

# The Directed Prediction Change - Efficient and Trustworthy Fidelity Assessment for Local Feature Attribution Methods

Kevin Iselborn\*, David Dembinsky\*, Adriano Lucieri, Andreas Dengel

German Research Center for Artificial Intelligence (DFKI) &  
Department of Computer Science, RPTU University Kaiserslautern-Landau  
67663 Kaiserslautern, Germany  
firstname.lastname@dfki.de

## Abstract

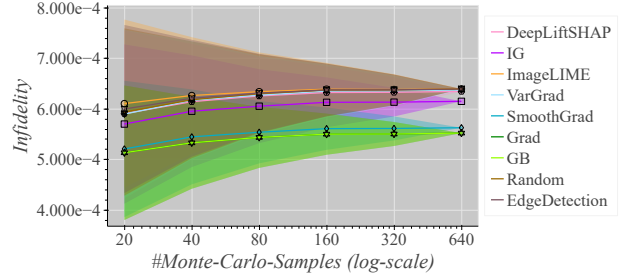
The utility of an explanation method critically depends on its fidelity to the underlying machine learning model. Especially in high-stakes medical settings, clinicians and regulators require explanations that faithfully reflect the model’s decision process. Existing fidelity metrics such as *Infidelity* rely on Monte Carlo approximation, which demands numerous model evaluations and introduces uncertainty due to random sampling. This work proposes a novel metric for evaluating the fidelity of local feature attribution methods by modifying the existing *Prediction Change (PC)* metric within the Guided Perturbation Experiment. By incorporating the direction of both perturbation and attribution, the proposed *Directed Prediction Change (DPC)* metric achieves an almost tenfold speedup and eliminates randomness, resulting in a deterministic and trustworthy evaluation procedure that measures the same property as local *Infidelity*. *DPC* is evaluated on two datasets (skin lesion images and financial tabular data), two black-box models, seven explanation algorithms, and a wide range of hyperparameters. Across 4 744 distinct explanations, the results demonstrate that *DPC*, together with *PC*, enables a holistic and computationally efficient evaluation of both baseline-oriented and local feature attribution methods, while providing deterministic and reproducible outcomes.

## 1 Introduction

The application of artificial intelligence (AI) in high-risk domains such as healthcare bears great potential, yet also entails substantial risks that must be mitigated through regulations demanding trustworthiness. Beyond thorough performance evaluation, explainable AI (XAI) methods aim to enhance the transparency of modern black-box models and facilitate their integration into clinical practice. The rapid development of AI has led to a vast variety of XAI techniques (Arrieta et al. 2020; Yang et al. 2023), making it increasingly difficult to select an appropriate method. Moreover, explanation algorithms in general, and feature attribution (FA) methods in particular, often involve numerous hyperparameters. Consequently, an extensive evaluation across many configurations is required to meaningfully assess the performance of any single algorithm.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

\*These authors contributed equally.



(a) ↓ Infidelity mean and approximation std. for all tested methods

	Number of Model Evaluations	Runtime [s] Mean / Median ± std
Infidelity	640	4924.83/5714.21 ± 1170.52
DPC (ours)	40	593.54/576.67 ± 42.35
Speedup		8.30/9.91

(b) Running time for an A100-80GB GPU when evaluating with *Infidelity* and *DPC* (the proposed metric) on skin lesion image data.

Figure 1: (a) The Monte Carlo sampling used by *Infidelity* exhibits high variance, requiring many samples for reliable FA evaluation (i.e., low standard deviation). (b) The proposed *DPC* metric achieves an almost tenfold speedup while providing a deterministic, and hence trustworthy, evaluation.

Existing analyses of explanation algorithms often rely on anecdotal evidence, that is, the qualitative inspection of exemplary explanations (Nauta et al. 2023; Dembinsky et al. 2025). While such anecdotal evidence can provide users with an intuition about explanation behavior, it is inherently subjective (Bućinca et al. 2020; Nauta et al. 2023; Dembinsky et al. 2025). Furthermore, human evaluation is prone to confirmation bias, focusing predominantly on plausibility while neglecting aspects such as the faithfulness of the explanation to the underlying decision process (Doshi-Velez and Kim 2017). To enable systematic and trustworthy evaluation, quantitative and functionally grounded evaluation metrics are therefore required to provide proxies for measuring explanation quality.

Recently, the *eValuation of XAI (VXAI)* framework extensively categorized such metrics (Dembinsky et al. 2025). Among the described desiderata (i.e., desirable atomic prop-

erties of explanations), fidelity is particularly important: without fidelity, even explanations that score well on other desiderata cannot provide meaningful insight into the model (Dembinsky et al. 2025) and are thus unsuitable for high-risk applications such as medicine or finance. The most common metrics in this category are based on the *Unguided* and *Guided Perturbation Experiment*, which evaluate fidelity through *input intervention* by perturbing input features and comparing the resulting change in model prediction with the effect predicted by the explanation (Dembinsky et al. 2025).

A widely used metric performing the *Unguided Perturbation Experiment* to compare FA methods is *Infidelity*. *Infidelity* estimates the expected agreement between an explanation and the corresponding change in model prediction under small perturbations using Monte Carlo sampling. However, ensuring a trustworthy evaluation of FA methods requires a large number of Monte Carlo samples, and thus repeated model evaluations, resulting in high computational cost (see Figure 1a and Section 4.2). This makes the scalable and trustworthy selection of FA methods impractical.

To address this limitation, this paper proposes a modification to the closely related *Guided Perturbation Experiment*. This work demonstrates that while the original Guided Perturbation Experiment is suitable only for FA methods that compare an input to a reference baseline (i.e., *baseline-oriented* methods), incorporating the direction of perturbation and attribution yields a metric that also applies to local FA methods. The resulting *Directed Prediction Change (DPC)* is a novel, computationally efficient, and deterministic metric that achieves an almost tenfold median speedup (see Figure 1b) compared to *Infidelity* (Yeh et al. 2019), which, to the best of our knowledge, is the only other metric targeting these explanation approaches.

## 2 Background

### 2.1 Feature attribution explanations

A FA method is a mapping that assigns each feature of an input a significance score, indicating which parts of the input were relevant for the predictor’s outcome. Let  $y$  be the target class of interest, scored by a black-box model  $f$ . Furthermore, let  $s_f^y : \mathbb{R}^d \rightarrow \mathbb{R}$  denote the scalar scoring function for class  $y$  and model  $f$ . A FA method is a function  $\mathcal{A}_f^y : \mathbb{R}^d \rightarrow \mathbb{R}^d$  that assigns a relevance score  $a_i := \mathcal{A}_f^y(x)_i$  to each feature  $i \in \{1, \dots, d\}$  of an input  $x \in \mathbb{R}^d$ .

Positive values ( $a_i > 0$ ) provide supporting evidence for the score of  $f$ , negative values ( $a_i < 0$ ) provide contradicting evidence. The magnitude  $|a_i|$  reflects the strength of the contribution. The interpretation of supporting or contradicting evidence depends on whether the FA method provides local or baseline-oriented explanations<sup>1</sup>.

For *local* methods, in a sufficiently small neighborhood around the data point, increasing  $x_i$  is expected to increase (support) or decrease (contradict) the score (Simonyan,

Vedaldi, and Zisserman 2013; Springenberg et al. 2014; Ribeiro, Singh, and Guestrin 2016).

For *baseline-oriented* methods, support (contradiction) indicates that the difference between the data point and the baseline in feature  $x_i$  increases (decreases) the score relative to the baseline (Sundararajan, Taly, and Yan 2017; Lundberg and Lee 2017). Baseline-oriented methods are therefore inherently contrastive (Sundararajan, Taly, and Yan 2017) and follow a different paradigm than local FA methods.

This analysis considers both local and baseline-oriented methods. To cover a wide range of approaches and include widely adopted techniques, we use the following methods:

- **Vanilla Gradient (Grad)** (Simonyan, Vedaldi, and Zisserman 2013):

$$\text{Grad}_f^y(x) = \nabla_x s_f^y(x) \quad (1)$$

- **Guided Backpropagation (GB)** (Springenberg et al. 2014): Modifies the backpropagation rules of ReLU activations as follows

$$\frac{\partial s_f^y(x)}{\partial g(x)} := R \cdot \mathbb{1}[g(x) > 0] \cdot \mathbb{1}[R > 0], \quad (2)$$

where  $g(x)$  denotes the input to a ReLU activation function, and  $R := \left( \frac{\partial s_f^y(x)}{\partial \text{ReLU}(g(x))} \right) / \left( \frac{\partial \text{ReLU}(g(x))}{\partial g(x)} \right)$  is the upstream gradient at that node.

- **SmoothGrad (SG)** (Smilkov et al. 2017) and **VarGrad (VG)** (Adebayo et al. 2018a): Compute the average or the variance of the FAs over noisy inputs, respectively.

$$\text{SG}_f^y(x) = \mathbb{E}_{X \sim \mathcal{N}(x, \sigma I)} [\nabla_X s_f^y(X)] \quad (3)$$

$$\text{VG}_f^y(x) = \mathbb{V}_{X \sim \mathcal{N}(x, \sigma I)} [\nabla_X s_f^y(X)] \quad (4)$$

- **Integrated Gradients (IG)** (Sundararajan, Taly, and Yan 2017): Compute the path integral of the gradients between a baseline  $x_0$  and the input to be explained.

$$\text{IG}_f^y(x) = (x - x_0) \int_0^1 \nabla_x s_f^y(x_0 + \alpha(x - x_0)) d\alpha, \quad (5)$$

- **LIME** (Ribeiro, Singh, and Guestrin 2016): Trains a linear model  $g \in G$  to locally approximate the model prediction and then uses the model as an explanation.

$$\text{LIME}_f^y(x) := \arg \min_{g \in G} \mathcal{L}(f, g, x) + \Omega(g), \quad (6)$$

where  $\mathcal{L}$  is a loss function and  $\Omega$  is a measure of complexity for the resulting explanation. This results in a large number of hyperparameters for LIME.

- **DeepLiftSHAP** (Lundberg and Lee 2017): Efficiently approximates Shapely Values (SV) by adapting the back-propagation rules of DeepLift.

$$\text{SV}_f^y(x)_i := \sum_{S \subseteq \{1, \dots, d\} \setminus \{i\}} \frac{|S|!(d-|S|-1)!}{d!} \left( s_f^y(x_{S \cup \{i\}}) - s_f^y(x_S) \right) \quad (7)$$

In addition to these FA methods, we also employ two model-agnostic reference methods, which we expect to be exceeded by the FA methods. We use importance values drawn from a standard gaussian distribution as *random explanations* for both considered data modalities. On image data, we additionally apply an *edge detection algorithm*.

A wide variety of hyperparameters is considered for all methods, with further details provided in the Appendix.

<sup>1</sup>Ancona et al. refer to these approaches as *local* and *global attributions* (Ancona et al. 2017). However, we find these terms misleading with respect to *local* and *global model explanations* as defined by other frameworks (Speith 2022; Molnar 2025).

## 2.2 Evaluation of XAI

**Unguided Perturbation Fidelity** Let  $\pi : \mathbb{R}^d \rightarrow \mathbb{R}^d$  denote a perturbation function sampled from a distribution  $\Pi$ , and let  $s_f^y : \mathbb{R}^d \rightarrow \mathbb{R}$  be the scoring function for class  $y$  and model  $f$ . For a FA method  $\mathcal{A}_f^y$ , the *Infidelity* metric (Yeh et al. 2019) on input  $x \in \mathbb{R}^d$  is defined as

$$P_f^y(\pi, x) := (x - \pi(x))^T \mathcal{A}_f^y(x) \quad (8)$$

$$S_f^y(\pi, x) := s_f^y(x) - s_f^y(\pi(x)) \quad (9)$$

$$\text{Inf}_f^y(x) := \mathbb{E}_{\pi \sim \Pi} \left[ (P_f^y(\pi, x) - S_f^y(\pi, x))^2 \right]. \quad (10)$$

*Infidelity* represents the expected mean squared error between the predicted perturbation effect  $P$  and the actual score change  $S$ . The choice of perturbation distribution  $\Pi$  is therefore crucial, as suitable choices enable the evaluation of both baseline-oriented and local FA methods. We consider adding gaussian noise with  $\sigma = 0.2$  to measure local *Infidelity*. Smaller values indicate a smaller mismatch and therefore higher fidelity ( $\downarrow$ ).

**Guided Perturbation Fidelity** In guided approaches, perturbations are applied sequentially to features according to their attribution ranking rather than being sampled randomly. At each step  $t$ , the *Prediction Change (PC)* is defined as the drop in class score when the next feature is removed:

$$\text{PC}_t^y(x) = s_f^y(\pi_t(x)) - s_f^y(\pi_{t-1}(x)), \quad (11)$$

where  $\pi_t : \mathbb{R}^d \rightarrow \mathbb{R}^d$  denotes the perturbation operator after  $t$  steps. Faithful explanations should cause large prediction drops (i.e., negative or low  $PC$ ) early when features are removed in descending order of importance (MoRF), and delayed drops when removed in the reverse order (LeRF). This behavior is typically quantified by the *Area Under the Perturbation Curve* (AUPC). To capture both perspectives, we report the weighted *Area Between Perturbation Curves* (ABPC) (Šimić, Sabot, and Veas 2022), which measures the gap between LeRF and MoRF curves, weighting early changes more strongly. For the resulting measure, larger values indicate a better FA ( $\uparrow$ ).

## 3 The Directed Prediction Change (DPC)

### 3.1 Prediction Change (PC) is insufficient for the evaluation of local FA methods

To motivate our novel metric, we first revisit the Sensitivity- $N$  property introduced by Ancona et al. (2019). A FA method satisfies Sensitivity- $N$  for an  $x \in \mathbb{R}^d$  if and only if, for a baseline  $x' \in \mathbb{R}^d$ , any subset  $S$  of the feature set  $X = \{1, \dots, d\}$  with  $|S| = N$  satisfies

$$\sum_{i \in S} \mathcal{A}_f^y(x)_i = s_f(x) - s_f(x_{X/S}; x'), \quad (12)$$

where  $x_{X/S}; x'$  denotes replacing the values in  $x$  by those from  $x'$  for the features in  $S$ .

Satisfying Sensitivity- $N$  for all  $N$  is a sufficient condition for baseline-oriented FA methods. Sensitivity-1 implies that

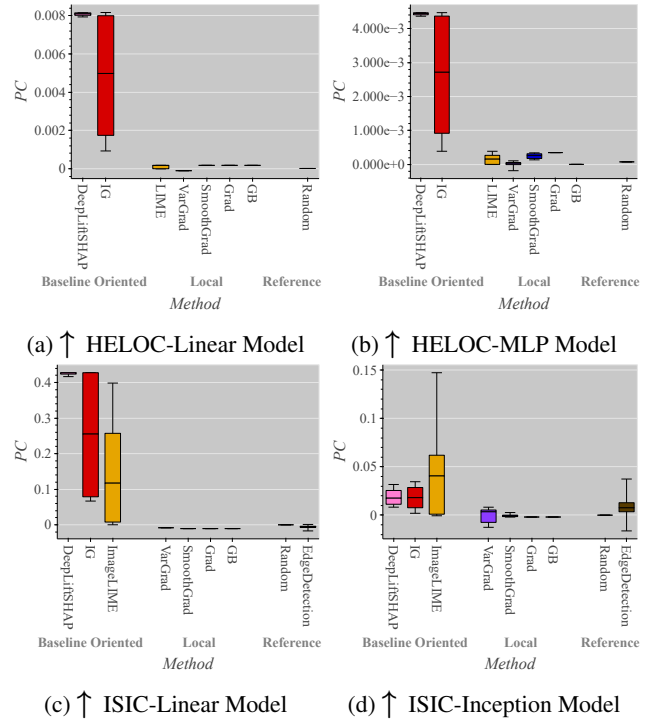


Figure 2: Overview of *Prediction Change (PC)* results for all considered models. The  $PC$  ranks baseline-oriented FA methods higher than other methods, whereas the *Gradient* (an optimal local FA method for linear models) performs comparably to random attribution.

each feature’s attribution exactly corresponds to the difference in model prediction, thereby fulfilling the condition of a baseline-oriented FA. Analogously,  $N > 1$  extends this to combinations of multiple features. Ancona et al. (2019) show that several FA methods, including *Integrated Gradients* and *DeepLift*, satisfy Sensitivity- $N$  for all  $N$  and  $x \in \mathbb{R}^d$  if applied to linear models. These can therefore be regarded as theoretically confirmed baseline-oriented FA methods.

Under mild assumptions (which hold for the FA methods analyzed by Ancona et al. (2019) in linear settings), Sensitivity- $N$  is also a sufficient condition for optimality according to the  $PC$  metric. We formalize this in the following theorem, with the proof provided in the Appendix:

**Theorem 1** Let  $x$  be the considered data point,  $x' \in \mathbb{R}^d$  be a baseline,  $\mathcal{A}$  a FA method with consistent rankings on all perturbation paths between  $x$  and  $x'$ , and  $s_f^y$  the scoring function for model  $f$  and class  $y$ .

If  $\mathcal{A}$  satisfies Sensitivity- $N$  for  $s_f^y$  on all perturbation paths between  $x$  and  $x'$  for all  $1 \leq N \leq d$ , then  $\mathcal{A}$  is optimal under evaluation by the *Prediction Change (PC)* on  $x$ .

This shows that the baseline-oriented FA methods analyzed by Ancona et al. (2019) are rated as optimal by  $PC$ . In this work, we consider *Integrated Gradients* and *DeepLiftSHAP*, which have been shown to satisfy Sensitivity- $N$  and

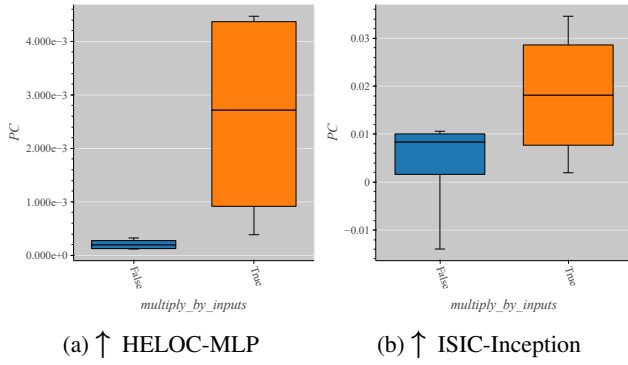


Figure 3: *Prediction Change (PC)* evaluation of *Integrated Gradients* for nonlinear models when baseline-oriented methods are converted into pseudo-local variants (*multiply\_by\_inputs* = False) following Ancona et al. (2019). As with the Sensitivity- $N$  property, the pseudo-local variants are consistently rated lower by *PC* than their baseline-oriented counterparts.

are thus optimal under *PC* for linear models<sup>2</sup>.

While this theoretical analysis does not extend to nonlinear models, we experimentally assess the FA methods across all hyperparameters described in the Appendix in the boxplots presented in Figure 2. As is done for all boxplots in this work, whiskers indicate maximal and minimal values to highlight the best and worst hyperparameter configurations observed. It is observed, that for all considered models (linear and nonlinear), baseline-oriented FA methods are rated superior by *PC*.

Ancona et al. (2019) further show that many baseline-oriented methods can be expressed using input multiplication. Therefore they can be reformulated as  $(x - x') \cdot z$  for some  $z \in \mathbb{R}^d$  and baseline  $x' \in \mathbb{R}^d$ , where  $\cdot$  denotes the Hadamard product. Retaining only  $z$  (and thus “removing the input multiplication”) then transforms a method into a pseudo-local FA method (Ancona et al. 2019), enabling a more detailed analysis of whether a metric favors local or baseline-oriented attributions.

Figure 3 presents the results for nonlinear models using *Integrated Gradients*. The pseudo-local variants are consistently rated worse than their baseline-oriented counterparts. Together, both experiments show that for both linear and nonlinear models, baseline-oriented FA methods are systematically favored by *PC*.

However, the gradient is ranked only marginally better than random attributions, which is problematic for linear models since it represents a well-established correct explanation of model behavior (see, e.g., Molnar (2025)). Furthermore, local FA methods are by design better suited to capture local model behavior than baseline-oriented methods. Accordingly, *PC* is not suitable when the goal is to evaluate explanations of local model behavior.

<sup>2</sup>DeepLiftSHAP inherits the Sensitivity- $N$  property from DeepLift through the use of the mean baseline.

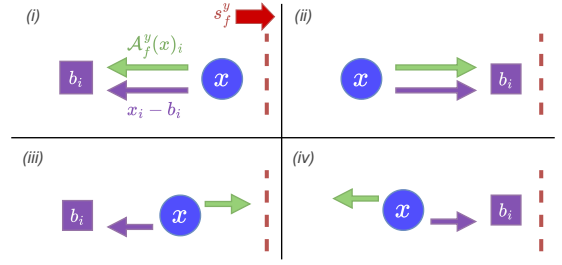


Figure 4: For a perturbation step  $t$  in which feature  $i$  of data point  $x$  is replaced with a baseline value  $b$ , four relevant evaluation scenarios arise. Cases (i) and (ii) represent situations where the *Prediction Change* correctly ranks the local FA method, whereas (iii) and (iv) illustrate failure cases.

### 3.2 An intuition for the *Directed Prediction Change* metric

We now demonstrate how the existing *PC* metric can be adapted to effectively evaluate local FA methods.

Recall that the *class score* is given by  $s_f^y(x)$ , where a higher value represents a higher likelihood for class  $y$ . For a local FA method, the *attribution score*  $a_i(x)$  for feature  $i$  is defined such that a positive value indicates that increasing  $x_i$  supports the score of class  $y$ , whereas a negative value indicates that decreasing  $x_i$  reduces  $s_f^y(x)$ .

We now analyze how the *PC* behaves when evaluating a local FA method. In a Guided Perturbation Experiment, attributions must rank features such that at each step the feature causing the minimum (MoRF order) or maximum (LeRF order) *PC* is selected. However, we observe that *PC* can only recognize this correctly if the attribution and perturbation directions align.

Consider the cases illustrated in Figure 4, where features are successively replaced with baseline values. Without loss of generality, assume that increasing feature  $x_i$  raises the model score  $s_f^y(x)$ . Under this assumption, we obtain:

- (i) If the attribution predicts a *decreasing model score* ( $a_i < 0$ ) and  $x_i$  *decreases*, then  $PC_t^y(x) > 0$ . Hence, *PC* correctly recognizes the FA method as erroneous. Case (ii) is analogous for correctly identified features
- (iii) If  $a_i > 0$  and  $x_i$  *decreases*, then  $PC_t^y(x) > 0$ , meaning that *PC* incorrectly evaluates the FA as erroneous for this feature. Case (iv) is the analogous incorrect recognition for the opposite case.

The *PC* thus evaluates the FA method as correct only when the perturbation and attribution point in the same direction. Because it relies solely on the change in model prediction, it cannot capture directional agreement. This observation provides the intuition for our new metric: we extend *PC* by incorporating information about the relative direction of the FA and the perturbation.

### 3.3 Formal definition of DPC

Mathematically, this directional information is fully represented by the signs of the perturbation and attribution. If the FA and perturbation have different signs, the regular *PC* must be inverted to yield a correct evaluation. This can be achieved by multiplying the *PC* with the signs of both the attribution and perturbation directions using the sign function  $\sigma(\cdot)$ . The resulting *Directed Prediction Change (DPC)* for perturbation step  $t$  is defined as:

$$DPC_t^y(x) := \sigma(a_i(x)) \cdot \sigma(\pi_t(x)_i) \cdot PC_t^y(x) \quad (13)$$

For perturbations affecting multiple features simultaneously (e.g., image data), we sum over all affected attributions and perturbations before computing their sign.

As in the case of *PC*, faithful explanations cause larger prediction drops (lower *DPC*) earlier or later depending on the evaluation order (MoRF or LeRF). We again report the weighted *Area Between Perturbation Curves* (ABPC) (Šimić, Sabol, and Veas 2022) as the resulting metric. Thus higher values indicate better performance ( $\uparrow$ ).

## 4 Experiments and Results

### 4.1 Setup

**Datasets** We employ two datasets from distinct modalities: a skin-lesion image dataset and a financial tabular dataset. All data samples are standardized, and both datasets are split into training, validation, and test sets using a 60:20:20 ratio.

For the tabular data, we use the Home Equity Line of Credit (HELOC) dataset (FICO 2018), which contains credit records labeled according to whether a loan was repaid or defaulted. To avoid distortions in the explanations caused by imputation strategies, we exclude the three features with the highest proportion of missing values and remove all remaining data points with missing entries. This yields 8 290 samples across 20 features. Models trained on this reduced dataset achieve performance comparable to those trained on the full version with imputation.

For the image data, we use the International Skin Imaging Collaboration (ISIC) challenge datasets covering the years 2016–2020 (Gutman et al. 2016; Codella et al. 2018, 2019; Tschandl, Rosendahl, and Kittler 2018; Rotemberg et al. 2021). The original task is multi-class classification of skin lesion types, which we simplify into a binary task distinguishing malignant from benign lesions. Following Cassidy et al. (2022), we merge the datasets across years, remove duplicates, and resize/crop all images to  $224 \times 224$  pixels. This results in 77 227 images with a strong class imbalance of approximately 80% benign samples.

**Models** We train two models per dataset: a simple linear model, offering higher interpretability, and a Deep Neural Network (DNN), representing a more complex nonlinear predictor. For the tabular data, the DNN is a custom Multi-Layer Perceptron (MLP) with seven hidden layers. For image data we employ the *InceptionV1* architecture (Szegedy et al. 2014). All models use ReLU activations and are described in more detail in the Appendix.

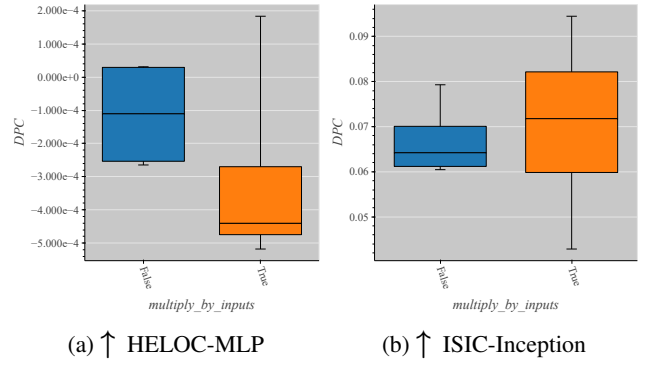


Figure 5: *Directed Prediction Change (DPC)* evaluation similar to Figure 3. *DPC* improves the scoring of local FA methods, with a weaker effect on image data than on tabular data.

**Explanation methods and metrics** Following common practice, we use model logits as the score function  $s_f^y$  for computing explanations and the predicted probabilities for evaluating explanation methods (Kokhlikyan et al. 2020). This simplifies the FA task for the considered piecewise linear neural networks.

As discussed in Section 2, all tested explanation methods expose a large number of hyperparameters, detailed in the Appendix. To enable a comprehensive analysis of the evaluated metrics, we compare FA methods across a total of 4 744 hyperparameter configurations. Unless otherwise noted, all experiments are performed on the validation splits to mimic a realistic hyperparameter-selection scenario for FA methods.

We employ the *Infidelity* metric implementation from the *Captum* framework (Kokhlikyan et al. 2020). For local FA evaluation, we follow the approach by Yeh et al. (2019), where a subset of  $k$  features is perturbed by adding Gaussian noise with  $\sigma = 0.2$ .

Guided Perturbation Experiments are implemented as successive replacement with the zero (mean) baseline. Unless otherwise specified, we perform 20 perturbation steps for Guided Perturbation Fidelity. This ensures that on tabular data, each feature is perturbed separately (since HELOC contains 20 features), while on image data, a comparable number of model evaluations is performed. We estimate *Infidelity* by sampling  $2^6 \cdot 20 = 1\,280$  perturbations on HELOC and  $2^5 \cdot 20 = 640$  on ISIC. The reduced sample count on ISIC is necessary due to the computational cost of *Infidelity*.

For all metrics, we employ a file-based cache for FA results, which substantially reduces computational demand and ensures that reported runtimes are independent of the explanation method used. Nevertheless, due to computational constraints, we subsample the ISIC validation split to 3 072 samples (from 11 487) when evaluating *Infidelity*.

### 4.2 Results and Analysis

In general, evaluating the quality of FA metrics is challenging. It is difficult to define a ground truth for the desired property and also difficult to measure it, even in controlled



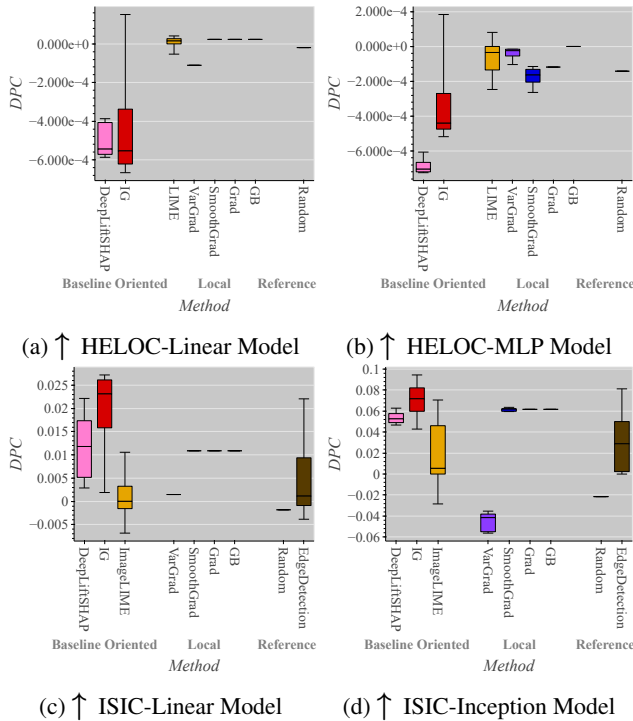


Figure 6: Evaluation of FA methods according to the *Directed Prediction Change (DPC)*. *DPC* favors local FA methods compared to baseline-oriented methods.

environments using synthetic datasets (Nauta et al. 2023; Dembinsky et al. 2025). Furthermore, when using synthetic data, it is unclear to what extent the results transfer to real data (Dembinsky et al. 2025).

***DPC* prefers local FA methods** First, we again resort to the experiments based on the approach by Ancona et al. (2019), where baseline-oriented FA methods are converted into pseudo-local variants. As shown in Figure 5, for non-linear models with *Integrated Gradients*, *DPC* consistently scores the pseudo-local variants higher (multiply\_by\_inputs = *False*). Results for additional method and model combinations are provided in the Appendix.

On the ISIC dataset this effect is less pronounced. Although *DPC* increases the relative scores of pseudo-local FA methods compared to *PC*, this increase can still leave relevant attributions ranked lower. One possible reason is the limited number of perturbations for this modality, which introduces variance due to summation over many perturbed regions. Another is the out-of-distribution issue described by Hooker et al. (2019), which may affect the perturbation experiment. Hence, *DPC* focuses more strongly on local FA than *PC*, but both metrics may require additional work to mitigate out-of-distribution effects, which is out of scope for this paper.

We next inspect overall performance rankings obtained from *DPC*. Figure 6 compares the methods under consideration according to *DPC*. For the linear model, *Gradient* is now evaluated substantially better than the other methods.

Model	Spearman Correlation
HELOC-Linear	−0.71
HELOC-MLP	−0.42
ISIC-Linear	−0.57
ISIC-Inception	−0.42

Table 1: Spearman correlations between the overall validation scores of local *Infidelity* and *DPC* for all four models. All models exhibit significant anti-correlations.

Model	Spearman Correlation	Pareto Size	Total Size
HELOC-Linear	−0.83	9	729
HELOC-MLP	0.70	21	729
ISIC-Linear	−0.07	11	1536
ISIC-Inception	−0.64	13	1536

Table 2: Spearman correlations between the overall validation scores of local *Infidelity* and *DPC* when evaluating only *LIME*. All models show a small Pareto set relative to the total number of configurations, and, except for ISIC-Linear and HELOC-MLP, a strong anti-correlation as expected.

Especially on HELOC, *DeepLiftSHAP* is often rated worse than a random attribution for the linear model. This indicates that *DPC* prefers more local methods than *PC*.

For the nonlinear models, the performances of individual FA methods overlap more strongly. Across all dataset and model combinations, the random baseline is generally rated worse under *DPC* than most FAs, which suggests that all analyzed methods encode some information about local model behavior.

**Comparing *DPC* and local *Infidelity*** We compare *DPC* with local *Infidelity*. First, we analyze the correlations between the two metrics. Since we test far more hyperparameter configurations for *LIME* than for the other methods, we consider *LIME* separately to avoid bias. Table 1 shows that *Infidelity* and *DPC* exhibit significant anti-correlations according to the Spearman rank coefficient. Because both metrics have opposing directions for optimal values, this suggests that they measure similar properties.

A detailed examination of *LIME* confirms this result. We analyze the Pareto set (Ehrgott 2005) for both metrics in Table 2. The Pareto set contains all configurations that are better in at least one metric and at least as good in the other than all other configurations. The small Pareto sets relative to the total number of configurations indicate strong agreement between the metrics on optimal hyperparameters.

Correlations are generally stronger in absolute value for *LIME* than for the model-wide aggregates excluding *LIME*. However, *LIME* shows little correlation for the linear model on ISIC and a positive correlation for the MLP on HELOC.

To investigate this more closely, we examine the scatter plots in Figure 7. No simple monotonic relationships are evident, which explains why the Spearman coefficient can be misleading in this context and motivates a more nuanced analysis.

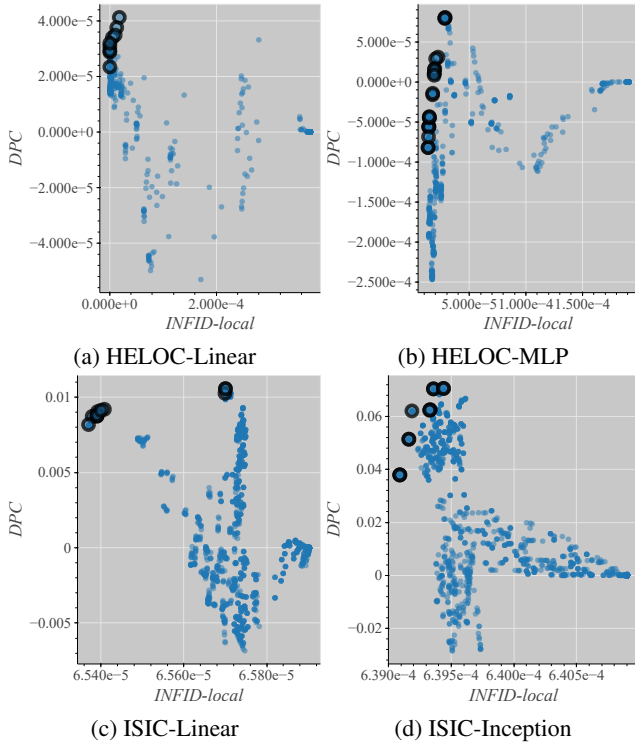


Figure 7: Scatter plots of local *InFidelity* (x-axis) and *DPC* (y-axis) across all tested *LIME* hyperparameters. Good configurations have low *InFidelity* ( $\downarrow$ ) and high *DPC* ( $\uparrow$ ). Pareto set elements are highlighted.

As a first observation, outlier hyperparameter configurations are identified for all four models. These are characterized by being rated as the worst configurations according to *InFidelity*, while *DPC* assigns them scores near zero (indicating that even worse configurations exist under *DPC*). Inspection of these cases reveals that the linear model trained within *LIME* assigns near-zero importance to all features due to excessive  $L_2$  regularization. In contrast to incorrect but seemingly plausible attributions, such degenerate explanations can be easily recognized by the users. This work hence concludes that *DPC* provides the more appropriate evaluation: *InFidelity* can rank erroneous explanations higher than configurations where the FA method clearly fails to produce a meaningful result, whereas *DPC* pushes such non-informative configurations toward the bottom of the ranking. This avoids scenarios in which users might prefer incorrect yet plausible-looking explanations over implausible ones.

For the ISIC-Linear model, we find additional outliers that receive high scores under *InFidelity*. These use the dataset mean as the pixel replacement value for *LIME* perturbations. Configurations ranked highly by *DPC* (which receive medium scores under *InFidelity*) instead use the mean of the explained image. Since the image mean is a more local perturbation than the dataset mean, we argue that these configurations better reflect local model behavior. These outliers illustrate that *InFidelity* can produce inaccurate evaluations relative to *DPC*.

Overall, most *InFidelity* scores lie in a very narrow range. In particular, for the HELOC-MLP model many configurations take nearly identical *InFidelity* values. Given the uncertainty observed in *InFidelity* evaluations in Figure 1, this lack of spread is problematic.

We suspect that this limited spread, together with the outliers, explains the unexpected positive correlation for HELOC-MLP and the weak anti-correlation ISIC-Linear. These effects support the effectiveness of *DPC* relative to *InFidelity*, rather than indicating errors by our metric.

**Determinism and computational efficiency** The computational efficiency of *InFidelity* depends strongly on the number of samples. We therefore study the variance as a function of the sample count, based on the Monte Carlo approximation in equation 10. Repeated evaluation with very large sample counts is infeasible when considering many FA methods and hyperparameters.

We approximate the variance as a measure of the Monte Carlo uncertainty as follows. First generate  $N = 640$  perturbations for ISIC are generated and the corresponding model scores are computed. We then draw 64 times  $N_{\text{perturb}} \in \{20, 40, 80, 160, 320, 640\}$  perturbations from these 640 and compute *InFidelity* for each draw. The standard deviation across these evaluations estimates the uncertainty. To summarize uncertainty at the method level, we average the aggregates (mean and standard deviation) across all hyperparameter settings of each method. Note, that larger ratios  $N_{\text{perturb}}/N$  induce greater overlap among sampled sets, which results in our procedure underestimating the true uncertainty.

Figure 1a shows the results. Despite the aforementioned bias, the  $\pm\sigma$  intervals show strong overlap with each other and their respective means. It is therefore difficult to distinguish methods or to differentiate between hyperparameter configurations without using large numbers of model evaluations.

Our *DPC* analysis uses 20 perturbation steps<sup>3</sup>, which is significantly more efficient than *InFidelity* where 640 perturbations are required for a reliable evaluation. We measure wall-clock runtime for both metrics when evaluating FAs of the Inception model on ISIC on a single A100-80GB GPU with the maximum feasible batch size. The mean, median, and standard deviation across 20 runs are reported in Table 1b. Even though our *DPC* implementation does not employ the same batching optimizations as Captum’s *InFidelity*, we observe a mean speedup of 8.30 and a median speedup of 9.91.

Finally, *PC* and *DPC* can be computed from the same Guided Perturbation Experiment without additional model evaluations, since only the scoring differs. By contrast, *InFidelity* requires separate perturbation sets to evaluate both baseline-oriented and local behavior. In scenarios where a holistic evaluation is desired, *DPC* together with *PC* yields an additional speedup factor of approximately 2, for a total median speedup of approximately 20 relative to *InFidelity*.

<sup>3</sup>This corresponds to 40 model evaluations because ABPC evaluates both LeRF and MoRF.

## 5 Conclusion

Motivated by the observation that existing metrics for evaluating local FA methods require many model evaluations due to Monte Carlo sampling, we developed an alternative metric that is both deterministic, and hence also trustworthy, and efficient. To this end, we analyzed the Guided Perturbation Experiment and found that evaluation with *Prediction Change (PC)* is unsuitable for evaluating local FA methods. Our main contribution is the *Directed Prediction Change (DPC)*, which integrates the direction of the attribution and the applied perturbations into the evaluation within the Guided Perturbation Experiment.

This modification enables effective evaluation of local FA methods and helps avoid error cases observed with the *Infidelity* metric. Furthermore, we achieve an almost tenfold median speedup over *Infidelity*, because *DPC* employs a deterministic procedure that does not require random sampling to obtain reliable results. The efficient and trustworthy evaluation enabled by *DPC* supports not only the final assessment of FA methods but also a holistic hyperparameter optimization, thereby contributing to the practical application of explainability methods in high-risk scenarios such as healthcare and finance.

**Limitations** A limitation of our metric is reduced accuracy when measuring the performance of local FA methods on complex data types with many features, such as image data. We hypothesize two causes. First, summation during aggregation across multiple perturbation steps can combine conflicting perturbation directions. Second, out-of-distribution effects may hamper the expressiveness for highly nonlinear models, even though our use of the weighted *Area Between Perturbation Curves* (ABPC) already alleviates this issue.

**Future Work** To analyze and potentially mitigate or solve these issues, future work could study the effect of the number of perturbation steps on *DPC* and explore adaptive choices for the perturbation step size. It would further be valuable to examine local perturbations for *DPC* instead of baseline replacement to counter out-of-distribution effects. Our analysis therefore opens several promising directions in the rapidly evolving field of VXAI. Since these issues only slightly reduced the effectiveness of our metric, we conclude that *DPC* is nonetheless a significant step toward an efficient and trustworthy evaluation of FA methods.

## 6 Acknowledgements

This research is funded by the German Federal Ministry for Digitalization and Government Modernization (BMDS) as part of the project MISSION KI - Nationale Initiative für Künstliche Intelligenz und Datenökonomie with the funding code 45KI22B021.

## References

Adebayo, J.; Gilmer, J.; Goodfellow, I. J.; and Kim, B. 2018a. Local Explanation Methods for Deep Neural Networks Lack Sensitivity to Parameter Values. *CoRR*, abs/1810.03307.

Adebayo, J.; Gilmer, J.; Muelly, M.; Goodfellow, I. J.; Hardt, M.; and Kim, B. 2018b. Sanity Checks for Saliency Maps. *CoRR*, abs/1810.03292.

Ancona, M.; Ceolini, E.; Öztireli, C.; and Gross, M. 2017. Towards better understanding of gradient-based attribution methods for deep neural networks. *arXiv preprint arXiv:1711.06104*.

Ancona, M.; Ceolini, E.; Öztireli, C.; and Gross, M. 2019. Gradient-Based Attribution Methods. In Samek, W.; Montavon, G.; Vedaldi, A.; Hansen, L. K.; and Müller, K.-R., eds., *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, 169–191. Cham: Springer International Publishing. ISBN 978-3-030-28954-6. [https://doi.org/10.1007/978-3-030-28954-6\\_9](https://doi.org/10.1007/978-3-030-28954-6_9).

Arrieta, A. B.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; García, S.; Gil-López, S.; Molina, D.; Benjamins, R.; et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion*, 58: 82–115.

Buçinca, Z.; Lin, P.; Gajos, K. Z.; and Glassman, E. L. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. *IUI '20*, 454–464. New York, NY, USA: Association for Computing Machinery. ISBN 9781450371186.

Canny, J. 1986. A Computational Approach to Edge Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(6): 679–698.

Cassidy, B.; Kendrick, C.; Brodzicki, A.; Jaworek-Korjakowska, J.; and Yap, M. H. 2022. Analysis of the ISIC image datasets: Usage, benchmarks and recommendations. *Medical Image Analysis*, 75: 102305.

Codella, N. C. F.; Gutman, D.; Celebi, M. E.; Helba, B.; Marchetti, M. A.; Dusza, S. W.; Kalloo, A.; Liopyris, K.; Mishra, N.; Kittler, H.; and Halpern, A. 2018. Skin lesion analysis toward melanoma detection: A challenge at the 2017 International symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC). In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, 168–172.

Codella, N. C. F.; Rotemberg, V.; Tschandl, P.; Celebi, M. E.; Dusza, S. W.; Gutman, D. A.; Helba, B.; Kalloo, A.; Liopyris, K.; Marchetti, M. A.; Kittler, H.; and Halpern, A. 2019. Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC). *CoRR*, abs/1902.03368.

Dembinsky, D.; Lucieri, A.; Frolov, S.; Najjar, H.; Watanabe, K.; and Dengel, A. 2025. Unifying VXAI: A Systematic Review and Framework for the Evaluation of Explainable AI. *arXiv:2506.15408*.

Doshi-Velez, F.; and Kim, B. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Ehrgott, M. 2005. Efficiency and Nondominance. In *Multi-criteria Optimization*, 23–64. Berlin, Heidelberg: Springer. ISBN 978-3-540-27659-3. [https://doi.org/10.1007/3-540-27659-9\\_2](https://doi.org/10.1007/3-540-27659-9_2).



- FICO. 2018. Home Equity Line of Credit (HELOC) Dataset (FICO Explainable Machine Learning Challenge). <https://community.fico.com/s/explainable-machine-learning-challenge>.
- Gutman, D. A.; Codella, N. C. F.; Celebi, M. E.; Helba, B.; Marchetti, M. A.; Mishra, N. K.; and Halpern, A. 2016. Skin Lesion Analysis toward Melanoma Detection: A Challenge at the International Symposium on Biomedical Imaging (ISBI) 2016, hosted by the International Skin Imaging Collaboration (ISIC). *CoRR*, abs/1605.01397.
- Ha, Q.; Liu, B.; and Liu, F. 2020. Identifying Melanoma Images using EfficientNet Ensemble: Winning Solution to the SIIM-ISIC Melanoma Classification Challenge. *CoRR*, abs/2010.05351.
- Hooker, S.; Erhan, D.; Kindermans, P.-J.; and Kim, B. 2019. A Benchmark for Interpretability Methods in Deep Neural Networks. <http://arxiv.org/abs/1806.10758>. arXiv:1806.10758.
- Ioffe, S.; and Szegedy, C. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *CoRR*, abs/1502.03167.
- Kingma, D. P.; and Ba, J. 2017. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*. <http://arxiv.org/abs/1412.6980>.
- Kokhlikyan, N.; Miglani, V.; Martin, M.; Wang, E.; Alsallakh, B.; Reynolds, J.; Melnikov, A.; Kliushkina, N.; Araya, C.; Yan, S.; and Reblitz-Richardson, O. 2020. Captum: A Unified and Generic Model Interpretability Library for PyTorch. <http://arxiv.org/abs/2009.07896>. arXiv:2009.07896.
- Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. <http://arxiv.org/abs/1711.05101>. arXiv:1711.05101.
- Lundberg, S. M.; and Lee, S. 2017. A unified approach to interpreting model predictions. *CoRR*, abs/1705.07874.
- Molnar, C. 2025. *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. 3 edition. ISBN 978-3-911578-03-5. <https://christophm.github.io/interpretable-ml-book>.
- Nauta, M.; Trienes, J.; Pathak, S.; Nguyen, E.; Peters, M.; Schmitt, Y.; Schlöterer, J.; Van Keulen, M.; and Seifert, C. 2023. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *ACM Computing Surveys*, 55(13s): 1–42.
- Nie, W.; Zhang, Y.; and Patel, A. 2018. A Theoretical Explanation for Perplexing Behaviors of Backpropagation-based Visualizations. In Dy, J.; and Krause, A., eds., *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, 3809–3818. PMLR.
- Reddi, S. J.; Kale, S.; and Kumar, S. 2019. On the Convergence of Adam and Beyond. <http://arxiv.org/abs/1904.09237>. arXiv:1904.09237.
- Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *CoRR*, abs/1602.04938.
- Rotemberg, V.; Kurtansky, N.; Betz-Stablein, B.; Caffery, L.; Chousakos, E.; Codella, N.; Combalia, M.; Dusza, S.; Guitera, P.; Gutman, D.; et al. 2021. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Scientific data*, 8(1): 34.
- Seo, J.; Choe, J.; Koo, J.; Jeon, S.; Kim, B.; and Jeon, T. 2018. Noise-Adding Methods of Saliency Map as Series of Higher Order Partial Derivative. <http://arxiv.org/abs/1806.03000>. arXiv:1806.03000.
- Shapley, L. S.; et al. 1953. A value for n-person games.
- Shrikumar, A.; Greenside, P.; and Kundaje, A. 2017. Learning important features through propagating activation differences. In *International conference on machine learning*, 3145–3153. PMIR.
- Šimić, I.; Sabol, V.; and Veas, E. 2022. Perturbation Effect: A Metric to Counter Misleading Validation of Feature Attribution. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, 1798–1807. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-9236-5. <https://dl.acm.org/doi/10.1145/3511808.3557418>.
- Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.
- Smilkov, D.; Thorat, N.; Kim, B.; Viégas, F. B.; and Wattenberg, M. 2017. SmoothGrad: removing noise by adding noise. *ArXiv*, abs/1706.03825.
- Sobel, I. 2014. An Isotropic 3x3 Image Gradient Operator. *Presentation at Stanford A.I. Project 1968*.
- Speith, T. 2022. A review of taxonomies of explainable artificial intelligence (XAI) methods. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, 2239–2250.
- Springenberg, J. T.; Dosovitskiy, A.; Brox, T.; and Riedmiller, M. 2014. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*.
- Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15(56): 1929–1958.
- Sundararajan, M.; and Najmi, A. 2020. The Many Shapley Values for Model Explanation. In *Proceedings of the 37th International Conference on Machine Learning*, 9269–9278. PMLR. <https://proceedings.mlr.press/v119/sundararajan20b.html>.
- Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic Attribution for Deep Networks. *CoRR*, abs/1703.01365.
- Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S. E.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; and Rabinovich, A. 2014. Going Deeper with Convolutions. *CoRR*, abs/1409.4842.
- Tschandl, P.; Rosendahl, C.; and Kittler, H. 2018. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1): 1–9.

Yang, W.; Wei, Y.; Wei, H.; Chen, Y.; Huang, G.; Li, X.; Li, R.; Yao, N.; Wang, X.; Gu, X.; et al. 2023. Survey on explainable AI: From approaches, limitations and applications aspects. *Human-Centric Intelligent Systems*, 3(3): 161–188.

Yeh, C.; Hsieh, C.; Suggala, A. S.; Inouye, D. I.; and Ravikumar, P. 2019. How Sensitive are Sensitivity-Based Explanations? *CoRR*, abs/1901.09392.

## 7 Appendix

### 7.1 Detailed Descriptions and Hyperparameters of the employed Explanation Algorithms

**Vanilla Gradient** The (Vanilla) Gradient method, introduced by Simonyan et al. (Simonyan, Vedaldi, and Zisserman 2013), explains a prediction by computing the sensitivity of the model output with respect to infinitesimal changes to the input features. It is defined as:

$$\text{Grad}_f^y(x) = \nabla_x s_f^y(x). \quad (14)$$

Features with larger gradients are assigned higher importance, as small changes in these features would cause larger changes in the prediction score.

**Guided Backpropagation** The Guided Backpropagation (GB) method, introduced by Springenberg et al. (Springenberg et al. 2014), extends the Vanilla Gradient method by modifying the backward pass through ReLU. Whereas the standard ReLU subgradient propagates gradients when the forward activation is positive, GB additionally suppresses negative upstream gradients. Formally, let  $g(x)$  denote the pre-activation input to a ReLU, and let  $R = \frac{\partial s_f^y(x)}{\partial \text{ReLU}(g(x))}$  be the upstream gradient at that node. GB replaces the standard ReLU derivative with

$$\frac{\partial s_f^y(x)}{\partial g(x)} := R \cdot \mathbb{1}[g(x) > 0] \cdot \mathbb{1}[R > 0], \quad (15)$$

i.e., the gradient is zeroed whenever the forward activation is non-positive or the incoming gradient is non-positive. Originally proposed as a visualization heuristic (Springenberg et al. 2014), GB has been criticized for producing edge-detector-like outputs and for limited class sensitivity (Nie, Zhang, and Patel 2018).

**Integrated Gradient** Sundararajan et al. propose the Integrated Gradients (IG) method as an attribution method that satisfies a set of desirable theoretical properties (Sundararajan, Taly, and Yan 2017). Rather than computing a single gradient, it computes attributions by integrating gradients along a path from a given baseline  $x_0$  to the input:

$$\text{IG}_{f,x_0}^y(x) = (x - x_0) \int_0^1 \nabla_x s_f^y(x_0 + \alpha(x - x_0)) d\alpha. \quad (16)$$

Sundararajan et al. show that this method distributes the difference in model score between the baseline and the input among the considered features, thereby providing a baseline-oriented FA method.

**SmoothGrad & VarGrad** A strategy to reduce artifacts in noisy FAs is presented by Smilkov et al. with SmoothGrad (SG) (Smilkov et al. 2017). They extend (in principle) any FA method by averaging the attributions over multiple small perturbations of the input. For simplicity, we implement their method by wrapping the Gradient method and employ Gaussian noise for generating perturbations, as in Smilkov et al., resulting in the explanation method:

$$\text{SG}_f^y(x) = \mathbb{E}_{X \sim \mathcal{N}(x, \sigma I)} [\nabla_X s_f^y(X)]. \quad (17)$$

A variance-based analog to SmoothGrad is presented by Adebayo et al. (Adebayo et al. 2018a) in the form of VarGrad:

$$\text{VG}_f^y(x) = \mathbb{V}_{X \sim \mathcal{N}(x, \sigma I)} [\nabla_X s_f^y(X)]. \quad (18)$$

While SmoothGrad reduces noise through averaging, VarGrad highlights high variance across perturbations, thereby estimating higher-order derivatives of the model score for the considered class (Seo et al. 2018).

**LIME** The Local Interpretable Model-Agnostic Explanations (LIME) method was introduced by Ribeiro et al. (Ribeiro, Singh, and Guestrin 2016) and approximates local model behavior without relying on the model’s gradient. It extracts attributions from an inherently interpretable model  $g$  with complexity  $\Omega(g)$ , which is trained through loss  $\mathcal{L}$  to mimic the predictive behavior of  $f$  in a neighborhood around the input  $x$ :

$$\text{LIME}_f^y(x) = \arg \min_{g \in \mathcal{G}} \mathcal{L}^y(f, g, x) + \Omega(g). \quad (19)$$

Similar to Ribeiro et al., we train a linear regression model as the interpretable model  $g$  and use its weights as attribution scores.

While this approach does not require the model to be differentiable, unlike the previously discussed methods, it introduces several parameters. First, the perturbation scheme and the number of perturbed samples must be defined. Second, a complexity measure has to be chosen. Finally, perturbations that are too far from the data point may no longer reflect local model behavior, which is why Ribeiro et al. propose a weighting function for distant samples (Ribeiro, Singh, and Guestrin 2016). We account for all these parameters in our experiments.

**SHAP** The SHapley Additive exPlanations (SHAP) method, introduced by Lundberg and Lee (Lundberg and Lee 2017), calculates attributions w.r.t. the average model-prediction baseline by leveraging cooperative game theory. Each feature is interpreted as a player in a cooperative game, and the payout (i.e., the difference in model prediction compared to the average model prediction) is distributed fairly among players. This is achieved using Shapley Values (Shapley et al. 1953); for details on how SHAP computes the average marginal contribution of a given feature across all feature subsets, we refer to Lundberg and Lee (Lundberg and Lee 2017).

Although SHAP satisfies desirable axiomatic properties (Lundberg and Lee 2017; Sundararajan and Najmi 2020), in practice an exact computation is intractable for more than a

handful of features. In this work we therefore use DeepLiftSHAP (Lundberg and Lee 2017), a model-specific implementation that combines the DeepLIFT modified back-propagation procedure (Shrikumar, Greenside, and Kundaje 2017) with Shapley-value-based weighting to approximate SHAP values efficiently for deep neural networks.

### Model-agnostic baselines

**Random attributions** Random explanations act as data-independent references that any meaningful explanation algorithm should outperform. We implement two variants: (1) a constant random attribution **RndC**, where a random vector is sampled once and used for all inputs; (2) a non-constant variant **RndNC**, where a new random attribution is sampled for each input:

$$\mathbf{RndC}_f^y(x) = r, \quad r \sim \mathcal{N}(0, I), \quad (20)$$

$$\mathbf{RndNC}_f^y(x) \sim \mathcal{N}(0, I). \quad (21)$$

**Edge detection** Since some FA methods behave similarly to edge detectors on image data (Adebayo et al. 2018b), we also include edge detection as a data-dependent naïve explanation. Specifically, we use two established computer vision techniques: (1) the Sobel gradient filter (Sobel 2014) and (2) the Canny edge detection algorithm (Canny 1986). To reduce sparsity and increase object coverage, we apply a Gaussian filter as a post-processing step and vary the filter width as a hyperparameter.

**Hyperparameters** We explore a large variety of hyperparameters across all FA methods. This includes, on the one hand, different baseline choices for Integrated Gradients and artificial biasing toward either class for the reference distribution of DeepLiftSHAP, and, on the other hand, the large variety of hyperparameters used by LIME as well as different choices for the noise used by SmoothGrad and VarGrad.

For all methods, except for the individual LIME approaches and EdgeDetection, two models on two datasets are considered (for the individual LIME approaches and EdgeDetection, two models on one dataset each are considered). This amounts to an analysis of a total of 4744 setups on the corresponding validation splits, presented in Table 3.

## 7.2 Proof of Theorem 1

Let  $\mathcal{A}$  be a FA method,  $x$  be the considered data point,  $x' \in \mathbb{R}^d$  be a baseline, and  $s_f^y$  be a scoring function for a model  $f$  and a class  $y$ . Further, let  $\mathcal{A}$  fulfill Sensitivity- $N$  with the baseline  $x'$  for all  $x'' \in \mathbb{R}^d$  on all perturbation paths between  $x$  and  $x'$  (inclusive) for all  $1 \leq N \leq d$ .

Let  $x'$  denote the baseline used by the Guided Perturbation Experiment,  $X := \{1, \dots, d\}$ , and

$$\tau \in X^d : \forall i, j \in X : i \neq j \Leftrightarrow \tau_i \neq \tau_j \quad (22)$$

We define the perturbation function  $\pi_t(x) = \pi_{t-1}(x_{x \setminus \{\tau_t\}}; x')$ , where  $x_{\{\tau_t\}}; x'$  denotes the replacement of the feature  $\tau_t$  with the corresponding feature of  $x'$ . We observe that  $\pi$  encompasses all valid perturbation functions used by the Guided Perturbation Experiment.

Next, we investigate an arbitrary set  $S \subseteq X, |S| \leq d$ . By the definition of Sensitivity- $N$  and by using a telescoping sum, we obtain:

$$\sum_{i \in S} \mathcal{A}_f^y(x)_i = s_f(x) - s_f(x_{X \setminus S}; x') \quad (23)$$

$$= s_f(x) - s_f(\pi_{|S|}(x)) \quad (24)$$

$$= \sum_{i=1}^{|S|} s_f^y(\pi_{i-1}(x)) - s_f^y(\pi_i(x)) \quad (25)$$

$$= \sum_{i=1}^{|S|} -\text{PC}_i^y(x) \quad (26)$$

We utilize this result to show the theorem by induction over the performed perturbation steps.

Since we can choose  $\tau_1$  freely, we obtain for the first perturbation step:

$$\mathcal{A}_f^y(x)_i = -\text{PC}_i^y(x) \forall i \in \{1, \dots, d\} \quad (27)$$

Hence, choosing the feature to perturb based on the largest attribution value will choose the feature causing the largest drop in prediction.

Now for any  $\tau_i, i > 1$  we first observe that  $\pi_{i-1}(x)$  is again a data point for which our previous analysis applies, and thus the attribution value of that data point is indicative of the feature causing the largest drop in prediction. Since, by assumption, the order of features by the FA method stays consistent along the perturbation path, we have that for all steps up to  $i$  the Prediction Change is maximal.

□

## 7.3 Models

For the HELOC dataset, the nonlinear model is a custom fully-connected neural network with sufficient capacity to achieve 99.9% training accuracy<sup>4</sup>. For the ISIC dataset, we employ the *InceptionV1* architecture (Szegedy et al. 2014) as a representative deep model. Both networks use ReLU activations.

To improve model performance, we apply suitable data augmentation strategies. On the HELOC dataset, we add Gaussian noise to the features. On the ISIC dataset, we follow the augmentation scheme of the winning submission to the ISIC 2020 Challenge (Ha, Liu, and Liu 2020).

All models are optimized on the training split of their respective datasets using a logistic regression loss and the AdamW optimizer with *amsgrad* (Kingma and Ba 2017; Reddi, Kale, and Kumar 2019; Loshchilov and Hutter 2019). We performed extensive hyperparameter tuning, ensuring that all models achieved competitive performance on their respective test splits, as shown in Table 4.

<sup>4</sup>The network has hidden dimensions [32, 128, 256, 128, 256, 128, 32] with batch-normalization (Ioffe and Szegedy 2015) and dropout (Srivastava et al. 2014).

Method Type	Method	Tested Hyperparameters	# Configurations
Propagation Based	Gradient	—	1
	GB	—	1
	IG	min, mean (0), median, and max baselines Input multiplication (True or False) 64 samples along a straight path	8
Perturbation Based	(Tabular) LIME	$\alpha \in \{0.00055, 0.0001, 0.00055, 0.001, 0.0055, 0.01, 0.055, 0.1, 0.55\}$ $\sigma_k \in \{0.25, 0.375, 0.5, 0.625, 0.75, 0.875, 1.0, 1.125, 1.25\}$ $\sigma_s \in \{0.1, 0.5, 1\}$ $n_{\text{samples}} \in \{64, 256, 1024\}$	729
	(Image) LIME	$\alpha \in \{0.00055, 0.001, 0.0055, 0.01, 0.055, 0.1, 0.55, 1\}$ $\sigma_k \in \{0.125, 0.25, 0.375, 0.5\}$ $n_{\text{samples}} = 1024$ Replacement value: Segment-Mean, Image-Mean, Dataset-Mean Segmentation algorithm: Quickshift or SLIC Seg. preprocess $\sigma \in \{0, 4\}$ Seg. Quickshift: max_dist $\in \{5, 7.5, 10, 200\}$ Seg. SLIC: $n_{\text{segments}} \in \{48, 64, 96, 128\}$	1536
	DeepLiftSHAP	Stratified baseline distribution with expected label $y \in \{0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$ or Random Input multiplication (True or False) $n_{\text{samples}} = 1024$ for tabular data $n_{\text{samples}} = 128$ for image data	24
Wrapping Approaches	SmoothGrad	$\sigma \in \{0.01, 0.1, 0.25, 0.5, 1\}$	5
	VarGrad	$\sigma \in \{0.01, 0.1, 0.25, 0.5, 1\}$	5
(Model-Agnostic) Baselines	Random Attribution	Use a constant value (True or False)	2
	Edge Detection	Postprocess $\sigma_{\text{post}} \in \{0.0, 2.0, 4.0, 8.0\}$ Sobel or Canny edge-detection algorithm Canny smoothing $\sigma_{\text{smooth}} \in \{1.0, 2.0, 4.0\}$	16

Table 3: List of all tested hyperparameter configurations for all FA methods explored in this work.

Model	AUROC	Accuracy
ISIC-Inception	92.91%	88.34%
ISIC-Linear	82.99%	69.29%
HELOC-MLP	79.72%	73.01%
HELOC-Linear	79.46%	73.43%

Table 4: Performance of all trained models.

#### 7.4 Additional input-multiplication experiments

We provide extended results when converting baseline-oriented FA methods into pseudo-local approaches as described in Section 3.1. Figure 8 shows the results of Integrated Gradients for both metrics and the employed linear models. Similarly, Figure 9 and Figure 10 show the results when investigating *DeepLiftSHAP*. We observe in these extended evaluations the same patterns described in the main paper: *PC* prefers baseline-oriented approaches, whereas *DPC* more strongly prefers local FA methods. We again find that the effect of our modification is weaker on image data but still visible.

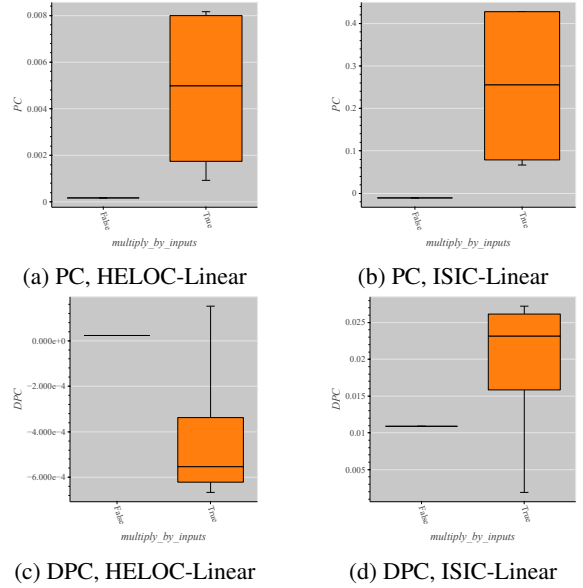


Figure 8: Directed Prediction Change (DPC) and Prediction Change (PC) evaluation as in Figure 3 for Integrated Gradients on linear models.

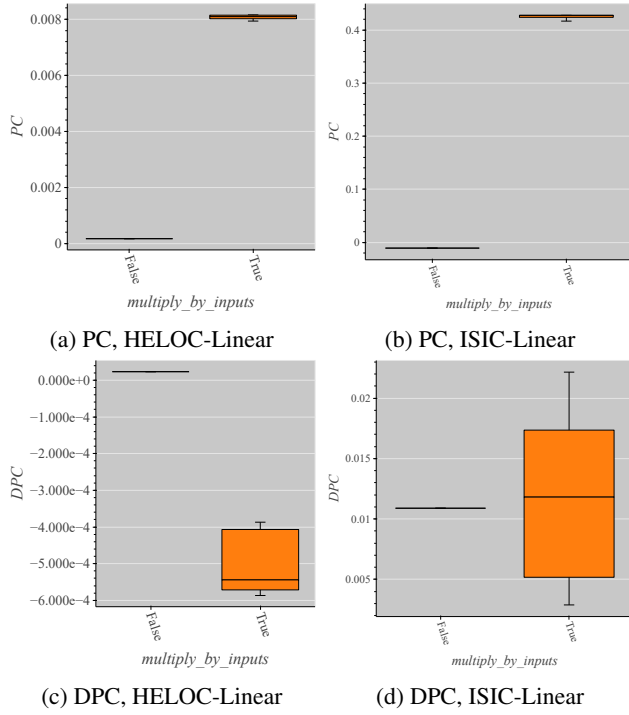


Figure 9: Directed Prediction Change (DPC) and Prediction Change (PC) evaluation as in Figure 3 for DeepLiftSHAP on linear models.

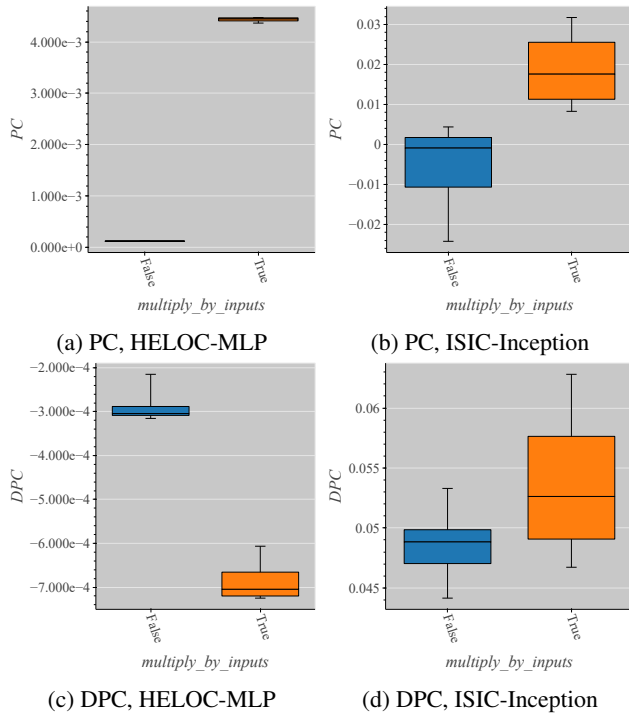


Figure 10: Directed Prediction Change (DPC) and Prediction Change (PC) evaluation as in Figure 3 for DeepLiftSHAP on nonlinear models.