/.

Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote Sensing Image Scene Classification: Benchmark and State of the Art. *Proceedings of the IEEE*, 105(10):1865–1883, October 2017. ISSN 1558-2256. doi: 10.1109/JPROC.2017.2675998. Conference Name: Proceedings of the IEEE.

Sanjoy Dasgupta, Nave Frost, and Michal Moshkovitz. Framework for Evaluating Faithfulness of Local Explanations. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 4794–4815. PMLR, June 2022. URL https://proceedings.mlr.press/v162/dasgupta22a.html. ISSN: 2640-3498.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, June 2009. doi: 10.1109/CVPR.2009.5206848. ISSN: 1063-6919.

Ann-Kathrin Dombrowski, Christopher J. Anders, Klaus-Robert Müller, and Pan Kessel. Towards robust explanations for deep neural networks. *Pattern Recognition*, 121(C), January 2022. ISSN 0031-3203. doi: 10.1016/j.patcog.2021.108194. URL https://doi.org/10.1016/j.patcog.2021.108194.

Finale Doshi-Velez and Been Kim. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv:1702.08608 [cs, stat]*, March 2017. URL http://arxiv.org/abs/1702.08608. arXiv: 1702.08608.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Gabriel Erion, Joseph D. Janizek, Pascal Sturmfels, Scott Lundberg, and Su-In Lee. Improving performance of deep learning models with axiomatic attribution priors and expected gradients, November 2020. URL http://arxiv.org/abs/1906.10670. arXiv:1906.10670 [cs, stat].

Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards Automatic Concept-based Explanations. In *dvances in Neural Information Processing Systems*, volume 32, pp. 10, 2019. URL https://proceedings.neurips.cc/paper/2019/file/77d2afcb31f6493e350fca61764efb9a-Paper.pdf.

Jacob Gildenblat. PyTorch library for CAM methods, January 2024. URL https://github.com/jacobgil/pytorch-grad-cam.

Meng-Hao Guo, Jun-Xiong Cai, Zheng-Ning Liu, Tai-Jiang Mu, Ralph R. Martin, and Shi-Min Hu. PCT: Point cloud transformer. *Computational Visual Media*, 7(2):187–199, June 2021. ISSN 2096-0662. doi: 10.1007/s41095-021-0229-5. URL https://doi.org/10.1007/s41095-021-0229-5.

Anna Hedström, Leander Weber, Dilyara Bareeva, Franz Motzkus, Wojciech Samek, Sebastian Lapuschkin, and Marina M.-C. Höhne. Quantus: An Explainable AI Toolkit for Responsible Evaluation of Neural Network Explanations, February 2022. URL http://arxiv.org/abs/2202.06861. arXiv:2202.06861 [cs].

Anna Hedström, Leander Weber, Daniel Krakowczyk, Dilyara Bareeva, Franz Motzkus, Wojciech Samek, Sebastian Lapuschkin, and Marina Marina M.-C. Höhne. Quantus: An Explainable AI Toolkit for Responsible Evaluation of Neural Network Explanations and Beyond. *Journal of Machine Learning Research*, 24 (34):1–11, 2023. URL http://jmlr.org/papers/v24/22-0142.html.

Robin Hesse, Simone Schaub-Meyer, and Stefan Roth. FunnyBirds: A Synthetic Vision Dataset for a Part-Based Analysis of Explainable AI Methods. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, October 2-6, 2023*, pp. 3981–3991. IEEE, 2023.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A Benchmark for Interpretability Methods in Deep Neural Networks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://papers.nips.cc/paper_files/paper/2019/hash/fe4b8556000d0f0cae99daa5c5c5a410-Abstract.html.

Ioannis Kakogeorgiou and Konstantinos Karantzalos. Evaluating explainable artificial intelligence methods for multi-label deep learning classification tasks in remote sensing. *International Journal of Applied Earth Observation and Geoinformation*, 103:102520, December 2021. ISSN 1569-8432. doi: 10.1016/j.jag.2021.102520. URL https://www.sciencedirect.com/science/article/pii/S0303243421002270.

Daniel S. Kermany, Michael Goldbaum, Wenjia Cai, Carolina C. S. Valentim, Huiying Liang, Sally L. Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, Justin Dong, Made K. Prasadha, Jacqueline Pei, Magdalene Y. L. Ting, Jie Zhu, Christina Li, Sierra Hewett, Jason Dong, Ian Ziyar, Alexander Shi, Runze Zhang, Lianghong Zheng, Rui Hou, William Shi, Xin Fu, Yaou Duan, Viet A. N. Huu, Cindy Wen, Edward D. Zhang, Charlotte L. Zhang, Oulan Li, Xiaobo Wang, Michael A. Singer, Xiaodong Sun, Jie Xu, Ali Tafreshi, M. Anthony Lewis, Huimin Xia, and Kang Zhang. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. *Cell*, 172(5):1122–1131.e9, February 2018. ISSN 0092-8674. doi: 10.1016/j.cell.2018.02.010. URL https://www.sciencedirect.com/science/article/pii/S0092867418301545.

Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T. Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. The (Un)reliability of Saliency Methods. In Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller (eds.), *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, Lecture Notes in Computer Science, pp. 267–280. Springer International Publishing, Cham, 2019. ISBN 978-3-030-28954-6. doi: 10.1007/978-3-030-28954-6_14. URL https://doi.org/10.1007/978-3-030-28954-6_14.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. Captum: A unified and generic model interpretability library for PyTorch, September 2020. URL http://arxiv.org/abs/2009.07896. arXiv:2009.07896 [cs, stat].

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, November 1998. ISSN 1558-2256. doi: 10.1109/5.726791. Conference Name: Proceedings of the IEEE.

Howard Levene. *Robust tests for equality of variances*. Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling. Stanford University Press, 1960. ISBN 0-8047-0596-8.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 4768–4777, Red Hook, NY, USA, December 2017. Curran Associates Inc. ISBN 978-1-5108-6096-4.

Christoph Molnar, Gunnar König, Julia Herbinger, Timo Freiesleben, Susanne Dandl, Christian A. Scholbeck, Giuseppe Casalicchio, Moritz Grosse-Wentrup, and Bernd Bischl. General Pitfalls of Model-Agnostic Interpretation Methods for Machine Learning Models. In Andreas Holzinger, Randy Goebel, Ruth Fong, Taesup Moon, Klaus-Robert Müller, and Wojciech Samek (eds.), *xxAI - Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers*, Lecture Notes in Computer Science, pp. 39–68. Springer International Publishing, Cham, 2022. ISBN 978-3-031-04083-2. doi: 10.1007/978-3-031-04083-2_4. URL https://doi.org/10.1007/978-3-031-04083-2_4.

Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1–15, February 2018. ISSN 1051-2004. doi: 10.1016/j.dsp.2017.10.011. URL https://www.sciencedirect.com/science/article/pii/S1051200417302385.

Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. Layer-Wise Relevance Propagation: An Overview. In Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller (eds.), *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, Lecture Notes in Computer Science, pp. 193–209. Springer International Publishing, Cham, 2019. ISBN 978-3-030-28954-6. doi: 10.1007/978-3-030-28954-6_10. URL https://doi.org/10.1007/978-3-030-28954-6_10.

An-phi Nguyen and María Rodríguez Martínez. On quantitative aspects of model interpretability, July 2020. URL http://arxiv.org/abs/2007.07584. arXiv:2007.07584 [cs, stat].

Weili Nie, Yang Zhang, and Ankit Patel. A Theoretical Explanation for Perplexing Behaviors of Backpropagation-based Visualizations. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 3809–3818. PMLR, July 2018. URL https://proceedings.mlr.press/v80/nie18a.html. ISSN: 2640-3498.

Namuk Park and Songkuk Kim. How Do Vision Transformers Work?, June 2022. URL http://arxiv.org/abs/2202.06709. arXiv:2202.06709 [cs].

Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: Randomized Input Sampling for Explanation of Black-box Models, September 2018. URL http://arxiv.org/abs/1806.07421. arXiv:1806.07421 [cs].

Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do Vision Transformers See Like Convolutional Neural Networks?, March 2022. URL http://arxiv.org/abs/2108.08810. arXiv:2108.08810 [cs, stat].

Anurag Ranjan, Timo Bolkart, Soubhik Sanyal, and Michael J. Black. Generating 3D faces using Convolutional Mesh Autoencoders. In *European Conference on Computer Vision (ECCV)*, pp. 725–741, 2018. URL http://coma.is.tue.mpg.de/.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 1135–1144, New York, NY, USA, August 2016. Association for Computing Machinery. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939778. URL https://dl.acm.org/doi/10.1145/2939672.2939778.

Laura Rieger and Lars Kai Hansen. IROF: a low resource evaluation metric for explanation methods, March 2020. URL http://arxiv.org/abs/2003.08747. arXiv:2003.08747 [cs].

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*, pp. 18770–18795. PMLR, 2022.

Avi Rosenfeld. Better Metrics for Evaluating Explainable Artificial Intelligence. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '21, pp. 45–50, Richland, SC, May 2021. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 978-1-4503-8307-3.

Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. Evaluating the Visualization of What a Deep Neural Network Has Learned. *IEEE Transactions on Neural Networks and Learning Systems*, 28(11):2660–2673, November 2017. ISSN 2162-2388. doi: 10.1109/TNNLS.2016.2599820. Conference Name: IEEE Transactions on Neural Networks and Learning Systems.

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)*, pp. 618–626, October 2017. doi: 10.1109/ICCV.2017.74. ISSN: 2380-7504.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pp. 3145–3153, Sydney, NSW, Australia, August 2017. JMLR.org.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *CoRR*, abs/1312.6034, 2013. URL https://api.semanticscholar.org/CorpusID:1450294.

Amitojdeep Singh, Janarthanam Jothi Balaji, Mohammed Abdul Rasheed, Varadharajan Jayakumar, Rajiv Raman, and Vasudevan Lakshminarayanan. Evaluation of Explainable Deep Learning Methods for Ophthalmic Diagnosis. *Clinical Ophthalmology (Auckland, N.Z.)*, 15:2573–2581, 2021. ISSN 1177-5467. doi: 10.2147/OPTH.S312236.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda B. Viégas, and Martin Wattenberg. SmoothGrad: removing noise by adding noise. *CoRR*, abs/1706.03825, 2017. URL http://arxiv.org/abs/1706.03825. arXiv: 1706.03825.

Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin A. Riedmiller. Striving for Simplicity: The All Convolutional Net. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6806.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic Attribution for Deep Networks. *arXiv:1703.01365 [cs]*, June 2017. URL http://arxiv.org/abs/1703.01365. arXiv: 1703.01365.

Hugo Touvron, Matthieu Cord, and Hervé Jégou. DeiT III: Revenge of the ViT. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV*, pp. 516–533, Berlin, Heidelberg, October 2022. Springer-Verlag. ISBN 978-3-031-20052-6. doi: 10.1007/978-3-031-20053-3_30. URL https://doi.org/10.1007/978-3-031-20053-3_30.

Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. Designing Theory-Driven User-Centric Explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pp. 1–15, New York, NY, USA, May 2019. Association for Computing Machinery. ISBN 978-1-4503-5970-2. doi: 10.1145/3290605.3300831. URL https://doi.org/10.1145/3290605.3300831.

H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, and X. Hu. Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 111–119, Los Alamitos, CA, USA, June 2020. IEEE Computer Society. doi: 10.1109/CVPRW50498.2020.00020. URL https://doi.ieeecomputersociety.org/10.1109/CVPRW50498.2020.00020.

Yi Wang, Zhiwen Fan, Tianlong Chen, Hehe Fan, and Zhangyang Wang. Can We Solve 3D Vision Tasks Starting from A 2D Vision Transformer?, September 2022. URL http://arxiv.org/abs/2209.07026. arXiv:2209.07026 [cs].

Zeyu Wang, Yutong Bai, Yuyin Zhou, and Cihang Xie. Can CNNs Be More Robust Than Transformers?, March 2023. URL http://arxiv.org/abs/2206.03452. arXiv:2206.03452 [cs].

Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3D ShapeNets: A deep representation for volumetric shapes. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1912–1920, 2014. URL https://api.semanticscholar.org/CorpusID:206592833.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. MedMNIST v2 - A large-scale lightweight benchmark for 2D and 3D biomedical image classification. *Scientific Data*, 10(1):41, January 2023. ISSN 2052-4463. doi: 10.1038/s41597-022-01721-8. URL https://www.nature.com/articles/s41597-022-01721-8. Number: 1 Publisher: Nature Publishing Group.

Mengjiao Yang and Been Kim. Benchmarking Attribution Methods with Relative Feature Importance. *arXiv:1907.09701 [cs, stat]*, November 2019. URL http://arxiv.org/abs/1907.09701. arXiv: 1907.09701.

Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Sai Suggala, David I. Inouye, and Pradeep Ravikumar. On the (In)fidelity and Sensitivity for Explanations. *arXiv:1901.09392 [cs, stat]*, November 2019. URL http://arxiv.org/abs/1901.09392. arXiv: 1901.09392.

Matthew D. Zeiler and Rob Fergus. Visualizing and Understanding Convolutional Networks. *ArXiv*, abs/1311.2901, 2013. URL https://api.semanticscholar.org/CorpusID:3960646.

Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-Down Neural Attention by Excitation Backprop. *International Journal of Computer Vision*, 126(10): 1084–1102, October 2018. ISSN 1573-1405. doi: 10.1007/s11263-017-1059-x. URL https://doi.org/10.1007/s11263-017-1059-x.

Daquan Zhou, Zhiding Yu, Enze Xie, Chaowei Xiao, Anima Anandkumar, Jiashi Feng, and Jose M. Alvarez. Understanding The Robustness in Vision Transformers, November 2022. URL http://arxiv.org/abs/2204.12451. arXiv:2204.12451 [cs].

Appendix

A Model performance and hyperparameter

A.1 Test set performance

a. Testset Performance on Image Modality

Dataset:	Model Architecture:	Accuracy	Precision	Recall	F1	AUROC
OCT (MC: 4)	ResNet 50	0.999	0.999	0.999	0.999	1.0
	EfficientNet b0	0.9969	0.9969	0.9969	0.9969	1.0
	DeiT ViT	0.999	0.999	0.999	0.999	1.0
R45 (MC: 45)	ResNet 50	0.9535	0.9536	0.9538	0.9535	0.9995
	EfficientNet b0	0.9554	0.9554	0.9549	0.9549	0.9995
	DeiT ViT	0.9568	0.957	0.9568	0.9567	0.9995

b. Testset Performance on Volume Modality

Dataset:	Model Architecture:	Accuracy	Precision	Recall	F1	AUROC
AMN (BC)	3D ResNet 18	0.8003	0.8013	0.7987	0.8	0.8699
	EfficientNet3D b0	0.8003	0.7954	0.8087	0.802	0.8647
	Simple3DFormer	0.7936	0.7907	0.7907	0.7907	0.8728
OMN (MC: 11)	3D ResNet 18	0.9115	0.9248	0.9248	0.9226	0.9953
	EfficientNet3D b0	0.8754	0.8924	0.8936	0.8914	0.9893
	Simple3DFormer	0.8131	0.8463	0.8381	0.84	0.9815
VMN (BC)	3D ResNet 18	0.9359	0.937	0.9346	0.9358	0.98
	EfficientNet3D b0	0.9162	0.9162	0.9162	0.9162	0.9229
	Simple3DFormer	0.8861	0.8871	0.8848	0.886	0.9394

c. Testset Performance on Point Cloud Modality

Dataset:	Model Architecture:	Accuracy	Precision	Recall	F1	AUROC
CMA (MC: 12)	PointNet	0.9852	0.9743	0.9876	0.98	0.998
	DGCNN	0.9535	0.9373	0.9498	0.9423	0.9989
	PC Transformer	0.9751	0.9645	0.9688	0.9662	0.9996
M40 (MC: 40)	PointNet	0.8914	0.8374	0.8564	0.8438	0.9958
	DGCNN	0.9177	0.8844	0.891	0.8864	0.9973
	PC Transformer	0.9149	0.8779	0.8842	0.8796	0.9969
SHN (MC: 16)	PointNet	0.9878	0.9673	0.9689	0.9668	0.9991
	DGCNN	0.9903	0.966	0.9847	0.9745	0.9995
	PC Transformer	0.9896	0.9642	0.9819	0.9716	0.9997

MC #: Multi-Class (# Classes), **BC**: Binary-Class

Table 3: **a., b. & c.** Test set performance measured with the metrics: accuracy, precision, recall, F1, and area under the receiver operating characteristic (AUROC) curve, for each modality. In the case of IMN we use pretrained weights for the Transformer architecture from Huggingface[1] and the CNN architectures from TorchHub[2,3].

[1] https://huggingface.co/facebook/deit-small-patch16-224

[2] https://pytorch.org/vision/stable/models/generated/torchvision.models.resnet50.html

[3] https://pytorch.org/vision/stable/models/generated/torchvision.models.efficientnet_b0.html

Architectures were chosen based on their popularity and, to a limited extent, comparability between modalities, e.g. ResNet-50 and 3D ResNet-18 which both emerge from the same family of ResNet architectures. While 3D volume architectures could also be applied to point cloud data, we choose point cloud specific architectures for the modality.

A.2 Hyperparameter

We tuned all hyperparameters on either the declared validation set or sampled a validation set based on 20% of the train set. The tuning was performed via grid search for each model. The primary metric for hyperparameter tuning was the F1 score.

Hyperparameter for Image Modality

		Utilized Computer Vision Datasets	
Model Architecture:	Hyperparameter:	OCT	R45
ResNet 50	Batch size	128	128
	Max Epochs	8	60
	Learning rate (LR)	0.0001	0.0001
	Optimizer	Madgrad	Madgrad
	LR Scheduler	Cosine Annealing	Cosine Annealing
	Weight Decay	0	0
	Momentum	0.9	0.9
	Augmentations	Train: Resize (256,256) RandomCrop (224,224) RandomAffine (shear=0.2, degrees=5) RandomHorizontalFlip Grayscale (channels=3) Test: Resize (256,256) CenterCrop (224,224) Grayscale (channels=3)	Train: Resize (256,256) RandomCrop (224,224) RandomHorizontalFlip RandAugment Normalize (mean=(0.485, 0.456, 0. 406), std=(0.229, 0.224, 0.225)) Test: Resize (256,256) Normalize (mean=(0.485, 0.456, 0. 406), std=(0.229, 0.224, 0.225))
	Sampling	Weighted Random Sampling	None
EfficientNet b0	Batch size	128	128
	Max Epochs	5	15
	Learning rate (LR)	0.0001	0.001
	Optimizer	Madgrad	Madgrad
	LR Scheduler	Cosine Annealing	Cosine Annealing
	Weight Decay	0	0
	Momentum	0.9	0.9
	Augmentations	Train: Resize (256,256) RandomCrop (224,224) RandomAffine (shear=0.2, degrees=5) RandomHorizontalFlip Grayscale (channels=3) Test: Resize (256,256) CenterCrop (224,224) Grayscale (channels=3)	Train: Resize (256,256) RandomCrop (224,224) RandomHorizontalFlip RandAugment Normalize (mean=(0.485, 0.456, 0. 406), std=(0.229, 0.224, 0.225)) Test: Resize (256,256) Normalize (mean=(0.485, 0.456, 0. 406), std=(0.229, 0.224, 0.225))
	Sampling	Weighted Random Sampling	None
DeIT ViT	Batch size	128	128
	Max Epochs	6	60
	Learning rate (LR)	0.0001	0.0001
	Optimizer	Madgrad	Madgrad
	LR Scheduler	Cosine Annealing	Cosine Annealing
	Weight Decay	0	0
	Momentum	0.9	0.9
	Augmentations	Train: Resize (256,256) RandomCrop (224,224) RandomAffine (shear=0.2, degrees=5) RandomHorizontalFlip Grayscale (channels=3) Test: Resize (256,256) CenterCrop (224,224) Grayscale (channels=3)	Train: Resize (256,256) RandomCrop (224,224) RandomHorizontalFlip RandAugment Normalize (mean=(0.485, 0.456, 0. 406), std=(0.229, 0.224, 0.225)) Test: Resize (256,256) Normalize (mean=(0.485, 0.456, 0. 406), std=(0.229, 0.224, 0.225))
	Sampling	Weighted Random Sampling	None

Table 4: Hyperparameter for all three architectures and CV datasets, excluding IMN as we load pretrained weights.

Hyperparameter for Volume Modality

Model Architecture:	Hyperparameter:	Utilized Computer Vision Datasets		
		AMN	OMN	VMN
3D ResNet18	Batch size	32	32	32
	Max Epochs	100	100	100
	Learning rate (LR)	0.001	0.001	0.001
	Optimizer	SGD	Adam	Adam
	LR Scheduler	Cosine Annealing	Cosine Annealing	Cosine Annealing
	Weight Decay	0	0	0
	Momentum	0.9	0	0
	Augmentations	Train: RandomBrightness($U(0,1)$) Test: FixedBrightness(0.5)	None	Train: RandomBrightness($U(0,1)$) Test: FixedBrightness(0.5)
	Sampling	Weighted Random Sampling	None	Weighted Random Sampling
3D EfficientNet b0	Batch size	32	32	64
	Max Epochs	100	100	100
	Learning rate (LR)	0.001	0.001	0.001
	Optimizer	SGD	AdamW	Adam
	LR Scheduler	Cosine Annealing	Cosine Annealing	Cosine Annealing
	Weight Decay	0.0005	0.0005	0
	Momentum	0.9	0	0
	Augmentations	Train: RandomBrightness($U(0,1)$) Test: FixedBrightness(0.5)	None	Train: RandomBrightness($U(0,1)$) Test: FixedBrightness(0.5)
	Sampling	Weighted Random Sampling	None	Weighted Random Sampling
Simple3DFormer	Batch size	32	32	64
	Max Epochs	150	100	100
	Learning rate (LR)	0.001	0.000001	0.001
	Optimizer	SGD	Madgrad	Adam
	LR Scheduler	Cosine Annealing	Cosine Annealing	Cosine Annealing
	Weight Decay	0.0005	0	0
	Momentum	0.9	0.9	0
	Augmentations	Train: RandomBrightness($U(0,1)$) Test: FixedBrightness(0.5)	None	Train: RandomBrightness($U(0,1)$) Test: FixedBrightness(0.5)
	Sampling	Weighted Random Sampling	None	Weighted Random Sampling

Table 5: Hyperparameters for all three architectures and CV datasets.

Hyperparameter for Point Cloud Modality

Model Architecture:	Hyperparameter:	Utilized Computer Vision Datasets		
		CMA	M40	SHN
PointNet	Batch size	32	24	32
	Max Epochs	100	200	200
	Learning rate (LR)	0.001	0.001	0.001
	Optimizer	AdamW	AdamW	AdamW
	LR Scheduler	Cosine Annealing	Cosine Annealing	Cosine Annealing
	Weight Decay	0.0001	0.0001	0.0001
	Momentum	0	0	0
	Augmentations	Pretransforms: NormalizeScale Train: SamplePoints (1024) RandomScale (0.67,1.5) RandomRotate (degrees=15) RandomJitter (0.02) Test: SamplePoints (1024)	Pretransforms: NormalizeScale Train: SamplePoints (1024) RandomScale (0.67,1.5) RandomJitter (0.02) Test: SamplePoints (1024)	Pretransforms: NormalizeScale Train: RandomScale (0.67,1.5) RandomJitter (0.01) RandomRotate(degress=15, axis = (0,1,2)) Test: None
	Sampling	None	None	None
DGCNN	Batch size	32	32	32
	Max Epochs	100	250	200
	Learning rate (LR)	0.001	0.001	0.001
	Optimizer	AdamW	AdamW	AdamW
	LR Scheduler	Cosine Annealing	Cosine Annealing	Cosine Annealing
	Weight Decay	0.0001	0.0001	0.0001
	Momentum	0	0	0
	Augmentations	Pretransforms: NormalizeScale Train: SamplePoints (1024) RandomScale (0.67,1.5) RandomRotate (degrees=15) RandomJitter (0.02) Test: SamplePoints (1024)	Pretransforms: NormalizeScale Train: SamplePoints (1024) RandomScale (0.67,1.5) RandomJitter (0.02) Test: SamplePoints (1024)	Pretransforms: NormalizeScale Train: RandomScale (0.67,1.5) RandomJitter (0.01) RandomRotate(degress=15, axis = (0,1,2)) Test: None
	Sampling	None	None	None
PC Transformer	Batch size	32	32	32
	Max Epochs	150	250	200
	Learning rate (LR)	0.01	0.01	0.01
	Optimizer	SGD	SGD	SGD
	LR Scheduler	Cosine Annealing	Cosine Annealing	Cosine Annealing
	Weight Decay	0.0005	0.0005	0.0005
	Momentum	0.9	0.9	0.9
	Augmentations	Pretransforms: NormalizeScale Train: SamplePoints (1024) RandomScale (0.67,1.5) RandomRotate (degrees=15) RandomJitter (0.02) Test: SamplePoints (1024)	Pretransforms: NormalizeScale Train: SamplePoints (1024) RandomScale (0.67,1.5) RandomJitter (0.02) Test: SamplePoints (1024)	Pretransforms: NormalizeScale Train: RandomScale (0.67,1.5) RandomJitter (0.01) RandomRotate(degress=15, axis = (0,1,2)) Test: None
	Sampling	None	None	None

Table 6: Hyperparameters for all three architectures and CV datasets.

B The LATEC dataset: Reference data for standardized evaluation

The resulting data of the three stages, which comprise the LATEC dataset, include pretrained model weights (excluding IMN), saliency maps, and evaluation scores. Thanks to the LATEC dataset, future experiments can start at a certain stage and use the results from the previous stage without recomputing everything again, e.g. when testing out a new evaluation metric on the existing saliency maps, preserving comparability. For the LATEC dataset, we compute per dataset saliency maps for the entire test set or 1000 observations depending on which size is smaller (on the validation set if the test set is unavailable), from which we sample 50 observations to compute evaluation scores for all 7,560 combinations. In total, the LATEC dataset consists of 326,790 saliency maps and 378,000 evaluation scores. As for such large datasets, the size can go into the hundreds of gigabytes. To save disk space, saliency maps could be cast from 64-bit precision to 32 or even 16-bit. We would, however, strongly advise against this, as even casting to 32-bit precision introduced numerical instability in our experiments due to the rounding of attribution and attention values, resulting in all-zero saliency maps and *nan* or *inf* evaluation scores. Further, as ranking lengths between CNN and Transformer architectures differ (attention methods only for Transformer architectures), we recompute rankings in the subsequent study, which aggregate over all three architectures by first combining the normalized evaluation

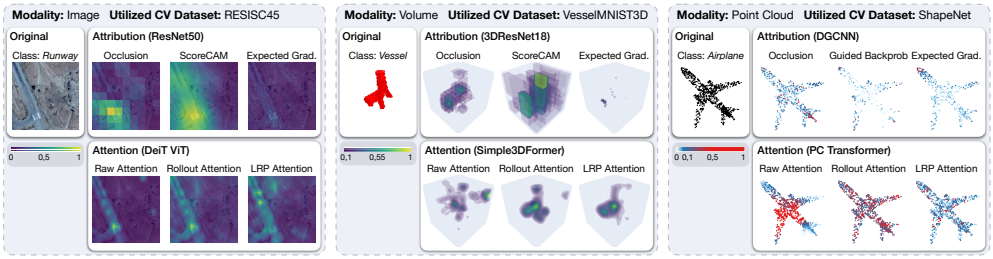


Figure 5: Illustrative saliency maps for all three modalities. The upper row shows three attributions, respectively, and the lower row, three attention-based methods. We observe how all XAI methods highlight the runway in the image and the vessel for the volume modality but with different granularity and focus. For the point cloud plane, explanations are less understandable, with attribution methods highlighting single points at the front tip, rudder, or wing tips.

scores per model architecture and then computing the ranking, preserving equal length between rankings (see Appendix G).

To ensure a standardized setting with fair comparability between XAI methods over all possible experiment set-ups and aggregation levels, we take precautions regarding e.g. different types of feature attributions or the conversion of all metrics to single scores (see Appendix F for all detailed procedures). LRP requires non-negative activation outputs Montavon et al. (2019), leading us to a replacement of such activation functions (i.e. GeLU, leakyReLU) in CNN models, but we keep them for Transformer models, as they are central to the architecture and therefore also to our benchmark, and apply the 0^+-rule instead.

C XAI methods overview and parameters

C.1 Overview

C.1.1 Attribution Methods

Occlusion [OC] (Zeiler & Fergus, 2013) Systematically obscures different parts of the input data and observes the resulting impact on the output, to determine which parts of the data are most important for the model's predictions.

LIME [LIME] (Ribeiro et al., 2016) Creates an interpretable model around the prediction of a complex model to explain individual predictions locally (patch-based in our case), using perturbations of the input data and observing the corresponding changes in the output.

Kernel SHAP [KS] (Lundberg & Lee, 2017) Using a weighted linear regression model as the local surrogate and selecting a suitable weighting kernel, the regression coefficients from the LIME surrogate can estimate the SHAP values.

Vanilla Gradient [VG] (Simonyan et al., 2013) The raw input gradients of the model.

Input x Gradient [IxG] (Shrikumar et al., 2017) Multiples the input features by their corresponding gradients with respect to the model's output.

Guided Backprob [GB] (Springenberg et al., 2015) Modifies the standard backpropagation process to only propagate positive gradients for positive inputs through the network, thereby creating visualizations that highlight the features that strongly activate certain neurons in relation to the target output.

GradCAM [GC] (Selvaraju et al., 2017) Uses the gradients of the target class flowing into the final convolutional layer to produce a coarse localization map by, highlighting the important regions in the image by up-scaling the map.

ScoreCAM [SC] (Wang et al., 2020) Eliminates the need for gradient information by determining the importance of each activation map based on its forward pass score for the target class, producing the final output through a weighted sum of these activation maps.

GradCAM++ [C+] (Chattopadhay et al., 2018) Generates a visual explanation for a given class label by employing a weighted sum of the positive partial derivatives from the final convolutional layer's feature maps, using them as weights with respect to the class score.

Integrated Gradients [IG] (Sundararajan et al., 2017) Explains model predictions by attributing the prediction to the input features, calculating the path integral of the gradients along the straight-line path from a baseline input to the actual input.

Expected Gradients [EG] (Erion et al., 2020) Also called Gradient SHAP. Avoids the selection of a baseline value compared to IG, by leveraging a probabilistic baseline computed over a sample of observations.

DeepLIFT [DL] (Shrikumar et al., 2017) Assigns contribution scores to each input feature based on the difference between the feature's activation and a reference activation, effectively measuring the feature's impact on the output compared to a baseline.

DeepLIFT SHAP [DLS] (Lundberg & Lee, 2017) Combines the DeepLIFT method with Shapley values to assign importance scores to input features by computing their contributions to the output relative to a reference input, while ensuring consistency with Shapley values.

Layer-Wise Relevance Propagation [LRP] (Binder et al., 2016) Explains neural network decisions by backpropagating the output prediction through the layers, redistributing relevance scores to the input features to visualize their contribution to the final decision. We use the ϵ-,γ- and 0^+-rules depending on the model architecture for relevance backpropagation.

C.1.2 Attention Methods

Raw Attention [RA] (Dosovitskiy et al., 2021) Rearranged and up-scaled attention values of the last attention head.

Rollout Attention [RoA] (Abnar & Zuidema, 2020) Averages attention weights of multiple heads to trace the contribution of each part of the input data through the network.

LRP Attention [LA] (Chefer et al., 2021) Assigns local relevance scores to attention weights based on the Deep Taylor Decomposition principle and propagates these relevancy scores through the model.

C.2 Parameters

XAI Method:	OC				LIME			KS			CAM (all)	SC	IG		EG		DL		LRP			RA
Parameter:	strides	sliding_window_shapes	baseline	perturbations_per_eval	alpha	n_samples	perturbations_per_eval	baseline	n_samples	perturbations_per_eval	layer	batch_size	baseline	n_steps	n_samples	std	eps	baseline	rule	eps	gamma	layer
Image	25	(50, 50)	0	1	1,0	10	5	0	10	5	ResNet50,layer4[-1] EfficientNetbo,features[-1] ViT,blocks[-1],norm1	32	0	30	40	0,001	1e-9	0	ε & γ-rule / 0^+-rule	0,0001	0,25	ViT,blocks[-1],attn
Volume	4	(7, 7, 7)	0	1	1,0	10	5	0	10	5	3DEfficientNetbo,blocks[-13] 3DResNet18,layer3 S3DF,blocks[-1],norm1	64	0	30	40	0,001	1e-9	0	ε & γ-rule / 0^+-rule	0,0001	0,25	S3DF[-1],attn
Point Cloud	1	(3,1)	0	5	4,0	10	5	0	10	5	PointNet,transform,bn1 DGCNN,conv5 PCT,sa4,after_norm	16	0	30	16	0,001	1e-9	0	ε & γ-rule / 0^+-rule	0,00001	0,25	PCT,sa4,attn

Table 7: Parameters for each XAI method and modality.

The parameters for each XAI method are derived for each modality via qualitative evaluation which we deem the most realistic scenario. We tuned the XAI methods on five observations per dataset and modality, which we argue is a fair trade-off between fitting the methods to the dataset but not overfitting them to bias the evaluation. We did not tune the parameters per dataset, as the parameters transfer very well between datasets and only needed minimal adjustments.

D Evaluation metrics overview and parameters

D.1 Overview

D.1.1 Faithfulness

Faithfulness Correlation (Bhatt et al., 2020) Gauges an explanation's fidelity to model behavior. It measures the linear correlation between predicted logits of modified test points and the average explanation for selected features, returning a score between -1 and 1. For each test, selected features are replaced with baseline values, and Pearson's correlation coefficient is determined, averaging results over multiple tests.

Faithfulness Estimate (Alvarez Melis & Jaakkola, 2018) Evaluates the accuracy of estimated feature relevances by using a proxy for the "true" influence of features, as the actual influence is often unavailable. This is done by observing how the model's prediction changes when certain features are removed or obscured. Specifically, for probabilistic classification models, the metric looks at how the probability of the predicted class drops when features are removed. This drop is then compared to the interpreter's prediction of that feature's relevance. The metric also computes correlations between these probability drops and relevance scores across various data points.

Monotonicity Correlation (Nguyen & Martínez, 2020) Evaluates the correlation between the absolute values of attributions and the uncertainty in probability estimation using Spearman's coefficient. If attributions are not monotonic the authors argue that they are not providing the correct importance of the features.

Pixel Flipping (Bach et al., 2015) The core concept involves flipping pixels with very high, very low, or near-zero attribution scores. The effect of these changes is then assessed on the prediction scores, with the average prediction being determined.

Region Perturbation (Samek et al., 2017) A step-by-step method where the class representation in the image, as determined by a function, diminishes as we gradually eliminate details from an image. This process, known as region perturbation, occurs at designated locations. Finally, the effect on the average prediction is calculated.

Insertion (Petsiuk et al., 2018) Gradually inserts features into a baseline input, which is a strongly blurred version of the image, to not create OOD examples. During this process, the change in prediction is measured and the correlation with the respective attribution value is calculated.

Deletion (Petsiuk et al., 2018) Deletes input features one at a time by replacing them with a baseline value based on their attribution score. During this process, the change in prediction is measured and the correlation with the respective attribution value is calculated.

Iterative Removal of Features (IROF) (Rieger & Hansen, 2020) The metric calculates the area under the curve for each class based on the sorted average importance of feature segments (superpixels). As these segments are progressively removed and prediction scores gathered, the results are averaged across multiple samples.

Remove and Debias (ROAD) (Rong et al., 2022) Evaluates the model's accuracy on a sample set during each phase of an iterative process where the k most attributed features are removed. To eliminate bias, in every step, the k most significant pixels, by the most relevant first order, are substituted with noise-infused linear imputations.

Sufficiency (Dasgupta et al., 2022) Assesses the likelihood that the prediction label for a specific observation matches the prediction labels of other observations which have similar saliency maps.

D.1.2 Robustness

Local Lipschitz Estimate (Alvarez Melis & Jaakkola, 2018) Lipschitz continuity in calculus is a concept that measures the relative changes in a function's output concerning its input. While the traditional definition of Lipschitz continuity is global, focusing on the largest relative deviations across the entire input space,

this global perspective isn't always meaningful in XAI. This is because expecting consistent explanations for vastly different inputs isn't realistic. Instead, a more localized approach, focusing on stability for neighboring inputs, is preferred, resulting in a point-wise, neighborhood-based local Lipschitz continuity metric.

Max Sensitivity (Yeh et al., 2019) Measures the largest shift in the explanation when the input is slightly altered. It specifically evaluates the utmost sensitivity of a saliency map by taking multiple samples from a defined L-infinity ball subspace with a set input neighborhood radius, using Monte Carlo sampling for approximation.

Continuity (Montavon et al., 2018) Evaluates, that if two observations are nearly equivalent, then the explanations of their predictions should also be nearly equivalent. It then measures the strongest variation of the explanation in the input domain.

Relative Input/Output/Representation Stability (Agarwal et al., 2022) All metrics leverage model information to evaluate the stability of a saliency map with respect to the change in the either, input data, intermediate representations, and output logits of the underlying prediction model.

Infidelity (Yeh et al., 2019) Calculates the expected mean-squared error (MSE) between the saliency map multiplied by a random variable input perturbation and the differences between the model at its input and perturbed input.

D.1.3 Complexity

Sparseness (Chalasani et al., 2020) Measures the Gini Index on the vector of absolute saliency map values. The assessment ensures that features genuinely influencing the output have substantial contributions, while insignificant or only slightly relevant features should have minimal contributions.

Complexity (Bhatt et al., 2020) Determines the entropy of the normalized saliency map.

Effective Complexity (Nguyen & Martínez, 2020) Evaluates the number of absolute saliency map values that surpass a threshold. Values above this threshold suggest the features are significant, while those below indicate they are not.

D.2 Parameters

We tuned the parameters of the evaluation metrics per dataset based on the distribution of their scores from Appendix N. We applied the suggested parameters from Hedström et al. (2022) or the respective papers. If the resulting score distributions were collapsed, almost uniform, or too indistinguishable between the XAI methods, we tuned the respective parameters. This step was completed prior to the ranking analysis, and no adjustments were made to the metrics once the ranking phase commenced.

		Image			Voxel			Point Cloud		
Evaluation Metric:	Parameter:	IMN	OCT	R45	AMN	OMN	VMN	CMA	M40	SHN
Faithfulness Correlation	nr_runs	100	100	100	100	100	100	100	100	100
	subset_size	224	224	224	56	56	56	32	32	32
	perturb_baseline	black	black	black	black	black	black	center	center	center
Faithfulness Estimate	features_in_step	224	224	224	56	56	56	32	32	32
	perturb_baseline	black	black	black	black	black	black	center	center	center
Monotonicity Correlation	nr_samples	10	10	10	10	10	10	10	10	10
	features_in_step	3136	3136	3136	392	392	392	256	256	256
	perturb_baseline	uniform	uniform	uniform	uniform	uniform	uniform	uniform	uniform	uniform
Pixel Flipping	features_in_step	224	224	224	56	56	56	32	32	32
	perturb_baseline	black	black	black	black	black	black	center	center	center
Region Perturbation	patch_size	14	14	18	4	4	4	3	3	3
	regions_evaluation	10	10	20	20	20	20	32	32	32
	perturb_baseline	uniform	uniform	uniform	uniform	uniform	uniform	uniform	uniform	uniform
Insertion	pixel_batch_size	50	50	50	50	50	50	50	50	50
	sigma	5.0	120.0	40.0	2.5	2.5	2.5	0.05	0.1	0.05
	kernel_size	15	39	19	1	1	1	1	1	1
Deletion	pixel_batch_size	50	50	50	50	50	50	50	50	50
IROF	segmentation	Slic	Slic	Slic	3D Slic	3D Slic	3D Slic	KMeans	KMeans	KMeans
	perturb_baseline	mean	mean	mean	black	black	black	center	center	center
ROAD	noise	0.1	0.1	0.1	4.0	2.5	50.0	0.02	0.15	0.3
	percentages_max	100	100	100	100	100	100	100	100	100
Sufficiency	threshold	0.9	0.6	0.6	0.02	0.75	0.0002	0.75	0.75	0.6
Local Lipschitz Estimate	nr_samples	5	5	5	10	10	10	5	5	5
	perturb_std	0.1	0.0002	0.1	0.2	0.2	0.2	0.1	0.1	0.1
	perturb_mean	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MaxSensitivity	nr_samples	10	10	10	10	10	10	10	10	10
	lower_bound	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2
Continuity	patch_size	56	56	56	7	7	7	3	3	3
	nr_steps	20	20	20	20	20	20	20	20	20
	perturb_baseline	uniform	uniform	uniform	uniform	uniform	uniform	uniform	uniform	uniform
RIS	nr_samples	10	10	10	10	10	10	10	10	10
ROS	nr_samples	10	10	10	10	10	10	10	10	10
RRS	nr_samples	10	10	10	10	10	10	10	10	10
Infidelity	n_perturb_samples	50	50	50	50	50	50	50	50	50
Effective Complexity	eps	0.01	0.01	0.01	0.001	0.001	0.001	0.001	0.001	0.001

Table 8: Parameters for all evaluation metrics on each CV dataset.

E Adapting current XAI methods and evaluation metrics for 3D data

While many XAI methods and evaluation metrics are independent of the input space dimensions, especially methods leveraging perturbations, interpolations for up- and down-scaling or segmentation are not. Our implementation builds upon the work from Kokhlikyan et al. (2020) and Hedström et al. (2023) for XAI methods and evaluation metrics for 1D and 2D images, and we extended it to 3D volume and point cloud data. Both modalities come with their own specifies, e.g. that local neighborhoods have to be defined via k-nearest neighbors (KNN) in point cloud data and not 2D or 3D patches as in image or volume data. For the XAI methods, we advanced e.g. OC, LI, and KS by the adoption of 3D patches, all three CAM methods with 3D interpolation, all attention-based methods with 3D and KNN-based interpolations, and LA with relevance backpropagation for the Simple3DFormer and PC Transformer architectures. As the adoption of the CAM methods for point cloud data and more complex architectures than PointNet is not trivial, we deem it out of scope for this paper and do not include them in our point cloud experiments. In the case of evaluation metrics, we adapted e.g. perturbation applying metrics to 3D patches or point-based perturbations, the superpixel segmentation in IROF by 3D Slic and KMeans clustering and padded x-axis transversal for the volume and point cloud data in Continuity. Additionally, we modified all methods and metrics to function with (x, y, z) volume and $(n, 3)$ point cloud dimensions. All adaptations were tested for their coherency, and illustrative saliency maps can be observed in Figure 5. We refer to Appendix C and Appendix subsection D.2 for all implementation details.

E.1 Adaption of XAI methods

In this section, we explain how we adapted XAI methods in our framework to seamlessly work with 3D modalities. We neglect the methods that did not need any adaption (besides e.g. unit tests etc.) as they work independently of the input dimensions. All XAI methods are adapted, such that they only return positive attribution.

Occlusion For the 3D modalities we implemented a 3D kernel as the perturbation baseline for volumes and a 1x3 mask (one point) for the point clouds. The image and volume mask transverse with overlap and the point cloud mask without overlap over all dimensions of the input object.

LIME & Kernel SHAP For both methods, we implemented feature masks for each modality, as training the linear surrogate models on the original input features is not informative and computationally very expensive. Each mask groups the input features to the same interpretable feature. We use predefined grids as feature masks, as superpixel computing algorithms are too computational and time-expensive, especially for 3D modalities and evaluation metrics that perturb the input space or refit the XAI method multiple times. For the image modality, we use a 16x16x3, for volume 7x7x7, and for point cloud 1x3 (one point) mask, which is distributed as a non-overlapping grid in all dimensions over the whole object. For point clouds we use ridge regression and for the other modalities lasso regression.

GradCAM, ScoreCAM & GradCAM++ For all CAM methods on volume data we adapted the gradient averaging and the subsequent weighting of the activations and used nearest-neighbor interpolation to upscale the weighted activations to 3D volumes. In the case of ScoreCAM we also use nearest neighbor up-sampling instead of bilinear up-sampling, to upscale the activations for weighting the output of the previous layer. To correctly reshape the upscaled images and volumes in the case of the Transformer architectures (taking the channels to the first dimension as for CNNs), we use two different reshape functions for images and volumes when the CAM methods are applied to Transformer architectures. Further, we use the absolute activation output, not the non-negative for Transformer architectures, as the leaky-ReLU/GeLU function output otherwise would sometimes be zero.

LRP For CNNs, we assigned the ϵ-rule to the linear or identity layers, the identity rule to all non-linear layers, and to all other layers (convolutions, pooling, batch normalization, etc.) the γ-rule. For Transformer architectures we implemented the 0^+-rule for all layers. However, for the Simple3DFormer and the PC Transformer, we had to add custom relevance propagation through the whole model, as the architectures come with several sub-modules such as "local gathering" for the PC Transformer, which are non-trivial to backpropagate through.

Raw Attention We always use the raw attention of the last Transformer block and use bilinear or trilinear interpolation to rescale the attention for image and volume data. For point cloud data, this procedure is more complicated as the PC Transformer projects the embeddings on which the Transformer acts via farthest point sampling and k-nearest neighbor grouping. Thus in each downsampling step, we save which k points are sampled to then use k-nearest neighbor interpolation to cast the attention values for these remaining points back into the input space onto all 1024 original points.

Rollout Attention Same procedure as for Raw Attention but before we interpolate back into the original input space, we use the rollout attention aggregation algorithm over all Transformer modules in the architecture.

LRP Attention As for LRP we use custom relevance backpropagation for the Simple3DFormer and PC Transformer architectures. Based on the relevance scores, we filter the attention of each Transformer module, aggregate the filtered attention with the rollout algorithm, and interpolate the resulting attention back into the input as described for Raw Attention.

E.2 Adaption of evaluation metrics

In this section, we explain how we adapted the evaluation metrics in our framework to seamlessly work with 3D modalities. All metrics were adapted for point cloud (n,d) and volume (x,y,z) dimensions besides classical image dimensions (w,h,c). We neglected the metrics which did not need any further adaption. All metrics leveraging threshold values expect normalized saliency maps on the observation level. Otherwise, thresholds have to be selected per observation.

Pixel Flipping We compute the Area Under the Curve (AUC) to receive a single score. For point cloud data acts on the single coordinates.

Region Perturbation We compute the AUC to receive a single score. Acts on a 3D kernel for volume data and single points for point cloud data. Compute the AUC to receive a single score.

Insertion Use Gaussian noise for 3D data instead of Gaussian blur for images. Inserting single points for point cloud data and voxels for volume data.

Deletion Deletes single points for point cloud data and voxels for volume data. Compute the AUC to receive a single score.

Iterative Removal of Features (IROF) Compute the Area Over the Curve (AOC) to receive a single score. We use 3D Slic for volume segmentation and KMeans clustering with fixed $k = 16$ clusters for point cloud segmentation. $k = 16$ was determined by visual inspection. See exemplary visualization in Figure 6.

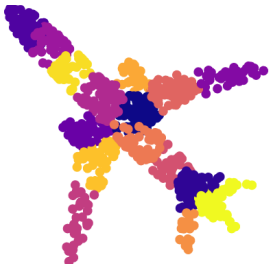


Figure 6: Example of KMeans clustering for point cloud data with k=16.

Remove and Debias (ROAD) We use Gaussian noise for 3D modalities. Compute the AUC to receive a single score.

Sufficiency Use the whole set of saliency maps for similarity comparison and not only the batch the metric is applied to (see Appendix K). For distance calculation between saliency maps, we use squared

Euclidean distance for volume data and standardized Euclidean distance for image and point cloud data due to numerical instability.

Continuity We implemented x-axis traversal for volume data along the x-axis with black padding in all dimensions and for the point cloud data by traversing all points along the x-axis position at $(n, d = 0)$ (see Figure 7). As removing points for point cloud data would change the input dimension of the object, we instead map them to the center (0,0,0). We did not observe any OOD behavior by implementing this solution. We use the Pearson Correlation Coefficient (PCC) between traversals to compute a single score.

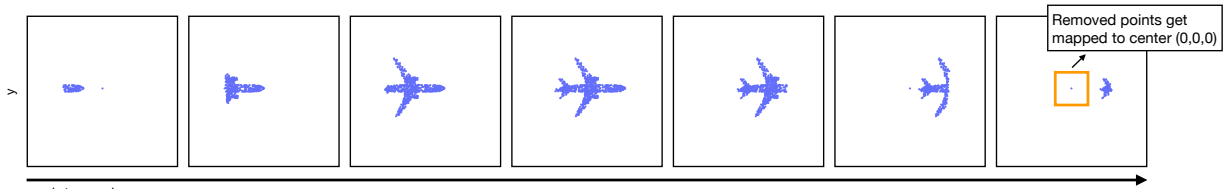


Figure 7: X-axis traversal of point clouds for continuity metric. We can not remove points as this would change the input dimensionality, thus we map them to the center (0,0,0), which is similar to black padding for image and volume data.

Relative Representation Stability We use uniform noise ($U(0, 0.05)$) due to numerical stability as Gaussian noise could generate infinity values.

F Ensuring comparability of results

To ensure fair comparability between XAI methods over all possible experiment set-ups and aggregation levels, we take precautions about the XAI methods, evaluation metrics, and model architectures. Attribution measures the positive or negative contribution of an input feature (e.g. pixel) into the predicted output class of the model. On the contrary, CAM methods only compute positive attribution, and attention highlights all general (or absolute) important input features independent of the output class. However, in practice, attention is only valuable in interpretation if it also highlights features that are used for prediction. New methods such as LA filter the attention to only show such class-relevant attention, and their possible better performance to unfiltered attention can only be shown by evaluating it as positive attribution. Thus we consider only positive attribution for saliency map comparison (also suggested by Zhang et al. (2018)).

Further, we normalize the saliency maps on the observation level as some metrics have nominal thresholds or noise intensities which depend on the scale of saliency maps. As not all metrics compute single scores we have to convert all metrics computing sequences or array of sequences into single scores either via the AUC for Pixel Flipping, Region Perturbation, Selectivity and ROAD, AOC for IROF, or the PCC for SensitivityN and Continuity. All scores are normalized on the metric and dataset level. Score backpropagation-based metrics such as LRP (excluding the 0^+-rule), DS or DLS, and the CAM methods expect non-negative activation outputs. Thus, we exchanged before the CNN model training all GeLU or leakyReLU activation functions with standard ReLU functions as they output negative values, biasing the XAI method. For the Transformer architectures, however, we keep all activation functions, as well as the skip connections and patchification, as they are central to the architecture. Their potential effect on different attribution methods is part of the benchmark. For CAM methods on the Transformer architectures, we interpolate the reshaped *absolute* cls token, as saliency maps would otherwise often be empty (also recommended by Chefer et al. (2021)).

G Ranking computation flow chart

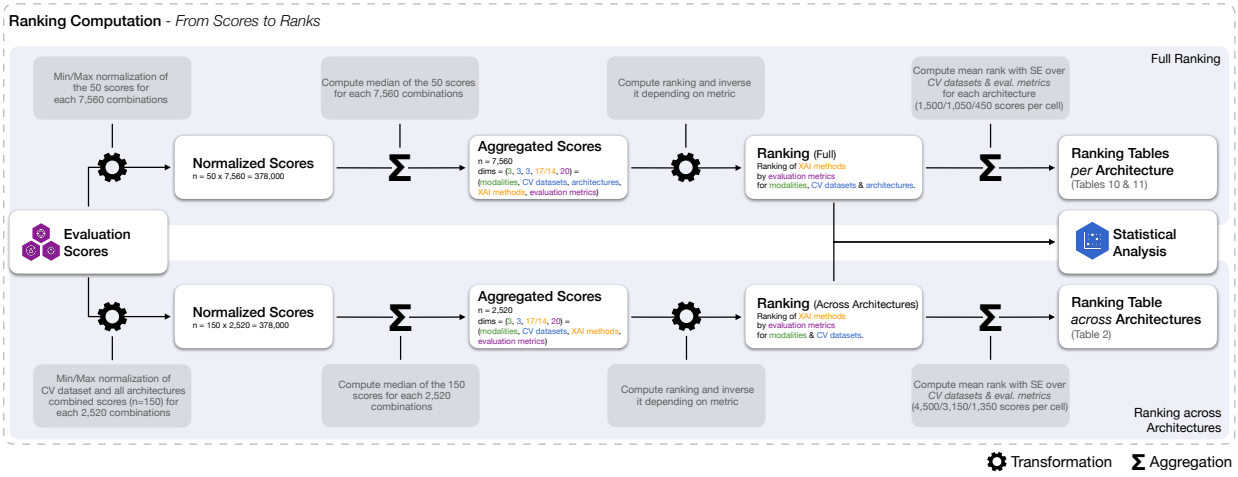


Figure 8: Transformation and aggregation steps from raw evaluation scores to final tables.

Figure 8 shows the transformation and aggregation steps from raw scores to final tables depending if we want to average across architectures or construct tables per architecture. In the calculation of the combinations, it must be taken into account that in the case of the Transformer architectures we have three more XAI methods (attention methods), and in the case of the point cloud modality, we have three fewer XAI methods (excluding CAM methods). In the case of the full ranking, we then have 7,560 combinations of CV datasets, architectures, XAI models, and evaluation metrics based on which we compute 50 scores for each combination, but always the same observations per dataset. If we average across architectures, we have to first normalize the 50 scores per architecture together, as the number of XAI methods differs between CNN and Transformer-based architectures. As we normalize across architectures we end up with 2,520 combinations but 150 scores per combination, which are in total again 378,000 scores. To receive the tables in the final step, we take the mean over the computer vision datasets and evaluation metrics per evaluation criteria, to receive one average rank per XAI method, evaluation criteria, modality, and depending if the ranking is full or across architectures, architecture. Values in the last aggregation step coincide with the number of scores per evaluation criteria, as each of the three criteria contains a different amount of metrics.

H Metric standard deviation for volume and point cloud data

a. Avg. Standard Deviation for Volume Model Architectures and Datasets

	Model Architectures			Utilized CV Datasets		
Evaluation Criteria:	3DResNet18	3DEffNetb0	S3DF	AMN	OMN	VMN
Faithfulness	3.07	3.41	3.61	3.34	3.19	3.55
Robustness	3.47	3.41	3.42	3.54	3.25	3.51
Complexity	0.45	0.48	0.64	0.51	0.63	0.43
Weighted Average	2.82	2.99	3.1	2.99	2.83	3.07

b. Avg. Standard Deviation for PC Model Architectures and Datasets

	Model Architectures			Utilized CV Datasets		
Evaluation Criteria:	PointNet	DGCNN	PCT	CMA	M40	SHN
Faithfulness	2.9	2.97	3.55	3.23	3.07	3.1
Robustness	2.52	2.74	2.91	2.9	2.8	2.48
Complexity	0.72	0.51	0.29	0.42	0.59	0.51
Weighted Average	2.44	2.52	2.84	2.69	2.6	2.49

Table 9: Average metric standard deviation per model architectures and utilized datasets for **a.** volume and **b.** point cloud modalities.

I Ranking-bias through metric subsets

Figure 9: Average Euclidean ranking distance between metric pairs for model architectures and the faithfulness and robustness criteria. More often agreeing metric pairs in their rankings appear more green, and disagreeing pairs more red.

In scenarios of high disagreement among metrics, potential biases could arise when considering only a limited subset of metrics, a concern we have raised in relation to other studies. Such scenarios further underscore our large-scale experimental design as it prevents undetected biases that could result from the selective use of individual metrics, intentional or accidental. Figure 9 shows the Euclidean distance, averaged over all attribution methods and datasets, between the ranking of two metrics for all model architectures and the faithfulness and robustness criteria. We observe that the general mean distance or disagreement between two metrics is around 4 ranks. We further observe around 8-10 outlier pairs for faithfulness and 2-5 for robustness, which have either substantially higher agreement ($\sim$ 2 ranks) or higher disagreement ($\sim$ 5.5 ranks). By first selecting a favorable metric and then pairing it with strongly similar ranking metrics based on this distance matrix (e.g. Faithfulness Estimate with Faithfulness Correlation), it is in theory possible to selectively pair similar ranking metrics to deliberately skew results towards a favorable outcome. However, this is only possible for a very small subset of metrics. Also the likelihood of any such "extremist subgroups" unduly influencing our large-scale study is small. The closest example of such subgroups would be the Relative Stability metrics for robustness.

J Additional ranking tables

J.1 Full ranking table with standard errors

Evaluation Criterion:	Faithfullness									Robustness									Complexity								
Modality:	Image			Volume			Point Cloud			Image			Volume			Point Cloud			Image			Volume			Point Cloud		
Utilized Computer Vision Datasets:	IMN	OCT	R45	AMN	OMN	VMN	CMA	M40	SHN	IMN	OCT	R45	AMN	OMN	VMN	CMA	M40	SHN	IMN	OCT	R45	AMN	OMN	VMN	CMA	M40	SHN
Attribution: OC	10 ±1.97	8 ±1.42	10 ±1.92	9 ±1.29	5 ±0.67	13 ±1.52	11 ±1.07	7 ±1.68	8 ±1.29	11 ±1.96	9 ±1.23	9 ±1.9	11 ±1.77	7 ±1.39	10 ±1.43	9 ±1.67	10 ±1.03	10 ±1.48	8 ±0.67	12 ±0.0	8 ±0.0	7 ±0.33	12 ±1.15	5 ±1.86	11 ±0.33	7 ±1.0	5 ±0.33
Attribution: LIME	13 ±1.29	14 ±1.16	13 ±1.48	12 ±1.22	13 ±0.84	8 ±1.29	5 ±0.86	6 ±1.3	6 ±1.18	14 ±1.53	15 ±1.08	14 ±1.93	11 ±1.63	13 ±1.62	10 ±1.4	11 ±0.95	11 ±1.3	12 ±1.07	7 ±1.33	13 ±0.88	6 ±0.33	12 ±0.33	6 ±0.33	12 ±1.53	9 ±0.88	7 ±0.88	9 ±0.88
Attribution: KS	12 ±1.2	15 ±0.6	13 ±1.27	12 ±1.1	12 ±0.7	8 ±1.52	7 ±0.98	5 ±1.25	7 ±0.89	15 ±1.31	15 ±1.62	15 ±1.35	13 ±1.89	14 ±2.08	13 ±1.81	10 ±0.65	10 ±0.81	9 ±0.87	10 ±0.33	16 ±0.33	12 ±0.67	12 ±1.0	9 ±0.67	12 ±1.2	9 ±0.33	9 ±0.67	8 ±0.0
Attribution: VG	8 ±1.36	12 ±0.89	9 ±1.19	14 ±1.19	12 ±1.56	8 ±1.57	8 ±0.99	8 ±1.07	8 ±1.24	7 ±0.9	7 ±0.94	8 ±1.3	7 ±1.66	7 ±2.24	7 ±2.29	7 ±1.04	8 ±0.81	9 ±1.27	11 ±0.88	14 ±0.67	9 ±0.0	15 ±0.33	11 ±0.88	13 ±0.58	7 ±0.33	8 ±0.58	9 ±0.88
Attribution: IxG	10 ±1.5	8 ±1.65	8 ±1.8	6 ±1.47	6 ±1.39	9 ±1.46	8 ±1.14	6 ±0.97	7 ±1.01	10 ±1.83	10 ±2.04	11 ±1.78	9 ±1.93	10 ±1.62	10 ±2.11	8 ±1.0	9 ±1.44	9 ±1.13	3 ±0.0	4 ±0.33	3 ±0.33	2 ±0.33	2 ±0.33	5 ±3.38	2 ±0.67	4 ±0.0	3 ±0.67
Attribution: GB	7 ±1.87	11 ±1.3	7 ±1.93	9 ±1.54	9 ±1.55	10 ±1.58	5 ±1.28	6 ±1.33	5 ±1.01	6 ±1.74	9 ±1.34	9 ±1.22	10 ±1.95	9 ±2.13	8 ±1.57	10 ±1.48	10 ±0.61	9 ±0.4	5 ±0.0	5 ±0.33	5 ±0.0	9 ±0.33	4 ±1.2	9 ±0.88	4 ±0.0	3 ±0.88	6 ±0.33
Attribution: GC	6 ±0.99	8 ±1.17	10 ±1.22	14 ±0.47	13 ±1.22	12 ±1.08	-	-	-	6 ±2.02	6 ±1.49	5 ±1.36	11 ±1.18	10 ±1.46	11 ±1.43	-	-	-	13 ±0.0	12 ±1.0	14 ±0.0	13 ±1.0	8 ±0.58	9 ±0.58	-	-	-
Attribution: SC	6 ±1.2	8 ±1.56	8 ±1.58	10 ±1.35	11 ±1.04	9 ±1.92	-	-	-	9 ±1.76	9 ±1.66	9 ±1.85	12 ±1.77	12 ±1.12	10 ±2.01	-	-	-	12 ±0.33	7 ±1.33	10 ±0.33	16 ±1.33	14 ±2.33	15 ±1.2	-	-	-
Attribution: C+	8 ±1.0	10 ±1.2	10 ±0.86	11 ±1.84	14 ±0.88	11 ±1.38	-	-	-	5 ±1.52	10 ±1.69	8 ±1.98	10 ±1.97	10 ±1.23	9 ±1.93	-	-	-	15 ±0.58	14 ±0.58	13 ±0.33	16 ±0.33	16 ±1.33	13 ±2.08	-	-	-
Attribution: IG	7 ±1.32	7 ±1.55	8 ±1.72	5 ±1.14	5 ±1.49	9 ±1.58	6 ±1.32	6 ±1.17	8 ±1.38	9 ±1.7	10 ±1.82	8 ±1.94	7 ±1.65	7 ±1.06	7 ±1.51	8 ±1.93	10 ±2.15	8 ±1.85	4 ±0.0	2 ±0.33	3 ±0.67	4 ±0.33	4 ±0.33	4 ±1.45	5 ±0.0	4 ±1.0	3 ±1.0
Attribution: EG	6 ±1.32	6 ±0.84	5 ±1.09	8 ±1.29	6 ±0.96	6 ±1.39	8 ±1.2	8 ±1.02	7 ±1.49	7 ±2.04	5 ±1.83	5 ±1.1	6 ±2.18	5 ±1.97	5 ±1.31	8 ±1.3	8 ±1.21	8 ±1.23	16 ±0.67	17 ±0.33	16 ±0.33	6 ±0.33	12 ±0.88	6 ±0.33	10 ±0.88	11 ±0.0	11 ±0.0
Attribution: DL	9 ±1.12	8 ±1.56	8 ±1.67	6 ±1.2	5 ±0.86	10 ±1.05	8 ±1.24	7 ±1.07	7 ±1.02	9 ±2.07	10 ±2.25	10 ±1.96	7 ±1.82	7 ±1.92	9 ±2.19	8 ±1.15	8 ±1.13	9 ±1.43	2 ±0.0	3 ±0.33	2 ±0.33	3 ±0.33	5 ±0.33	4 ±0.88	2 ±0.0	3 ±0.33	3 ±0.0
Attribution: DLS	6 ±1.27	7 ±0.78	7 ±1.31	5 ±0.7	7 ±0.69	9 ±1.22	8 ±0.84	9 ±0.53	9 ±1.08	8 ±1.19	9 ±1.46	7 ±1.3	10 ±1.49	8 ±1.54	10 ±1.91	9 ±0.97	7 ±0.34	7 ±0.44	7 ±0.67	8 ±1.2	7 ±0.33	5 ±0.0	8 ±1.33	5 ±0.33	6 ±0.0	8 ±0.88	9 ±0.88
Attribution: LRP	13 ±1.22	10 ±1.31	13 ±0.87	10 ±1.53	14 ±1.06	12 ±1.46	7 ±1.15	8 ±0.7	10 ±0.85	9 ±1.7	10 ±1.46	12 ±1.75	12 ±2.59	10 ±2.92	12 ±2.38	3 ±0.43	4 ±0.22	4 ±0.34	1 ±0.0	1 ±0.0	1 ±0.0	2 ±1.0	1 ±0.0	2 ±0.67	2 ±0.67	1 ±0.0	1 ±0.33
Attention: RA	12 ±1.52	7 ±1.7	11 ±1.35	8 ±1.69	7 ±1.27	9 ±1.9	8 ±1.77	9 ±1.8	7 ±1.53	7 ±2.12	3 ±1.32	6 ±1.93	6 ±1.55	7 ±1.98	9 ±1.94	5 ±1.52	4 ±1.64	3 ±1.6	8 ±0.58	8 ±0.67	11 ±0.67	8 ±0.0	13 ±1.45	11 ±4.7	12 ±0.0	12 ±0.0	12 ±0.0
Attention: RoA	11 ±1.4	8 ±2.01	7 ±1.18	8 ±1.75	8 ±1.8	6 ±1.49	9 ±1.38	9 ±1.34	8 ±1.74	9 ±1.6	9 ±2.07	8 ±1.84	6 ±0.95	7 ±1.23	6 ±1.8	5 ±1.95	3 ±1.62	4 ±1.27	16 ±0.58	9 ±0.33	17 ±0.0	11 ±1.53	14 ±0.88	12 ±2.03	14 ±0.33	14 ±0.33	14 ±0.33
Attention: LA	8 ±1.81	7 ±2.14	6 ±1.22	7 ±1.58	7 ±1.98	3 ±0.76	8 ±1.89	11 ±1.57	8 ±1.6	11 ±1.54	7 ±1.89	9 ±1.53	6 ±1.15	8 ±1.31	7 ±1.72	4 ±1.72	3 ±1.51	3 ±1.41	15 ±0.67	8 ±0.0	15 ±0.33	13 ±1.86	16 ±0.67	15 ±1.2	13 ±0.33	13 ±0.33	13 ±0.33

Top 3 | Top 4-6 | Bottom 4-6 | Bottom 3

Table 10: Full ranking table for all XAI methods and CV datasets with standard error (SE).

J.2 Ranking table CNNs only

Evaluation Criteria:	Faithfullness ($n_{(A,M)} = 60$)			Robustness ($n_{(A,M)} = 42$)			Complexity ($n_{(A,M)} = 18$)		
Modality:	Image	Volume	Point Cloud	Image	Volume	Point Cloud	Image	Volume	Point Cloud
Attribution: OC	8.5	5.5	9	11.5	5.5	4.5	9	10.5	7.5
Attribution: LIME	14	9.5	2.5	14	12.5	11	6	8.5	7.5
Attribution: KS	13	9.5	7	13	14	9	13.5	8.5	10
Attribution: VG	11.5	11	9	6	2	2	10	13	9
Attribution: IxG	10	2.5	5	11.5	9.5	7	3.5	1.5	3
Attribution: GB	8.5	7	1	10	8	8	2	5	5.5
Attribution: GC	3	14	-	2	11	-	12	7	-
Attribution: SC	5	8	-	8	9.5	-	7.5	14	-
Attribution: C+	4	13	-	4.5	7	-	11	12	-
Attribution: IG	6	1	5	8	3	4.5	5	4	4
Attribution: EG	1	5.5	11	2	1	10	13.5	10.5	11
Attribution: DL	7	2.5	2.5	8	4	4.5	1	3	2
Attribution: DLS	2	4	9	4.5	5.5	1	7.5	6	5.5
Attribution: LRP	11.5	12	5	2	12.5	4.5	3.5	1.5	1

Per modality: Top 1 | Top 2-4 | Bottom 2-4 | Bottom 1

Table 11: Ranking of the average rank over CNN architectures, datasets, and all evaluation metrics of the respective criteria, for each XAI method and modality (i.e. the rank of OC on image is based on $3*2*10 = 60$ ranks). Coloring coincides with top and bottom positions as no attention methods can be applied to CNN architectures.

J.3 Ranking table Transformer only

Evaluation Criteria:		Faithfullness ($n_{(A,M)} = 30$)			Robustness ($n_{(A,M)} = 21$)			Complexity ($n_{(A,M)} = 9$)		
	Modality:	Image	Volume	Point Cloud	Image	Volume	Point Cloud	Image	Volume	Point Cloud
Attribution	OC	13	9	11.5	12.5	12	13	11.5	8.5	6.5
	LIME	16.5	15	2	15.5	16	14	9	8.5	5
	KS	15	14	4.5	17	17	10	17	12.5	9
	VG	11.5	11	8.5	1.5	2	7	13.5	10	6.5
	IxG	9.5	3	1	7.5	10	10	1	1	3
	GB	5	5	3	10	13	12	4.5	5	4
	GC	9.5	16.5	-	12.5	10	-	10	14	-
	SC	6.5	11	-	6	15	-	6	16.5	-
	C+	14	16.5	-	15.5	10	-	13.5	16.5	-
	IG	6.5	2	6	5	5.5	8	3	3	2
	EG	1	1	7	1.5	1	6	15.5	7	12
	DL	11.5	4	4.5	10	7.5	10	4.5	2	1
	DLS	3	6.5	8.5	3	7.5	5	7	6	8
	LRP	16.5	13	13.5	14	14	4	2	4	10.5
Attention	RA	8	11	10	4	3.5	3	8	12.5	10.5
	RoA	4	6.5	11.5	10	3.5	2	15.5	11	14
	LA	2	8	13.5	7.5	5.5	1	11.5	15	13

Per modality: Top 1 | Top 2-4 | Bottom 2-4 | Bottom 1

Table 12: Ranking of the average rank over Transformer architectures, datasets, and all evaluation metrics of the respective criteria, for each XAI method and modality (i.e. the rank of OC on image is based on $3 * 1 * 10 = 30$ ranks). Coloring coincides with top and bottom positions.

K Shortcomings of evaluation metrics in practice

While all metrics are theoretically very well founded, we observed for some metrics shortcomings in applications:

Casting saliency maps from 64-bit to 32 or 16-bit to save disk space in such large evaluations is not recommended, as our experiments showed that even 32-bit precision can lead to numerical instability, resulting in all-zero saliency maps and *nan* or *inf* evaluation scores.

Sufficiency evaluates the likelihood that observations with the same saliency maps also share the same prediction label. In practice, this requires several saliency maps from observation with the same prediction label. While this works well on datasets with a small number of labels and balanced sampling, for datasets like IMN with 1000 labels, the probability is almost zero that at least 5-10 sampled observations in a set of sizes 50 or 100 have the same label (see Appendix N).

Sequence outputting metrics that alter the input space, such as Pixel Flipping, Region Perturbation, or ROAD, are only limited suitable for binary prediction tasks. When the input object is too noisy/perturbed to predict accurately, the probability for each class is 0.5 resulting in sequences converging against 0.5 and not 0. While the resulting AUC (or AOC in the case of Region Perturbation) can be compared between XAI methods within this task, between tasks the AUC would be biased as the area for the binary task would always be larger.

ROAD scores are arrays of binary sequences which are averaged to one sequence. The amount of noise has to be carefully tuned (also depending on the underlying model) as otherwise, all binary sequences in the array are only 0 or 1.

Local Lipschitz Estimate approximates the Lipschitz smoothness through several forward passes of a batch of observations. In application, this results in a large amount of RAM used (depending on modality) if the approximation should be stable. While the computation is relatively fast on a GPU, stable approximations exceed 40GB of VRAM by far and have to be partitioned. For the Transformer architectures, computation on the CPU for our amount of data was too slow to be feasible.

Effective complexity uses a nominal threshold value to determine attributed features. Even through normalization of the saliency maps, the threshold value can have a large effect on the results, differing between observations, and we would suggest tuning it per dataset.

IROF superpixel segmentation can result in very defined or binary structures such as in the AMN dataset in only two superpixels (object and background), ignoring finer structures.

As elaborated, all complexity metrics flatten the input object treating it as a vector and ignoring spatial dependencies.

L Ranking distance between model architectures

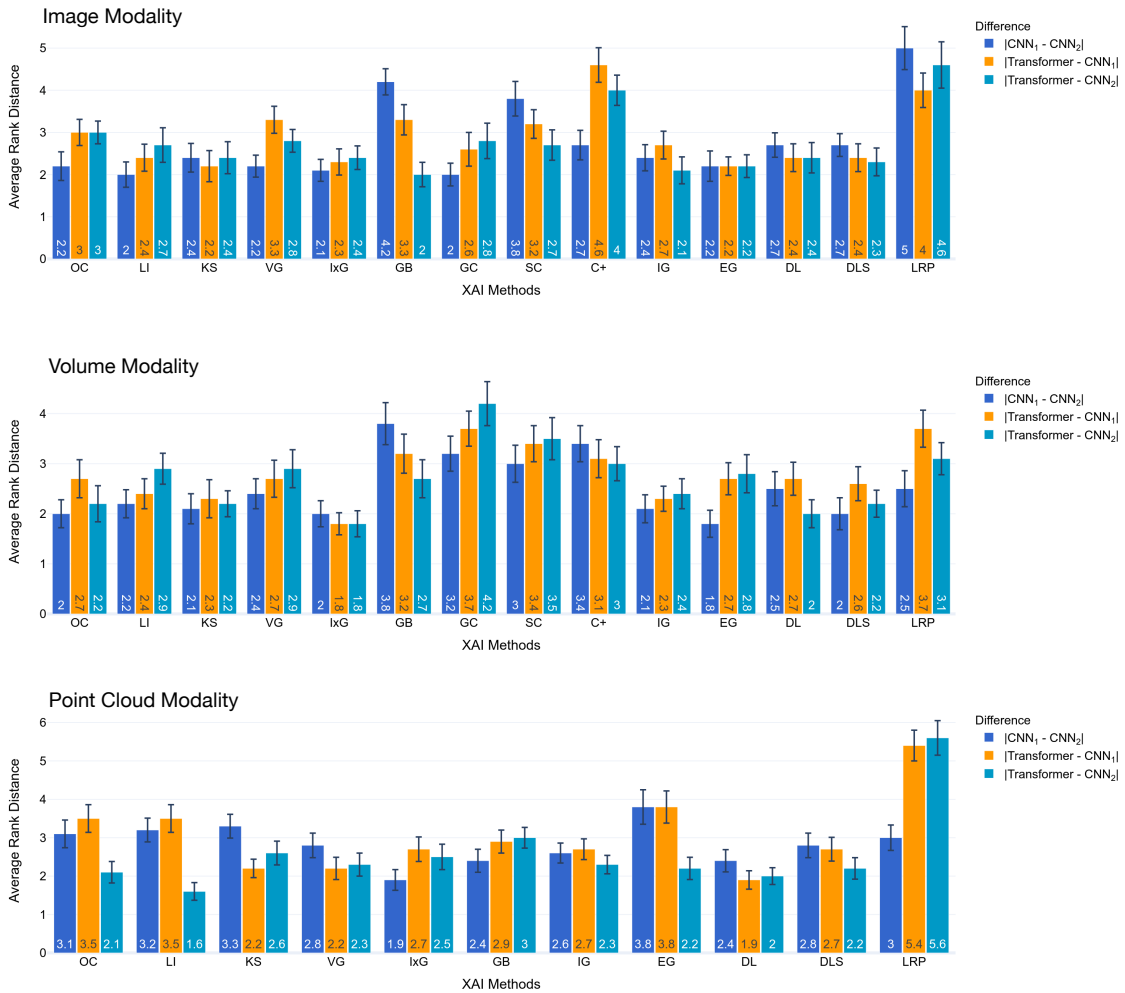


Figure 10: Average distance between ranks of XAI methods on different model architectures for all modalities.

M Differences in ranking order between model architectures

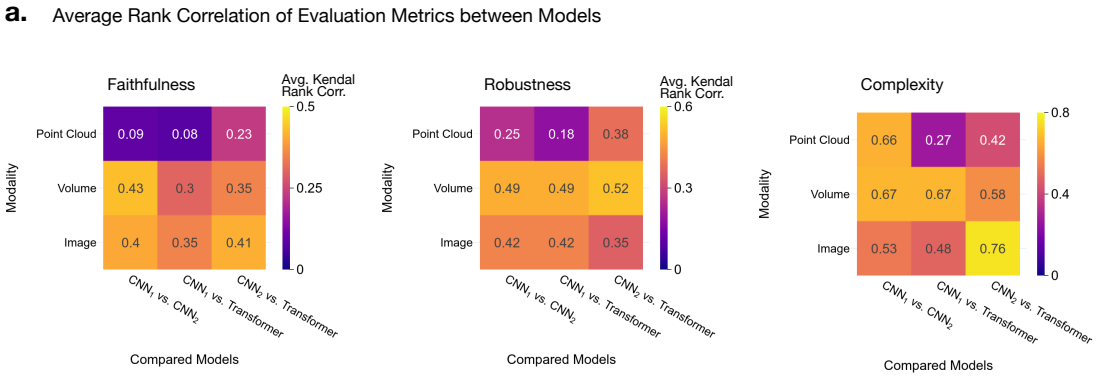


Figure 11: Kendall's-τ rank correlation between model architectures averaged over datasets and faithfulness criteria.

We compare the difference in faithfulness rankings of attribution methods between CNN and Transformer architectures, as biased methods should be less faithful to the model. To this end, we compute the Kendals-τ rank correlation between each of the three architectures per dataset and compute their average correlation per modality (see Figure 11). We observe a positive correlation between all rankings. For the point cloud modality, however, the correlation is significantly lower than for the other two modalities, indicating less similar rankings between model architectures. For volume and image modality, the similarity between CNN architectures is generally higher.

N Score distributions of evaluation metrics for all datasets

N.1 Image modality

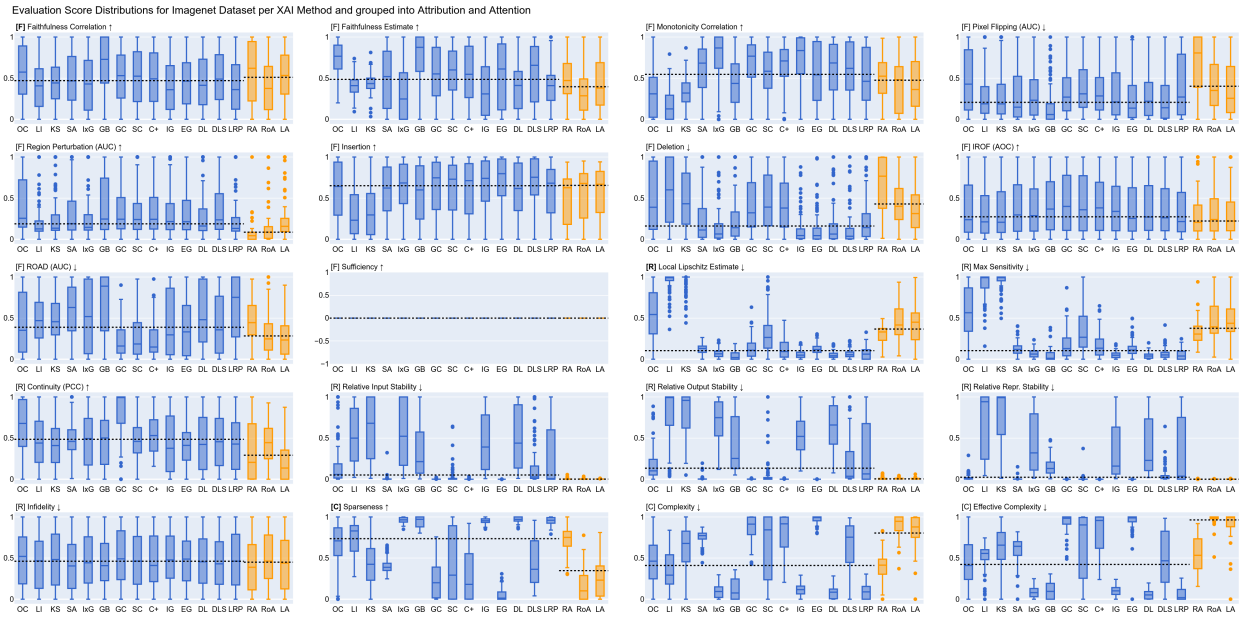


Figure 12: Score distributions of all evaluation metrics for each XAI method. Scores are normalized also for the Continuity PCC, as a negative correlation is worse than no correlation.

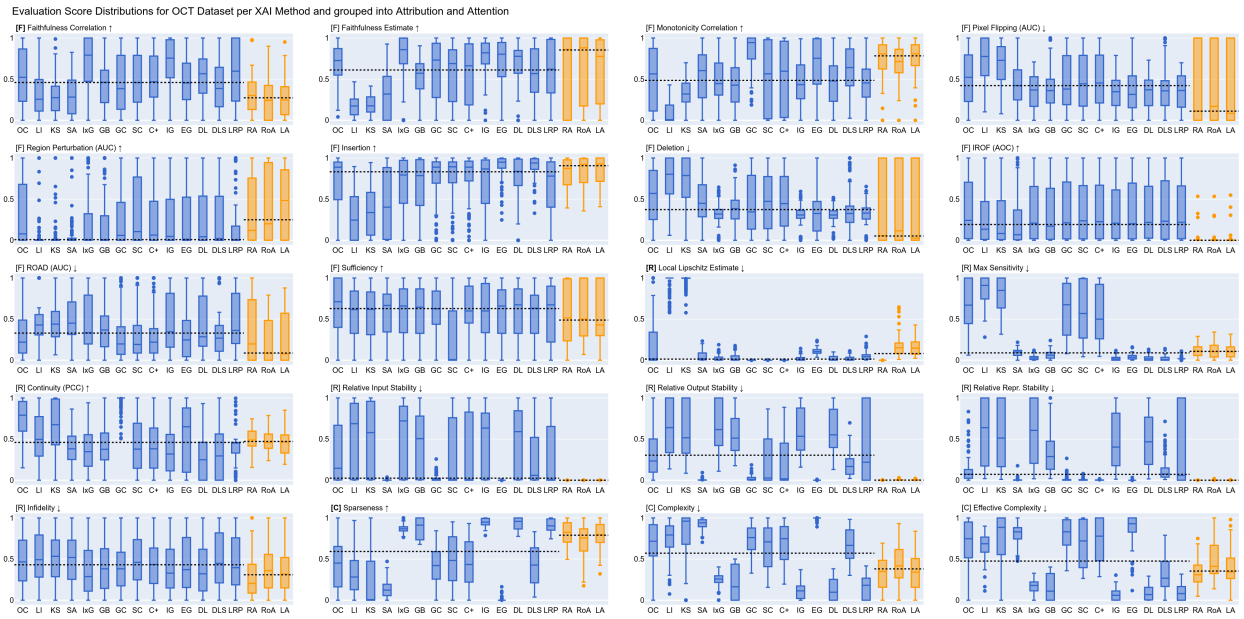


Figure 13: Score distributions of all evaluation metrics for each XAI method. Scores are normalized also for the Continuity PCC, as a negative correlation is worse than no correlation.

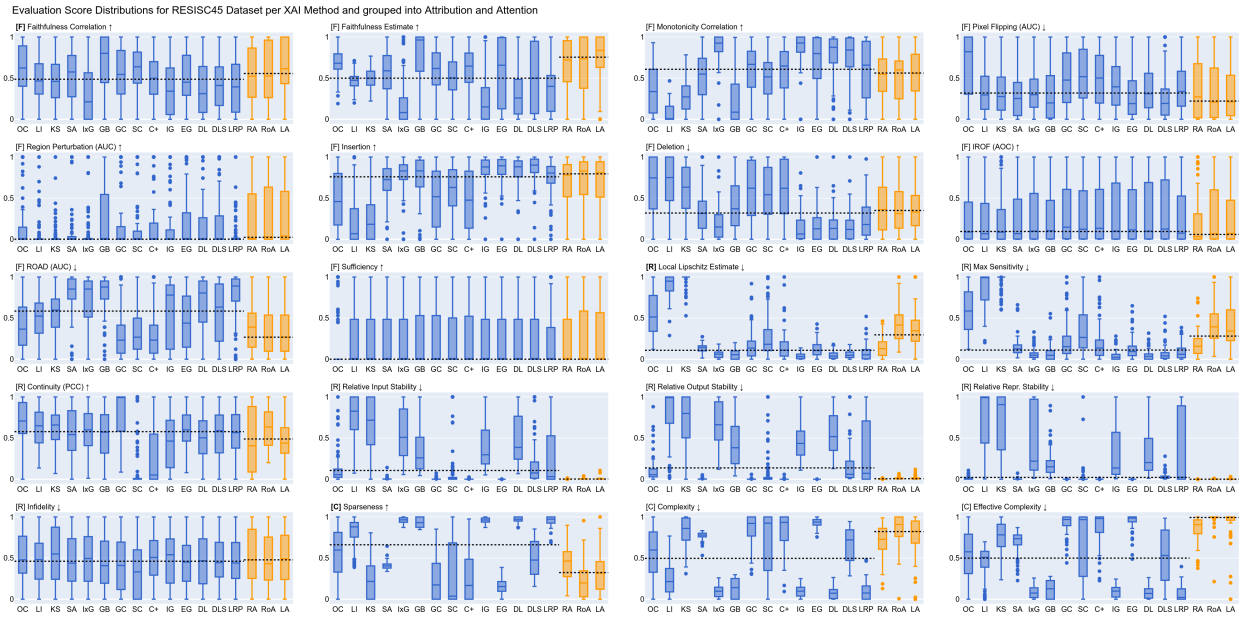


Figure 14: Score distributions of all evaluation metrics for each XAI method. Scores are normalized also for the Continuity PCC, as a negative correlation is worse than no correlation.

N.2 Volume modality

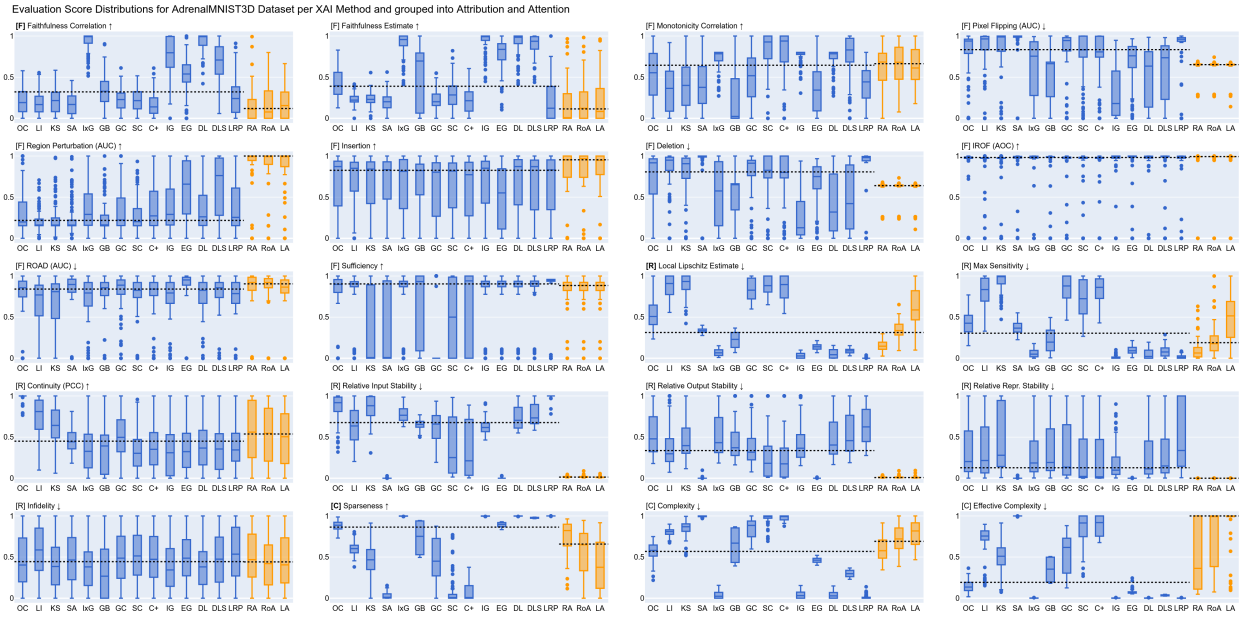


Figure 15: Score distributions of all evaluation metrics for each XAI method. Scores are normalized also for the Continuity PCC, as a negative correlation is worse than no correlation.

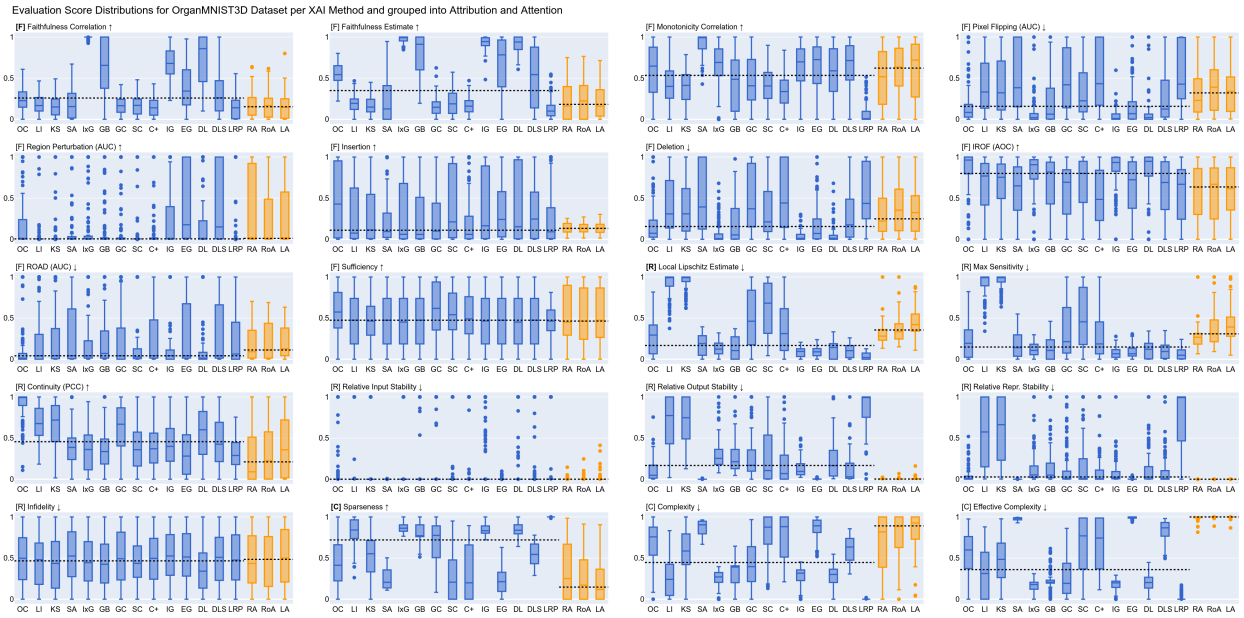


Figure 16: Score distributions of all evaluation metrics for each XAI method. Scores are normalized also for the Continuity PCC, as a negative correlation is worse than no correlation.

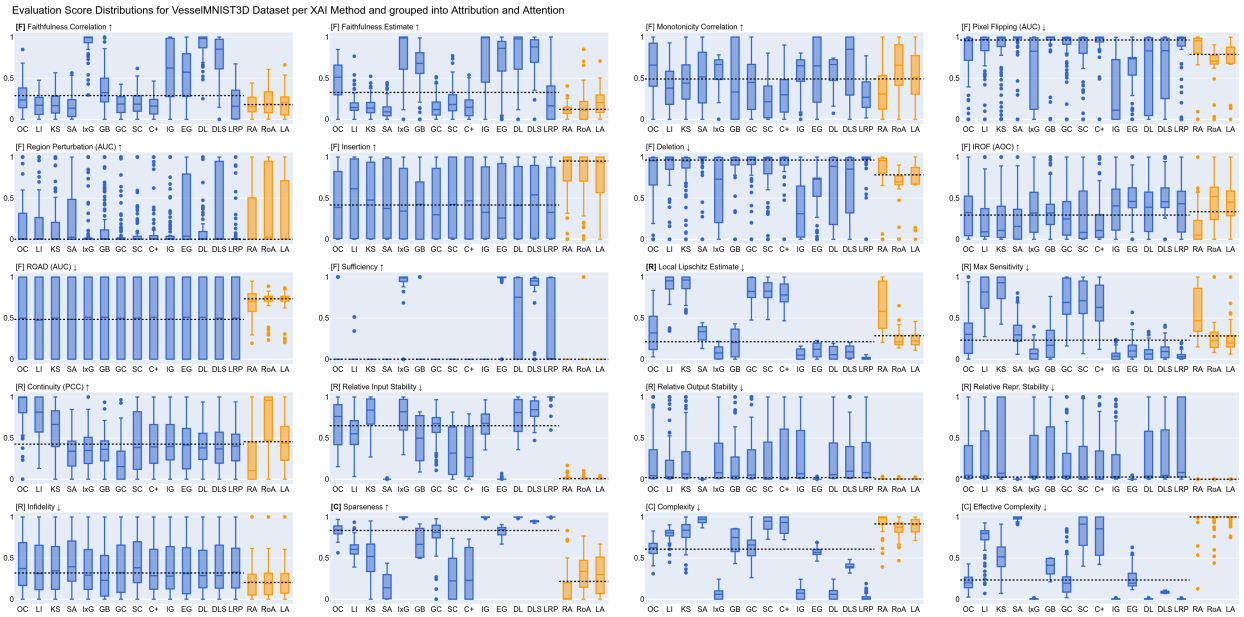


Figure 17: Score distributions of all evaluation metrics for each XAI method. Scores are normalized also for the Continuity PCC, as a negative correlation is worse than no correlation.

N.3 Point cloud modality

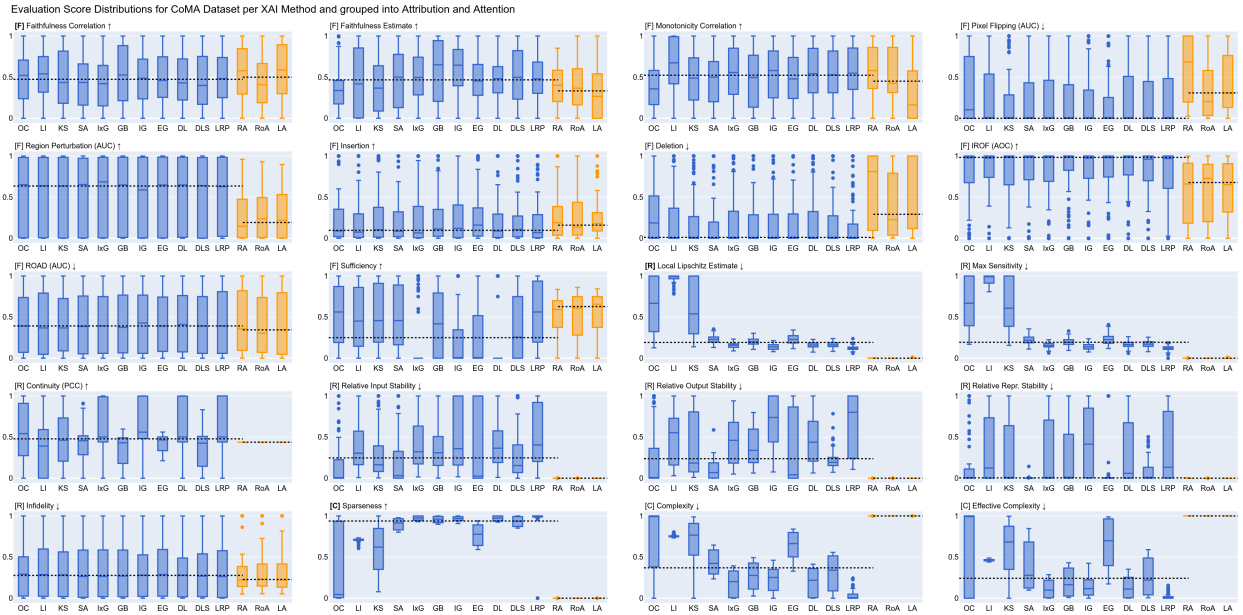


Figure 18: Score distributions of all evaluation metrics for each XAI method. Scores are normalized also for the Continuity PCC, as a negative correlation is worse than no correlation.

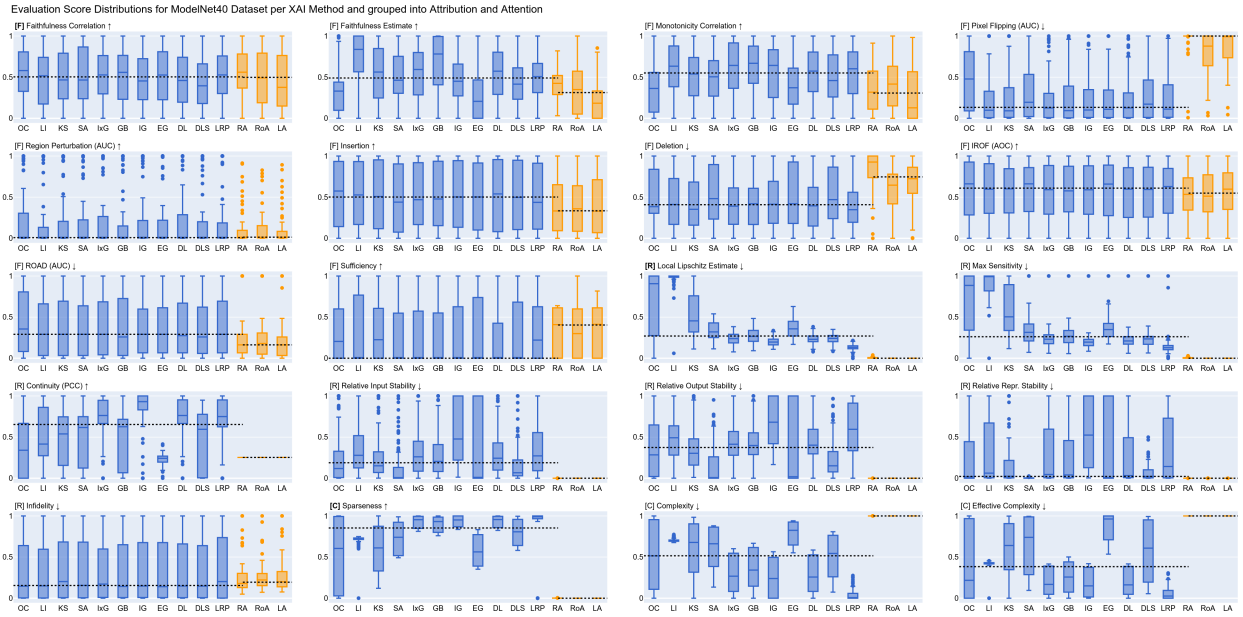


Figure 19: Score distributions of all evaluation metrics for each XAI method. Scores are normalized also for the Continuity PCC, as a negative correlation is worse than no correlation.

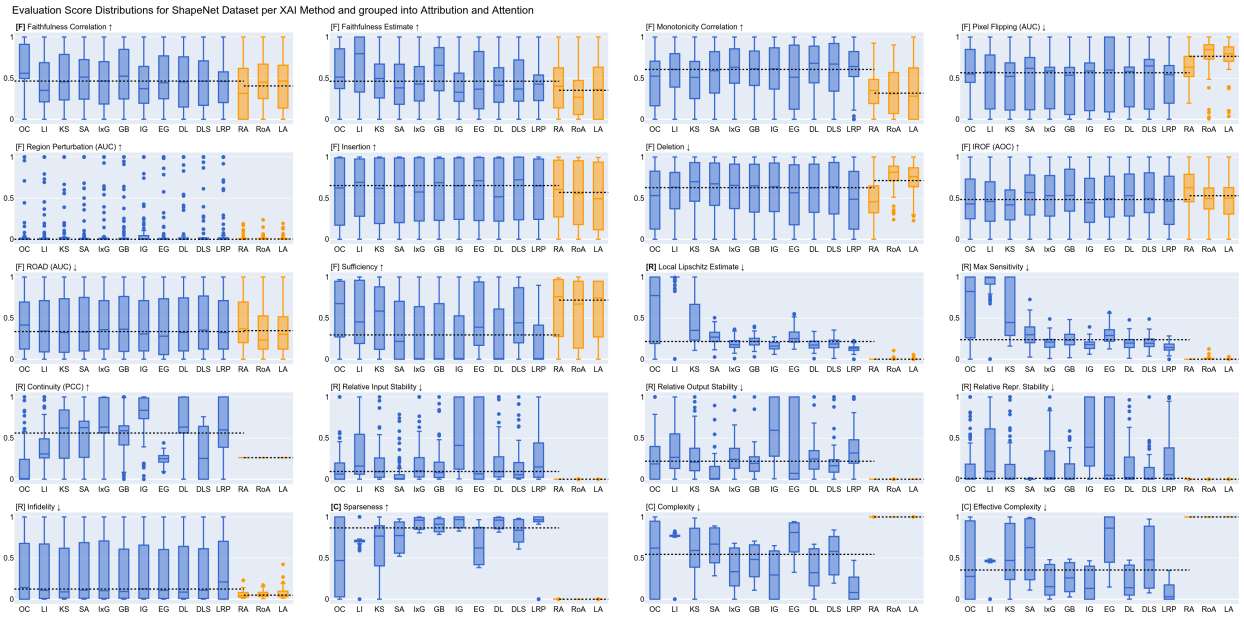


Figure 20: Score distributions of all evaluation metrics for each XAI method. Scores are normalized also for the Continuity PCC, as a negative correlation is worse than no correlation.