# The Price of Freedom: An Adversarial Attack on Interpretability Evaluation

**Kristoffer Wickstrøm**[1]† **Marina Höhne**[2,4,6] **Anna Hedström**[2,3,5]
[1]Department of Physics and Technology, UiT The Arctic University of Norway
[2]UMI Lab, ATB Potsdam, Germany [3]TU Berlin, Germany
[4]BIFOLD, Berlin, Germany [5] Fraunhofer HHI, Berlin, Germany
[6] University of Potsdam, Potsdam, Germany
† corresponding authors:
{kwi030@uit.no}

## Abstract

The absence of ground truth explanation labels poses a key challenge for quantitative evaluation in interpretable AI (IAI), particularly when evaluation methods involve numerous user-specified hyperparameters. Without ground truth, optimising hyperparameter selection is difficult, often leading researchers to make choices based on similar studies, which offers considerable flexibility. We show how this flexibility can be exploited to manipulate evaluation outcomes by framing it as an adversarial attack where minor hyperparameter adjustments lead to significant changes in results. Our experiments demonstrate substantial variations in evaluation outcomes across multiple datasets, explanation methods, and models. To counteract this, we propose a ranking-based mitigation strategy that enhances robustness against such manipulations. This work underscores the challenges of reliable evaluation in IAI. Code is available at `https://github.com/Wickstrom/quantitative-IAI-manipulation`.

## 1 Introduction

Interpretability in artificial intelligence (IAI) is crucial for ensuring trustworthiness (7; 17; 23; 20), especially in finance, law and healthcare, where numerous interpretability methods and evaluation metrics have been proposed (56; 50; 66; 65; 21; 44; 9; 22). While these advancements are helpful, it has resulted in significant disagreement and practitioner confusion regarding the best approaches for different problem settings (45; 35; 13; 36; 19; 31; 30). Quantitative evaluation of IAI methods has advanced, but the variety of methods and metrics makes it difficult for researchers to make informed choices (37; 2; 14).

A major challenge in IAI evaluation is the lack of ground truth explanation labels, leading to the use of proxy metrics to estimate explanation quality like faithfulness (55; 4; 52; 25; 10; 53), complexity (51; 10; 24), or robustness (49; 3; 68; 25). These evaluations depend heavily on hyperparameter choices, which are often ad-hoc and data-dependent. Faithfulness metrics (7; 58), for example, rely on how model behaviour changes with different input manipulations, but they are highly sensitive to hyperparameter choices such as baseline choice and perturbation order (55; 16; 15; 54; 64; 13; 48; 28). How do we mask out pixels and how large should the masks be? Since exhaustive hyperparameter search is impractical, researchers typically make subjective choices based on prior studies or accessibility (45; 35), introducing flexibility that can impact evaluation reliability. We urgently need to investigate the impact of hyperparameter choice on evaluations, making it an important area of research.

Interpretable AI: Past, Present and Future Workshop at NeurIPS 2024

| IAI METHOD | FAITHFULNESS SCORE (↓) |
|---|---|
| LRP | 25.19 |
| SALIENCY | **20.23** |
| KERNEL SHAP | 23.94 |

| IAI METHOD | FAITHFULNESS SCORE (↓) |
|---|---|
| LRP | **19.31** |
| SALIENCY | 22.96 |
| KERNEL SHAP | 24.87 |

Table 1: Faithfulness comparison of IAI methods on MNIST before (left) and after manipulation (right). The difference lies in the perturbation method used: uniform noise vs. blurring.

In this work, we demonstrate how the flexibility in IAI evaluation can be exploited to manipulate evaluation outcomes. Even minor changes to commonly used hyperparameters can significantly alter faithfulness evaluation results. As illustrated in Tab. 1, standard IAI methods show substantial differences in outcomes due to slight hyperparameter adjustments. We propose framing these adjustments as an optimisation problem, allowing for a manipulation of either a single IAI method's evaluation or the joint evaluation of multiple methods. Our contributions are:

- **C1** Two manipulation methods, one method-specific that can increase the evaluation score for a specific IAI method, entitled *intra-manipulation*, and one holistic that manipulates the quantitative comparison of several IAI methods, entitled *inter-manipulation*.
- **C3** A comprehensive experimental analysis on manipulation of faithfulness evaluation, demonstrating that the evaluation outcome can significantly change after manipulation.
- **C4** Towards improving the robustness of the quantitative IAI evaluation, we propose Mean Resilience Rank (MRR), a ranking-based procedure to reduce the sensitivity to manipulation.

Our findings carry significant importance for the IAI community. Quantitative evaluation is crucial to provide objective measurements of explanation quality, which can be used to select an appropriate method for a particular task (32). If evaluations can be easily manipulated, it reduces the trustworthiness of method selection. Therefore, our findings highlighting the issue of manipulation with mitigating solutions are of critical importance for the IAI community.

## 2 Manipulating IAI Evaluation

First, we present the core concepts and notation used in the work. Then, we propose two ways to manipulate IAI evaluation methods. Related works are found in Appendix A.1.

**Preliminaries** Let the input to a black-box classifier $f$ be denoted as $\mathbf{x} \in \mathbb{R}^d$ and the output of the classifier as $f(\mathbf{x}) = \hat{y}$. Local explanation methods (7; 61; 57; 18) interpret the decision of $f$ by attributing an importance score to each component of $\mathbf{x}$. We denote the explanation of $f$ for a given class $y$ as $\mathbf{e} \in \mathbb{R}^d$. Here, we present a generalized formulation of quantitative IAI evaluation to illustrate the static input parameters and tunable hyperparameters. We assume an evaluation function $F \rightarrow \mathbb{R}$ on the form:

$$F(f, \mathbf{x}, \mathbf{e}, a, b, c) = s. \tag{1}$$

Here, $f$, $\mathbf{x}$, and $\mathbf{e}$ are input parameters provided by the user, while $a$, $b$, and $c$ are hyperparameters that must be determined by the user. The output of the evaluation is represented by $s$, which is a scalar indicating the performance of the particular explanation. We keep the hyperparameters $a$, $b$, and $c$ general for the sake of clarity. But note that there could be more or less hyperparameters and they can take many different forms (*e.g.,* a number or a function), depending on the particular test and the data in question.

### 2.1 Manipulation methods

We introduce our manipulation strategies for altering IAI evaluation outcomes through small hyperparameter adjustments. This approach is motivated by multiple accepted hyperparameters for a given IAI evaluation method (see Sec. 4). For example, in faithfulness evaluations (7; 55; 4; 52; 25; 10), perturbing input pixels is a key step, with numerous choices available, making the selection of the appropriate one challenging (58; 54; 13). Evaluating multiple methods is computationally intensive, and without ground truth explanations, determining the best method is infeasible. Consequently,

practitioners often use a single perturbation method (49; 3; 11). However, as shown in Tab. 1, even slight hyperparameter changes can significantly impact the evaluation. Those aware of this sensitivity can potentially exploit it, which motivates our manipulation strategy.

**Intra-manipulation** First, we propose to focus on manipulating the evaluation outcome for a single IAI method, which we refer to as *intra-manipulation*.

**Definition 1** (Intra-Manipulation)**.** *Given an evaluation function $F$, an input sample $\mathbf{x}$, an explanation $\mathbf{e}$, hyperparameters $a$, $b$, and $c$, and a feasible set of hyperparameters $A_a^*$ for the hyperparameter $a$, the intra-manipulation method solves the following optimization problem to determine the hyperparameter $a$, which maximizes the evaluation score of $F$:*

$$\begin{aligned} \underset{a}{\textit{maximize}} \quad & F(f, \mathbf{x}, \mathbf{e}, a, b, c) \\ \textit{subject to} \quad & a \in A_a^*. \end{aligned}$$

Def. 1 defines an optimization problem where the goal is to find hyperparameters that maximize the evaluation outcome, constrained to lie within a feasible set ($A_a^*$ in this case) for the hyperparameters in questions. Determining this feasible set requires the researcher's judgement and an understanding of the specific IAI methods subject to manipulation. Fundamentally, this set depends on the learned functional response of the model. Sec. 2.2 further explains how to determine the feasible set. If the feasible set is large, Def. 1 can be solved through suitable optimization techniques. If the feasible set is small, an exhaustive search can be performed. Also note that Def. 1 can also be extended to optimise multiple hyperparameters, such as maximising both $a$ and $b$.

**Inter-manipulation** While Def. 1 improves the evaluation outcome of a single IAI method, altering the evaluation outcomes of multiple IAI methods simultaneously may be desirable. Our next approach, called *inter-manipulation*, involves jointly manipulating the evaluation of several IAI methods.

**Definition 2** (Inter-Manipulation)**.** *Given an evaluation function $F$, an input sample $\mathbf{x}$, a set of explanations $\{\mathbf{e}_1, \cdots, \mathbf{e}_M\}$ from $M$ different IAI methods, hyperparameters $a$, $b$, and $c$, and a feasible set of hyperparameters $A_a^*$ for the hyperparameter $a$, the inter-manipulation method solves the optimization problem to determine the hyperparameter $a$, maximizing the following objective:*

$$\begin{aligned} \underset{a}{\textit{maximize}} \quad & F(f, \mathbf{x}, \mathbf{e}_m, a, b, c) - \sum_{m' \neq m} F(f, \mathbf{x}, \mathbf{e}_{m'}, a, b, c) \\ \textit{subject to} \quad & a \in A_a^* \end{aligned}.$$

Here, $\mathbf{e}_m$ is the explanation from the IAI method we aim to improve, called the *focus method*. The explanation from a *non-focus method* is denoted as $\mathbf{e}'_m$, which we aim to degrade. The optimization problem presented in Def. 2 is more complex compared to Def. 1 due to the interplay between the different IAI methods. For example, the optimal solution could be found by a combination of increasing the performance of the focus-method while simultaneously decreasing the performance of the *non-focus methods*. Similarly, as Def. 1, the optimization problem can be solved in several ways (*e.g.,* Bayesian optimization) and can be extended to include several hyperparameters.

### 2.2 Manipulation Example: Faithfulness Evaluation

Some IAI evaluation methods are more prone to manipulation than others. For example, localisation metrics, which assess if an explanation falls within a region of interest, typically have only 1 or even 0 hyperparameters to set (63; 5), making them harder to manipulate. In contrast, faithfulness metrics (11; 3; 51) often involve at least 3 hyperparameters, if not more, and are among the most popular IAI evaluation methods (7; 55; 4; 52; 25; 10). Thus, manipulating faithfulness metrics is of significant interest. The next section provides an overview of key components in faithfulness evaluation.

**The fundamental components of faithfulness** Faithfulness measures to what extent explanations follow the predictive behaviour of the model by iteratively perturbing the input and monitoring the corresponding change in the output of the model. Explaining classification decisions is one of the most common use cases of IAI, and thus is our primary investigation focus. T this end, ket $S$ denote

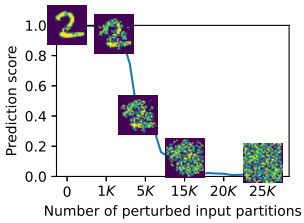


Figure 1: (a)

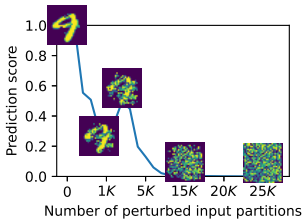


Figure 2: (b)

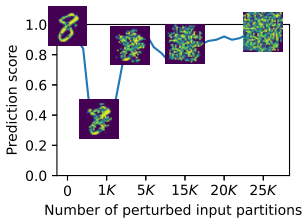


Figure 3: (c)

Figure 4: Example of possible faithfulness curves for digit classification. The leftmost curve illustrates how an "intuitive" faithfulness curve might look, while the remaining curves show the reality of high variation.

the set of indices $\{1, \cdots, d\}$ for each element in the input sample $\mathbf{x} \in \mathbb{R}^d$. Partition $S$ into $K$ sets $S_1, \cdots, S_K$ of equal cardinality $C$ and arranged such that

$$\sum_{i \in S_1} e_i \geq \cdots \geq \sum_{i \in S_K} e_i. \tag{2}$$

For convenient notation, we define the sum of attributions for one partition as

$$\tilde{e}_{S_k} = \sum_{i \in S_k} e_i. \tag{3}$$

The indices are ranked according to the input features with the highest descending importance and is perturbed accordingly. Note that some metrics sort the indices in an ascending fashion (6; 51; 4) and some perturb the input randomly (10), but the general approach in faithfulness metrics is to perturb the inputs according to the rank (55; 54; 52; 3). Let $\mathbf{x}_{S_1}$ denote a perturbed version of $\mathbf{x}$, where all $x_i$ for $i \in S_1$ are replaced by some baseline perturbation function $g_p$. We denote the output of the classifier based on $\mathbf{x}_{S_1}$ as $\hat{y}_{S_1}$. For $\mathbf{x}_{S_2}$, all $x_i$ for $i \in S_1 \cup S_2$ are perturbed. In general, $\mathbf{x}_{S_i}$ will have all have the indices in all sets up to set $S_i$ replaced by the baseline perturbation function.

**Illustrating the faithfulness curve** Based on the $K$ partitions of $S$, a set of progressively more perturbed inputs can be created, *i.e.,* $\{\mathbf{x}_{S_1}, \cdots, \mathbf{x}_{S_K}\}$. Each of the perturbed inputs is classified, which gives a set of model outputs $\{\hat{y}_{S_1}, \cdots, \hat{y}_{S_K}\}$. These model outputs are the fundamental components for faithfulness evaluation in IAI. The rationale is that a good explanation should remove the essential parts of an input first, which should lead to a steep drop in the classification score. A poor explanation will remove parts that are not important, which will allow the classification score to stay high. Fig. 1 shows an example where the classifier behaves as expected, with a sharp drop in accuracy when the important parts of the input are removed. To compare two explanations, one can inspect a plot such as in Fig. 1 and see which explanation has the sharpest drop in classification score. However, such a visual approach has many limitations. First, we generally would like to compare explanations across many samples to get a reliable estimate of how they perform. Inspecting numerous such plots is cumbersome, and the curves can look different for different visual objects in classification, which makes comparison challenging. Also, real-world data is not always as well-behaved as the plot shown in Fig. 1, as illustrated in Fig. 2 and Fig. 3. Another important aspect is ensuring that the curve is a genuine depiction of explanation quality and not an out-of-distribution (OOD) response (38; 54).

**Hyperparameters in Faithfulness Metrics** In faithfulness evaluation, several hyperparameters must be determined by the user, many of which are data-dependent, necessitating re-parameterisation for different tasks, and impeding comparability.

**(Size of partition)** The partition size affects how many features are removed and replaced in each step of the faithfulness curve, balancing computational efficiency and resolution. Smaller partitions increase computational load and risk adversarial effects when few features are removed (59), while larger partitions produce coarser curves. Some use the image dimensions as the partition size (37), though other choices exist (66; 7).

**(Perturbation Function)** The function replacing removed features varies based on the data type, such as Gaussian noise or zero values (3; 49). The choice is critical; for example, zeroing out pixels

may work for natural images but not for those with a black background, where it may fail to change the network's output (29). Diverse choices are documented across studies (49; 3; 11; 53).

**(Aggregation Function)** Aggregation of perturbed model outputs into a single score aids comparison. Two common approaches include calculating the AUC of the faithfulness curve (7; 55; 49) or correlating model outputs with attribution sums within each partition (4; 10).

**(Normalization Function)** To make attributions comparable across different methods, normalisation is often required. A simple approach is to standardise using the mean and standard deviation of the attributions, but more sophisticated methods are sometimes used (35; 12).

## 3 Towards Reliable Evaluations with Mean Resilience Rank

To mitigate hyperparameter manipulation, we propose to rank each IAI method for each hyperparameter setting in the feasible set, and average the ranking across the entire set. We refer to this ranking-approach as Mean Resilience Rank (MRR). For this, we assume to evaluate $M$ explanation methods, with a single hyperparameter $a$ in a feasible set $A_a^*$ that can be altered. We denote one element of $A_a^*$ as $a_i$, such that the evaluation outcome for all $M$ methods can be collected in the set:

$$S_F(a_i) = \{F(f, \mathbf{x}, \mathbf{e}_1, a_i, b, c), \cdots, F(f, \mathbf{x}, \mathbf{e}_M, a_i, b, c)\}. \tag{4}$$

Then, we define a function $R(\cdot)$ that takes in a set of scores and outputs a vector with integer elements, where 0 indicates the lowest score within the set and $M-1$ indicates the highest score within the set. Finally, we define the output of the MMR as the following ranking vector:

$$\mathbf{r} = \frac{1}{|A_a^*|} \sum_{a_i \in A_a^*} \frac{R(S_F(a_i))}{M}. \tag{5}$$

For clarity, we have focused on a single hyperparameter, but 5 can easily be extended to several hyperparameters. For evaluation methods where a high value is desirable, a high ranking indicates good performance, and vice versa for evaluation methods where a low value is desirable.

## 4 Experiments Setup

We evaluate our manipulation strategy across numerous datasets, models, and IAI methods, which are described below. We also define the feasible sets used in our manipulation methods.

**Models, Datasets and IAI Methods** We examine several widely used computer vision datasets; MNIST (27), FashionMNIST (67), PneumoniaMNIST (40), and ImageNet (26), and two common deep learning architectures: LeNet (46) and ResNet18 (34). The LeNet is used for classifying MNIST, FashionMNIST, and PneumoniaMNIST, while the Resnet18 is used for classifying ImageNet. For ImageNet, we randomly sample 100 samples to conduct the faithfulness evaluation, for PneumoniaMNIST we use 500 samples, and for the remaining datasets we use 1000 samples. We choose 100 samples for ImageNet because the larger size of these images increases the computational complexity. We choose 500 for PneumoniaMNIST as it does not have 1000 samples in its test set. We investigate the following IAI methods; Layer-wise relevance propagation (LRP) (7), Saliency (50), and KernelSHAP (47) using the `captum` library (43). We have selected these IAI methods for the experimental analysis since they represent common choices in the IAI field.

**Defining the Feasible Set of Hyperparameters for Faithfulness** A critical aspect of the manipulation methods outlined in Sec. 2 is to determine the feasible set of hyperparameters. This requires in-depth knowledge of the family of quantitative metrics that we aim to manipulate. In this work, we focus on the faithfulness family of evaluation metrics and the critical hyperparamters outlines in Sec. 2.2. We focus on a subset of hyperparameters to provide a clear and understandable evaluation of our manipulation strategies. The feasible set of hyperparameters considered in this work are shown in Tab. 2. This selection is based on common choices in the literature for partition size (7; 18; 35; 37; 66), perturbation function (52; 3; 58), and normalization function (11; 12; 35). We consider the aggregation function fixed as AUC aggregation, meaning that a lower faithfulness score is better. Specifically, AUC is computed on the set of perturbed model outputs $\{\hat{y}_{S_1}, \cdots, \hat{y}_{S_K}\}$.

| | MNIST | FASHIONMNIST | PNEUMMNIST | IMAGENET |
|---|---|---|---|---|
| PARTITION SIZE | {14, 28, 56} | {14, 28, 56} | {14, 28, 56} | {112, 224, 448} |
| PERTURBATION: | $\{\mathcal{N}(0,1), \mathcal{U}(0,1), \mathcal{G}(\cdot)\}$ | $\{\mathcal{N}(0,1), \mathcal{U}(0,1), \mathcal{G}(\cdot)\}$ | $\{\mathcal{N}(0,1), \mathcal{U}(0,1), \mathcal{G}(\cdot)\}$ | $\{\mathcal{N}(0,1), \mathcal{U}(0,1), \mathcal{G}(\cdot)\}$ |
| NORMALIZATION | {TRUE, FALSE} | {TRUE, FALSE} | {TRUE, FALSE} | {TRUE, FALSE} |

Table 2: The feasible set of hyperparameter in this work for different datasets. $\mathcal{G}(\cdot)$ denotes Gaussian blurring.

| | | MNIST | | FASHIONMNIST | | PNEUMMNIST | | IMAGENET | |
|---|---|---|---|---|---|---|---|---|---|
| | IAI METHOD | BASE | MANIP. | BASE | MANIP. | BASE | MANIP. | BASE | MANIP. |
| INTRA | LRP | 25.20 | 7.86 | 21.46 | 5.37 | 21.31 | 6.06 | 129.61 | 41.48 |
| | SALIENCY | 20.23 | 6.80 | 15.65 | 4.72 | 23.28 | 4.23 | 124.93 | 37.53 |
| | KERNELSHAP | 23.94 | 8.01 | 18.28 | 4.81 | 22.06 | 4.29 | 128.72 | 40.14 |
| INTER | LRP | 25.20 | 7.86 | 21.46 | 5.37 | 21.31 | 6.06 | 129.61 | 41.48 |
| | SALIENCY | 20.23 | 6.80 | 15.65 | 4.72 | 23.28 | 4.23 | 124.93 | 37.53 |
| | KERNELSHAP | 23.94 | 8.01 | 18.28 | 4.81 | 22.06 | 4.29 | 128.72 | 40.14 |

Table 3: Intra-results (top three rows) towards *LRP* and inter-results (bottom three rows). Lower is better.

## 5 Results

Here we present the results of our proposed inter-manipulation and intra-manipulation. We define a *base* set of hyperparameters from the literature: for MNIST, FashionMNIST, and PneumoniaMNIST, this includes a partition size of 28, uniform noise as perturbations, and no normalization; for ImageNet, it includes a partition size of 224, uniform noise, and no normalization. After applying the manipulations defined in Def.1 and Def. 2, we obtain a *manipulated* set of hyperparameters. Our results are centred around comparing the performance of the *base* set and the *manipulated* set.

**Intra-Manipulation Results** Tab. 3 (top three rows) shows the results of performing the intra-manipulation proposed in Def. 1, where *base* is the score obtained with the selected set of hyperparameters described above and *manipulated* is the score obtained after manipulation. These results show significant potential for altering the evaluation outcome of a single IAI method, with improvements of up to 130% from *base* to *manipulated*. Note that the *manipulated* scores aren't directly comparable since the manipulation is method-specific and hyperparameters can vary. For altering evaluation outcomes across methods, the inter-manipulation described in the next section should be used.

**Inter-Manipulation Results** Tab. 3 (bottom three rows), Tab. A.1, and Tab. A.2 show the results of performing the inter-manipulation proposed in Def. 2, where the scores are manipulated towards LRP, Saliency, and KernelSHAP, respectively. For some tasks, the evaluation outcome can be manipulated such that most of the three methods achieve the best performance. This is particularly apparent for PneumoniaMNIST, where all IAI methods can achieve the best performance after manipulation. For some datasets such as ImageNet, there is less room for manipulation. That said, the evaluation difference between explanation methods can still be reduced and thus make the IAI evaluation findings less conclusive (see *e.g.,* Imagenet results in Tab. A.2).

**Towards Reliable Faithfulness Evaluations** The results in Tab. 3 demonstrate that the evaluation outcome can be manipulated, eroding trust in the results. Here, we display the results of using MRR described in Sec. 3 towards mitigating the potential for manipulation. Tab. 4 displays the results of the ranking procedure, indicating that the top-performing IAI methods vary between datasets. However, when averaged across all datasets, LRP emerges as the top performer, closely followed by KernelSHAP, while Saliency is consistently ranked lower. There is notable variation in scores, further highlighted in Fig. A.1. The benefit of this ranking approach is that there is little room for manipulation since the top-performing methods will have to perform well across numerous hyperparameters and datasets. The downside of this ranking approach is that it requires a significant amount of computation including all methods, hyperparameters, and datasets. Also, while averaging across datasets adds robustness, it can obscure important dataset-specific insights, making it crucial to include dataset-wise rankings so that the reader can get an overview of the evaluation.

| IAI METHOD | MNIST | FASHIONMNIST | PNEUMMNIST | IMAGENET | ALL |
|---|---|---|---|---|---|
| LRP | $\mathbf{0.22 \pm 0.15}$ | $0.33 \pm 0.00$ | $0.21 \pm 0.00$ | $\mathbf{0.26 \pm 0.00}$ | $\mathbf{0.29 \pm 0.14}$ |
| SALIENCY | $0.41 \pm 0.26$ | $0.44 \pm 0.31$ | $0.37 \pm 0.31$ | $0.41 \pm 0.33$ | $0.41 \pm 0.30$ |
| KERNELSHAP | $0.37 \pm 0.33$ | $\mathbf{0.22} \pm 0.31$ | $\mathbf{0.33} \pm 0.27$ | $0.33 \pm 0.06$ | $0.31 \pm 0.31$ |

Table 4: MRR across feasible set for each dataset and across datasets (last column). Lower is better, a rank of 0 is best and 1 is worst. Results show that the top performing method can change significantly between datasets, but when averaging across datasets LRP and KernelSHAP are consistently higher ranked than Saliency.

# 6 Discussion

We have introduced two methods for manipulating the quantitative evaluation of explanation methods: intra-manipulation, which enhances the performance of a single method, and inter-manipulation, which affects comparative analyses of IAI methods. These methods are motivated by the absence of ground truth labels, making hyperparameter selection a challenge. We demonstrate the effectiveness of these manipulation strategies across various vision datasets and IAI methods, showing substantial potential for manipulating faithfulness outcomes. This raises concerns for the IAI community, suggesting that evaluation results cannot always be trusted. For this, we propose a new ranking-based procedure, underscoring the urgent need for reliable evaluation in IAI.

**Limitations and Future work** Our proposed MRR mitigates manipulation but has limitations. Its computational cost increases with more methods and hyperparameters. MRR also requires domain expertise to define the feasible hyperparameter set accurately. If misdefined, it could worsen manipulation issues by expanding hyperparameter choices. As a ranking-based method, MRR's scores are relative and dependent on the explanation methods set, limiting meaningful comparisons across tasks. To address this, we encourage the IAI community to build an open-source database using tools like `Quantus` (37) and `OpenXAI` (2) to standardise and store benchmarking results. For future work, we aim to expand the scope of IAI adversarial manipulation to other families of quantitative measures such as randomisation (1; 36) and robustness (3; 68; 25) which rely on parameters such as segmentation masks and noise perturbation methods, respectively.


# Acknowledgments and Disclosure of Funding

This work was partly funded by the German Ministry for Education and Research (BMBF) through the project Explaining 4.0 (ref. 01IS200551). Additionally, this work was supported by the European Union's Horizon Europe research and innovation programme (EU Horizon Europe) as grant TEMA (101093003); the European Union's Horizon 2020 research and innovation programme (EU Horizon 2020) as grant iToBoS (965221); and the state of Berlin within the innovation support programme ProFIT (IBB) as grant BerDiBa (10174498); and BIFOLD (ref. 01IS18025A and ref. 01IS18037A); and the Investitionsbank Berlin through BerDiBa (grant no. 10174498); and the Research Council of Norway (RCN), through the FRIPRO (grant no. 315029); the Centre for Research-based Innovation funding scheme (Visual Intelligence, grant no. 309439); and the Consortium Partners, RCN IKTPLUSS (grant no. 303514); and the UiT Thematic Initiative "Data-Driven Health Technology".


# References


[1] Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B.: Sanity checks for saliency maps. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. p. 9525–9536. NIPS'18, Curran Associates Inc., Red Hook, NY, USA (2018)

[2] Agarwal, C., Krishna, S., Saxena, E., Pawelczyk, M., Johnson, N., Puri, I., Zitnik, M., Lakkaraju, H.: OpenXAI: Towards a transparent evaluation of model explanations. In: Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2022), `https://openreview.net/forum?id=MU2495w47rz`

[3] Alvarez Melis, D., Jaakkola, T.: Towards robust interpretability with self-explaining neural networks. In: Advances in Neural Information Processing Systems. pp. – (2018)

[4] Ancona, M., Ceolini, E., Öztireli, C., Gross, M.: Towards better understanding of gradient-based attribution methods for deep neural networks. In: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net (2018)

[5] Arras, L., Osman, A., Samek, W.: Clevr-xai: A benchmark dataset for the ground truth evaluation of neural network explanations. Information Fusion **81**, 14–40 (2022). https://doi.org/https://doi.org/10.1016/j.inffus.2021.11.008, https://www.sciencedirect.com/science/article/pii/S1566253521002335

[6] Arya, V., Bellamy, R.K.E., Chen, P., Dhurandhar, A., Hind, M., Hoffman, S.C., Houde, S., Liao, Q.V., Luss, R., Mojsilovic, A., Mourad, S., Pedemonte, P., Raghavendra, R., Richards, J.T., Sattigeri, P., Shanmugam, K., Singh, M., Varshney, K.R., Wei, D., Zhang, Y.: One explanation does not fit all: A toolkit and taxonomy of AI explainability techniques. CoRR **abs/1909.03012** (2019), http://arxiv.org/abs/1909.03012

[7] Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLOS ONE **10**(7), e0130140 (Jul 2015). https://doi.org/10.1371/journal.pone.0130140, https://doi.org/10.1371/journal.pone.0130140

[8] Bansal, N., Agarwal, C., Nguyen, A.: SAM: the sensitivity of attribution methods to hyperparameters. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2020, Seattle, WA, USA, June 14-19, 2020. pp. 11–21. Computer Vision Foundation / IEEE (2020)

[9] Bareeva, D., Ümit Yolcu, G., Hedström, A., Schmolenski, N., Wiegand, T., Samek, W., Lapuschkin, S.: Quanda: An interpretability toolkit for training data attribution evaluation and beyond (2024), https://arxiv.org/abs/2410.07158

[10] Bhatt, U., Weller, A., Moura, J.M.F.: Evaluating and aggregating feature-based model explanations. In: Bessiere, C. (ed.) Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020. pp. 3016–3022. ijcai.org (2020)

[11] Bhatt, U., Weller, A., Moura, J.M.F.: Evaluating and aggregating feature-based model explanations. In: International Joint Conference on Artificial Intelligence. pp. 3016–3022 (2020). https://doi.org/10.24963/ijcai.2020/417

[12] Binder, A., Weber, L., Lapuschkin, S., Montavon, G., Müller, K.R., Samek, W.: Shortcomings of top-down randomization-based sanity checks for evaluations of deep neural network explanations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16143–16152 (2023). https://doi.org/10.1109/CVPR52729.2023.01549, https://doi.org/10.1109/CVPR52729.2023.01549

[13] Blücher, S., Vielhaben, J., Strodthoff, N.: Decoupling pixel flipping and occlusion strategy for consistent xai benchmarks (2024)

[14] Bommer, P., Kretschmer, M., Hedström, A., Bareeva, D., Höhne, M.: Finding the right xai method–a guide for the evaluation and ranking of explainable ai methods in climate science. arxiv. arXiv preprint arXiv:2303.00652 (2023)

[15] Brocki, L., Chung, N.C.: Evaluation of interpretability methods and perturbation artifacts in deep neural networks. CoRR **abs/2203.02928** (2022)

[16] Brunke, L., Agrawal, P., George, N.: Evaluating input perturbation methods for interpreting CNNs and saliency map comparison. In: Computer Vision – ECCV 2020 Workshops, pp. 120–134. Springer International Publishing (2020)

[17] Bykov, K., Deb, M., Grinwald, D., Müller, K.R., Höhne, M.M.C.: Dora: Exploring outlier representations in deep neural networks. arXiv preprint arXiv:2206.04530 (2022)

[18] Bykov, K., Hedström, A., Nakajima, S., Höhne, M.M.: Noisegrad - enhancing explanations by introducing stochasticity to model weights. In: Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022. pp. 6132–6140. AAAI Press (2022)

[19] Bykov, K., Höhne, M.M.C., Creosteanu, A., Müller, K.R., Klauschen, F., Nakajima, S., Kloft, M.: Explaining bayesian neural networks. arXiv preprint arXiv:2108.10346 (2021)

[20] Bykov, K., Kopf, L., Höhne, M.M.C.: Finding spurious correlations with function-semantic contrast analysis. In: World Conference on Explainable Artificial Intelligence. pp. 549–572. Springer (2023)

[21] Bykov, K., Kopf, L., Nakajima, S., Kloft, M., Höhne, M.: Labeling neural representations with inverse recognition. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems. vol. 36, pp. 24804–24828. Curran Associates, Inc. (2023), https://proceedings.neurips.cc/paper_files/paper/2023/file/4e52bbb99690d1e05c7ef7b4c8b3569a-Paper-Conference.pdf

[22] Bykov, K., Kopf, L., Nakajima, S., Kloft, M., Höhne, M.: Labeling neural representations with inverse recognition. Advances in Neural Information Processing Systems **36** (2024)

[23] Bykov, K., Müller, K.R., Höhne, M.M.C.: Mark my words: Dangers of watermarked images in imagenet. In: European Conference on Artificial Intelligence. pp. 426–434. Springer (2023)

[24] Chalasani, P., Chen, J., Chowdhury, A.R., Wu, X., Jha, S.: Concise explanations of neural networks using adversarial training. In: III, H.D., Singh, A. (eds.) Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 119, pp. 1383–1391. PMLR (13–18 Jul 2020), https://proceedings.mlr.press/v119/chalasani20a.html

[25] Dasgupta, S., Frost, N., Moshkovitz, M.: Framework for evaluating faithfulness of local explanations. In: International Conference on Machine Learning. pp. 4794–4815. PMLR (2022)

[26] Deng, J., et al.: Imagenet: A large-scale hierarchical image database. In: Computer Vision and Pattern Recognition. pp. 248–255 (2009)

[27] Deng, L.: The mnist database of handwritten digit images for machine learning research. IEEE Signal Processing Magazine **29**(6), 141–142 (2012)

[28] Dolci, G., Cruciani, F., Galazzo, I.B., Calhoun, V.D., Menegaz, G.: Objective assessment of the bias introduced by baseline signals in XAI attribution methods. In: IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering, MetroXRAINE 2023, Milano, Italy, October 25-27, 2023. pp. 266–271. IEEE (2023)

[29] Fong, R.C., Vedaldi, A.: Interpretable explanations of black boxes by meaningful perturbation. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 3449–3457 (2017). https://doi.org/10.1109/ICCV.2017.371

[30] Gautam, S., Boubekki, A., Hansen, S., Salahuddin, S., Jenssen, R., Höhne, M., Kampffmeyer, M.: Protovae: A trustworthy self-explainable prototypical variational model. Advances in Neural Information Processing Systems **35**, 17940–17952 (2022)

[31] Gautam, S., Höhne, M.M.C., Hansen, S., Jenssen, R., Kampffmeyer, M.: This looks more like that: Enhancing self-explaining models by prototypical relevance propagation. Pattern Recognition **136**, 109172 (2023)

[32] Han, T., Srinivas, S., Lakkaraju, H.: Which explanation should I choose? A function approximation perspective to characterizing post hoc explanations. In: NeurIPS (2022)

[33] Hase, P., Xie, H., Bansal, M.: The out-of-distribution problem in explainability and search methods for feature importance explanations. In: Ranzato, M., Beygelzimer, A., Dauphin, Y.N., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual. pp. 3650–3666 (2021)

[34] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 CVPR. pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90

[35] Hedström, A., Bommer, P.L., Wickstrøm, K.K., Samek, W., Lapuschkin, S., Höhne, M.M.: The meta-evaluation problem in explainable AI: Identifying reliable estimators with metaquantus. Transactions on Machine Learning Research (2023), https://openreview.net/forum?id=j3FK00HyfU

[36] Hedström, A., Weber, L., Lapuschkin, S., Höhne, M.: A fresh look at sanity checks for saliency maps. In: Explainable Artificial Intelligence. pp. 403–420. Springer Nature Switzerland, Cham (2024)

[37] Hedström, A., Weber, L., Krakowczyk, D., Bareeva, D., Motzkus, F., Samek, W., Lapuschkin, S., HÃ¶hne, M.M.C.: Quantus: An explainable ai toolkit for responsible evaluation of neural network explanations and beyond. Journal of Machine Learning Research **24**(34), 1–11 (2023), http://jmlr.org/papers/v24/22-0142.html

[38] Hooker, S., Erhan, D., Kindermans, P.J., Kim, B.: A benchmark for interpretability methods in deep neural networks. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 32. Curran Associates, Inc. (2019), https://proceedings.neurips.cc/paper_files/paper/2019/file/fe4b8556000d0f0cae99daa5c5c5a410-Paper.pdf

[39] Karimi, A.H., Muandet, K., Kornblith, S., Schölkopf, B., Kim, B.: On the relationship between explanation and prediction: A causal view. In: XAI in Action: Past, Present, and Future Applications (2023), https://openreview.net/forum?id=ag1CpSUjPS

[40] Kermany, D.S., Goldbaum, M., Cai, W., Valentim, C.C., Liang, H., Baxter, S.L., McKeown, A., Yang, G., Wu, X., Yan, F., Dong, J., Prasadha, M.K., Pei, J., Ting, M.Y., Zhu, J., Li, C., Hewett, S., Dong, J., Ziyar, I., Shi, A., Zhang, R., Zheng, L., Hou, R., Shi, W., Fu, X., Duan, Y., Huu, V.A., Wen, C., Zhang, E.D., Zhang, C.L., Li, O., Wang, X., Singer, M.A., Sun, X., Xu, J., Tafreshi, A., Lewis, M.A., Xia, H., Zhang, K.: Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell **172**(5), 1122–1131.e9 (2018). https://doi.org/https://doi.org/10.1016/j.cell.2018.02.010, https://www.sciencedirect.com/science/article/pii/S0092867418301545

[41] Kindermans, P.J., Hooker, S., Adebayo, J., Alber, M., Schütt, K.T., Dähne, S., Erhan, D., Kim, B.: The (Un)reliability of Saliency Methods, p. 267–280. Springer International Publishing (2019). https://doi.org/10.1007/978-3-030-28954-6$_1$4

[42] Koenen, N., Wright, M.N.: Toward understanding the disagreement problem in neural network feature attribution. CoRR **abs/2404.11330** (2024)

[43] Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., Melnikov, A., Kliushkina, N., Araya, C., Yan, S., Reblitz-Richardson, O.: Captum: A unified and generic model interpretability library for pytorch (2020)

[44] Kopf, L., Bommer, P.L., Hedström, A., Lapuschkin, S., Höhne, M.M.C., Bykov, K.: Cosy: Evaluating textual explanations of neurons (2024), https://arxiv.org/abs/2405.20331

[45] Krishna, S., Han, T., Gu, A., Pombra, J., Jabbari, S., Wu, S., Lakkaraju, H.: The disagreement problem in explainable machine learning: A practitioner's perspective. CoRR **abs/2202.01602** (2022), https://arxiv.org/abs/2202.01602

[46] Lecun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998). https://doi.org/10.1109/5.726791

[47] Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. p. 4768–4777. NIPS'17, Curran Associates Inc., Red Hook, NY, USA (2017)

[48] Mamalakis, A., Barnes, E.A., Ebert-Uphoff, I.: Carefully choose the baseline: Lessons learned from applying XAI attribution methods for regression tasks in geoscience. CoRR **abs/2208.09473** (2022)

[49] Montavon, G., Samek, W., Müller, K.R.: Methods for interpreting and understanding deep neural networks. Digital Signal Processing **73**, 1–15 (2018). https://doi.org/https://doi.org/10.1016/j.dsp.2017.10.011, https://www.sciencedirect.com/science/article/pii/S1051200417302385

[50] Morch, N., et al.: Visualization of neural networks using saliency maps. In: International Conference on Neural Networks. pp. 2085–2090 (1995)

[51] Nguyen, A., Martínez, M.R.: On quantitative aspects of model interpretability. CoRR **abs/2007.07584** (2020), https://arxiv.org/abs/2007.07584

[52] Rieger, L., Hansen, L.K.: IROF: a low resource evaluation metric for explanation methods. CoRR **abs/2003.08747** (2020), https://arxiv.org/abs/2003.08747

[53] Rong, Y., Leemann, T., Borisov, V., Kasneci, G., Kasneci, E.: A consistent and efficient evaluation strategy for attribution methods. In: Proceedings of the 39th International Conference on Machine Learning. pp. 18770–18795. PMLR (2022)

[54] Rong, Y., Leemann, T., Borisov, V., Kasneci, G., Kasneci, E.: A consistent and efficient evaluation strategy for attribution methods. In: International Conference on Machine Learning. pp. 18770–18795 (2022)

[55] Samek, W., Binder, A., Montavon, G., Lapuschkin, S., Müller, K.: Evaluating the visualization of what a deep neural network has learned. IEEE Trans. Neural Networks Learn. Syst. **28**(11), 2660–2673 (2017)

[56] Samek, W., Montavon, G., Vedaldi, A., Hansen, L.K., Müller, K.R. (eds.): Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Springer International Publishing (2019). https://doi.org/10.1007/978-3-030-28954-6, http://dx.doi.org/10.1007/978-3-030-28954-6

[57] Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M.: Striving for simplicity: The all convolutional net. In: ICLR Workshop (2015)

[58] Sturmfels, P., Lundberg, S., Lee, S.I.: Visualizing the impact of feature attribution baselines. Distill (2020). https://doi.org/10.23915/distill.00022, https://distill.pub/2020/attribution-baselines

[59] Su, J., Vargas, D.V., Sakurai, K.: One pixel attack for fooling deep neural networks. IEEE Transactions on Evolutionary Computation **23**(5), 828–841 (2019). https://doi.org/10.1109/TEVC.2019.2890858

[60] Sundararajan, M., Taly, A.: A note about: Local explanation methods for deep neural networks lack sensitivity to parameter values. CoRR **abs/1806.04205** (2018)

[61] Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: Precup, D., Teh, Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017. Proceedings of Machine Learning Research, vol. 70, pp. 3319–3328. PMLR (2017)

[62] Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: Precup, D., Teh, Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017. Proceedings of Machine Learning Research, vol. 70, pp. 3319–3328. PMLR (2017), http://proceedings.mlr.press/v70/sundararajan17a.html

[63] Theiner, J., Müller-Budack, E., Ewerth, R.: Interpretable semantic photo geolocalization. CoRR **abs/2104.14995** (2021), https://arxiv.org/abs/2104.14995

[64] Tomsett, R., Harborne, D., Chakraborty, S., Gurram, P., Preece, A.D.: Sanity checks for saliency metrics. In: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020. pp. 6021–6029. AAAI Press (2020)

[65] Vidovic, M.M.C., Kloft, M., Mueller, K.R., Goernitz, N.: Ml2motif—reliable extraction of discriminative sequence motifs from learning machines. PloS one **12**(3), e0174392 (2017)

[66] Wickstrøm, K.K., Trosten, D.J., Løkse, S., Boubekki, A., Mikalsen, K.Ø., Kampffmeyer, M.C., Jenssen, R.: RELAX: representation learning explainability. Int. J. Comput. Vis. pp. 1584–1610 (2023)

[67] Xiao, H., Rasul, K., Vollgraf, R.: Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms (2017)

[68] Yeh, C.K., Hsieh, C.Y., Suggala, A.S., Inouye, D.I., Ravikumar, P.: On the (in)fidelity and sensitivity of explanations. In: Neural Information Processing Systems (2019)

# A Appendix / supplemental material

## A.1 Related Works

**Metric-based Quality Estimation** Quantitative analysis of IAI explanation has improved considerably in recent years, and researchers now have a vast amount of evaluation metrics at their disposal (37; 2). Due to the lack of ground truth explanations, researchers try to quantify the quality of an explanation by measuring desirable properties, which can be categorized into 6 families of properties (37); faithfulness (11), robustness (3), localisation (63), complexity (24), randomisation (1), and axiomatic (41). Within each family, a variety of metrics exists.

**Prior Studies on Hyperparameter Sensitivity in IAI** Increasing attention has been given to the influence and potential confounding effects of hyperparameters in IAI evaluations (35). These studies vary in defining dependent versus independent variables and the hyperparameter space of intervention, be it model, explanation, or evaluation space. Studies have examined the sensitivity of attribution methods to explanation hyperparameters like random seed and number of samples (8), and the impact of baseline choices in methods like Integrated Gradients on explanation outcomes (58; 62). Additionally, the sensitivity of explanation outcomes concerning model performance variables such as optimizer, activation function, learning rate, and dataset split has been studied (39), along with the effects of model priors and random weight initialization on explanations and evaluations (33). Disagreement among different explanation methods regarding top-K features and ranking has also been investigated (45), while analyzing the impact of baselines (42).

Recently, researchers have explored how evaluation parameters affect outcomes, including the sensitivity of randomisation metrics to hyperparameters like normalisation, randomisation order, and similarity measures (12; 60; 36). Faithfulness metrics have been examined for hyperparameter influences such as baseline choice and perturbation order (55; 16; 15; 54; 64; 13; 48; 28). Unlike existing work, inspired by adversarial machine learning, we introduce a novel, general-purpose manipulation approach, applicable across a variety of evaluation approaches. Our findings reveal that faithfulness evaluation outcomes are highly susceptible to manipulation. This is a key issue for the IAI community to address. We put forward a preliminary mitigating solution for this in Sec. 3.

### A.1.1 Extended Results

**Exploring Variance in Target Manipulation** Fig A.1 shows the faithfulness score for each configuration in the feasible set for each dataset. This plot illustrates that the average faithfulness score across the feasible set can often be quite close. However, there is large spread in the scores, which is present for all datasets. This spread demonstrates the lack of robustness in the faithfulness evaluation and is part of the reason why manipulation is possible in this case. But, that alone would not be enough to allow for manipulation, since the different methods could have the same change in scores for different set of hyperparameters. However, the large standard deviation in Tab. 4 shows that is not the case, since the ranking change between sets of hyperparameters. In other words, the IAI methods react differently to different sets of hyperparameters. This, in combination with the variation shown in Fig. A.1, is what allows for manipulation in this study.

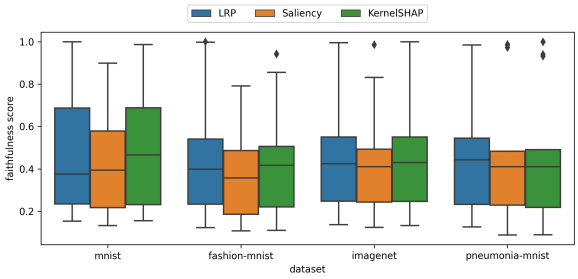


Figure A.1: Box plot showing faithfulness scores across all hyperparameter configurations in the feasible set for each dataset. The plot illustrates that the average faithfulness score is similar between different IAI methods across datasets. However the high variance enables a target manipulation. Note that the scores have been normalized dataset-wise by the highest score to allow for comparison across datasets.

**Inter-results** Below we include additional results for the inter-manipulation approach.

| | MNIST | | FASHIONMNIST | | PNEUMMNIST | | IMAGENET | |
|---|---|---|---|---|---|---|---|---|
| IAI METHOD | BASE | MANIP. | BASE | MANIP. | BASE | MANIP. | BASE | MANIP. |
| LRP | 25.19 | 51.41 | 21.46 | 43.80 | **21.31** | 25.86 | 129.61 | 167.14 |
| SALIENCY | **20.23** | **41.57** | **15.65** | **31.83** | 23.28 | **19.61** | **124.93** | **147.56** |
| KERNELSHAP | 23.94 | 49.25 | 21.45 | 37.36 | 22.06 | 19.99 | 128.72 | 167.74 |

Table A.1: Inter-results with manipulation towards *Saliency*. Lower is better.

| | MNIST | | FASHIONMNIST | | PNEUMMNIST | | IMAGENET | |
|---|---|---|---|---|---|---|---|---|
| IAI METHOD | BASE | MANIP. | BASE | MANIP. | BASE | MANIP. | BASE | MANIP. |
| LRP | 25.19 | 12.07 | 21.46 | 43.80 | **21.31** | 26.42 | 129.61 | 74.93 |
| SALIENCY | **20.23** | **9.72** | **15.65** | **31.83** | 23.28 | 19.95 | **124.93** | **74.21** |
| KERNELSHAP | 23.94 | 11.53 | 21.45 | 37.36 | 22.06 | **19.55** | 128.72 | 74.66 |

Table A.2: Inter-results with manipulation towards *KernelSHAP*. Lower is better.